Conférence Internationale Francophone sur la Science des Données (CIFSD) Actes de la 9e édition

Mohamed Quafafou

▶ **To cite this version:**

Mohamed Quafafou (Dir.). Conférence Internationale Francophone sur la Science des Données (CIFSD) Actes de la 9e édition. 2021. hal-03274095

HAL Id: hal-03274095
https://hal.science/hal-03274095

Submitted on 12 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

9e Conférence Internationale Francophone sur la Science des Données

9-11 juin 2021

Marseille

Aix-Marseille Université - LIS UMR 7020

https://cifsd-2021.sciencesconf.org

Avant-propos

De 2004 à 2016, la conférence a eu lieu tous les deux ans avec l'intitulé "Apprentissage Artificiel & Fouille de Données" (AAFD). Depuis 2018, la conférence a pris une dimension internationale et s'est transformée en une Conférence Internationale Francophone sur la Science des Données (CIFSD).

En raison de la situation sanitaire, la 9e édition de CIFSD se déroule en distanciel et est gratuite avec une inscription obligatoire. Pour cette édition, la thématique "science de données et santé" est mise en avant et ce choix prend tout son sens avec la période actuelle et l'émergence du vivant dans notre vie. Dans ce contexte, nous avons l'honneur et l'énorme plaisir de recevoir Thomas GRÉGORY, Jacques VAN HELDEN et Fabrizio MATURO, en tant que conférenciers invités pour présenter les avancés de l'intelligence artificielle dans trois domaines de la santé : la chirurgie, l'analyse de données multi-omiques et l'électrocardiogramme (ECG) :

- *La chirurgie guidée par l'Intelligence Artificielle*,
  par Thomas GRÉGORY
  Chef de service, Service de chirurgie de la main, du membre supérieur et du sport, Assistance Publique - Hôpitaux de Paris, Hôpital Avicenne, Bobigny,
  Professeur (PUPH) de chirurgie orthopédique et traumatologique, Université Sorbonne-Paris-Nord,
  Équipe Projet MOVEO (Intelligence Artificielle en Santé), LaMSN, Université Sorbonne-Paris-Nord,

- *Applications de l'apprentissage aux données multi-omiques de biologie-santé*,
  par Jacques VAN HELDEN
  Codirecteur de l'Institut Français de Bio-informatique (IFB),
  Professeur de bio-informatique, Université d'Aix-Marseille (AMU),

- *Supervised classification of ECG curves via a combined use of functional data analysis and tree-based methods to identify people affected by heart disease*,
  par Fabrizio MATURO
  Adjunct Professor at the Department of Mathematics and Physics,
  University of Campania "Luigi Vanvitelli" in Caserta, Italy.

En plus de ces trois conférences, 24 articles présentent les résultats d'équipes de chercheurs provenant de 8 pays différents dont la France. Ces articles sont organisés en 6 sessions dont 4 portant sur les aspects généraux de la science de données, une session sur ses applications et une dernière session sur la science de données et santé. Tous ces travaux sont regroupés dans ce document scientifique de très bonne qualité mis à disposition en ligne sur l'archive ouverte HAL (https://hal.archives-ouvertes.fr).

Cinq articles sont pré-sélectionnés pour le prix du meilleur article. Trois seront classés et annoncés à la fin de la conférence lors de la remise des prix.

Je tiens à remercier chaleureusement :

- Les auteurs pour leurs contributions,
- Le comité de programme pour la qualité de leurs rapports,
- Le comité d'organisation pour le très bon travail réalisé dans un contexte anxiogène.

Mohamed QUAFAFOU
Aix-Marseille Université, Laboratoire d'Informatique et Systèmes
Président du comité de programme de CIFSD 2021

Comité de pilotage

Younès BENNANI, Université Sorbonne Paris Nord
Abdelouahid LYHYAOUI, ENSAT
Mohamed QUAFAFOU, Aix-Marseille Université
Said RAGHAY, Université Cadi Ayyad - SASD
Abdelfattah TOUZANI, Université SMBA
Emmanuel VIENNET, Université Sorbonne Paris Nord

Comité de programme

Président du comité de programme :
Mohamed QUAFAFOU, Aix-Marseille Université

Massih-Reza AMINI, Université Grenoble Alpes
Thierry ARTIÈRES, Ecole Centrale Marseille
Hadj BATATIA, Heriot-Watt University, Dubai Campus
Khalid BENABDESLEM, Université Claude Bernard Lyon 1
Younès BENNANI, Université Sorbonne Paris Nord
Gilles BISSON, CNRS-LIG Grenoble
Paula BRITO, Université de Porto, Portugal
Stéphane CANU, INSA de Rouen
Guillaume CLEUZIOU, Université d'Orléans
Guy CUCUMEL, Université du Québec à Montréal, Canada
Jean DIATTA, Université de La Réunion
Richard EMILION, Université d'Orléans
Patrick GALLINARI, Sorbonne Université
Pierre GANÇARSKI, Université de Strasbourg
Eric GAUSSIER, Université Grenoble Alpes
Nadia GHAZZALI, Université du Québec à Trois-Rivières, Canada
Nistor GROZAVU, Université Sorbonne Paris Nord
Yann GUERMEUR, CNRS-LORIA Nancy
André HARDY, Université de Namur, Belgique
Zahi JARIR, Université Cadi Ayyad, Marrakech, Maroc
Léonard KWUIDA, Université de Bern, Suisse
Lazhar LABIOD, Université Paris Descartes
Philippe LERAY, Université de Nantes
Lotfi LAKHAL, Aix-Marseille Université
Vladinir MAKARENKOV, Université du Québec à Montréal, Canada
Franck MARZANI, Université de Bourgogne
Engelbert MEPHU NGUIFO, Université Clermont-Ferrand
Mathilde MOUGEOT, Université Paris Diderot
Mohamed NADIF, Université Paris Descartes
Bruno PINAUD, Université de Bordeaux

El Mostafa QANNARI, ONIRIS Nantes
Agus Budi RAHARJO, Institut de technologie Sepuluh Nopember, Surabaya, Indonésie
Gilbert RITSCHARD, Université de Genève, Suisse
Nicoleta ROGOVSCHI, Université Paris Descartes
Fabrice ROSSI, Université Paris Dauphine
Lorenza SAITTA, Université de Turin, Italie
Fabien TORRE, Université de Lille
Michel VERLEYSEN, Université de Louvain, Belgique
Emmanuel VIENNET, Université Sorbonne Paris Nord
Cédric WEMMERT, Université de Strasbourg
Djamel Abdelkader ZIGHED, Université de Lyon
Jean-Daniel ZUCKER, Institut de Recherche pour le Développement

Comité d'organisation

Nicolas DURAND, Aix-Marseille Université
Alain CASALI, Aix-Marseille Université
Sébastien MAVROMATIS, Aix-Marseille Université

Table des matières

Articles

Vers une régression Laplacienne semi-supervisée et multi-labels

Vivien Kraus*, Khalid Benabdeslem**, Bruno Canitia***

*Université Lyon 1, 43, Bd du 11 Novembre 1918, Villeurbanne, Cedex 69622, France
vivien@planete-kraus.eu
**Université Lyon 1, 43, Bd du 11 Novembre 1918, Villeurbanne, Cedex 69622, France
kbenabde@univ-lyon1.fr
***Lizeo IT, 42 Quai Rambaud, 69002 Lyon, France
bruno.canitia@lizeo-group.com

Résumé. Pour l'apprentissage multi-labels, il est nécessaire de trouver un modèle adapté pour la prédiction de valeurs multiples simultanément pour un même individu, à partir des mêmes variables. L'efficacité de la plupart des algorithmes multi-labels tient au fait qu'ils sont capables de prendre en considération les corrélations entre labels reliés. D'un autre côté, dans de nombreuses applications, une tâche d'apprentissage multi-labels induit un coût d'annotation élevée pour une seule observation. Ceci conduit à des jeux de données qui consistent en seulement quelques points labellisés, et de nombreux points non labellisés. Dans ce scénario, des méthodes semi-supervisées peuvent tirer avantage des points non labellisés. Dans cet article, nous proposons un nouvel algorithme pour la régression multi-labels semi-supervisée, LSMR, comme extension multi-labels d'un algorithme de régression semi-supervisée. Nous proposons des résultats expérimentaux sur des jeux de données de régression disponibles publiquement montrant l'intérêt de notre approche.

Mots-clés : Apprentissage multi-labels, Apprentissage semi-supervisé, Régression, Régularisation Laplacienne.

1 Introduction

L'un des défis principaux de l'apprentissage automatique consiste à apprendre à partir de données labellisées à la main ainsi que de données non labellisées, qui sont généralement plus faciles à obtenir. L'apprentissage faiblement labellisé (Li et al., 2013), et plus particulièrement l'apprentissage semi-supervisé (Chapelle et al., 2006; Zhu et al., 2003), abordent ce défi en utilisant soit des approches de propagation de labels pour suppléer des méthodes supervisées (Zhu et al., 2003; Zhao et al., 2015), ou utilisent l'information des labels comme contraintes pour des méthodes non supervisées (Basu et al., 2008; Zhang et al., 2012).

De nombreuses méthodes semi-supervisées fondées sur des méthodes supervisées ont été proposées. Par exemple, on retrouve des adaptations telles que le *self-training* (Scudder, 1965) ou *co-training* (Blum et Mitchell, 1998; Zhou et Li, 2005), des cas d'apprentissage *transductif* (Joachims, 1999; Bennett et Demiriz, 1999), ou des méthodes génératives (Nigam et al., 2000).

Dans le cas particulier de la régression, avec une prédiction numérique, des applications spécifiques sont proposées (Azriel et al., 2016; Ryan et Culp, 2015), et on retrouve principalement des approches fondées sur l'apprentissage de représentation de l'espace (Ji et al., 2012) ou utilisant un graphe entre individus tenant compte de la similarité (Moscovich et al., 2016), avec une insistance particulière sur la régularisation Laplacienne (Cai et al., 2006; Belkin et al., 2006).

De plus, l'apprentissage multi-labels vise à exploiter les relations entre *labels* pour apporter plus d'information à la tâche d'apprentissage. Des travaux théoriques ont été proposés (Gasse et al., 2015), et de nombreuses méthodes multi-labels ont été proposées (Borchani et al., 2015; Spyromitros-Xioufis et al., 2016). Certaines stratégies transforment un problème multi-labels en plusieurs problèmes mono-label (Read et al., 2011; Liu et al., 2017), et il est parfois possible d'étendre directement le problème au cadre multi-labels, comme pour les processus gaussiens (Yu et al., 2005).

De nombreux algorithmes d'apprentissage multi-labels minimisent une fonction objectif régularisée. Cela peut être effectuée via une régularisation non-convexe (Gong et al., 2012; Shi et al., 2018), mais les approches convexes sont plus fréquentes, avec l'adaptation de l'algorithme LASSO (Tibshirani, 1996a) au cadre multi-labels (Argyriou et al., 2007; Jian et al., 2016), ou le *Clustered Multik-Task Learning* (Zhou et al., 2011), ou d'autres approches d'apprentissage de représentation de l'espace des labels (Chen et Lin, 2012; Luo et al., 2017).

Pour l'apprentissage de régression, les méthodes spécifiques sont plus rares. Il est d'usage de modifier la fonction objectif pour utiliser des labels à valeurs réelles (Zhou et al., 2012).

2 Travaux liés

Pour l'apprentissage de régression mono-label, nous nous fondons sur l'algorithme *SSSL* Ji et al. (2012). La caractéristique principale de cet algorithmes réside dans son fonctionnement en deux étapes, qui permettent dans un premier temps de faire un changement d'espace *non-supervisé* propice à l'apprentissage de régression de certains jeux de données, puis d'adopter une *régression linéaire* dans ce nouvel espace.

Nous fondons également notre travail sur la régularisation Laplacienne Belkin et al. (2006), qui permet de lier les individus non labellisés dans le but de satisfaire l'hypothèse suivante :

Si deux individus sont proches dans l'espace réel, alors l'écart de prédiction doit être faible.

Comme évoqué ci-dessus, il est aussi possible d'utiliser la régression Laplacienne pour traiter le problème d'apprentissage multi-labels (Zhou et al., 2012). Dans cette approche, on construit un graphe des labels, et il s'agit de régulariser le modèle de façon à introduire l'hypothèse suivante :

Si deux labels sont similaires, les valeurs du modèle pour ces deux labels doivent être similaires.

3 Approche proposée : LSMR

3.1 Notations

Les individus sont décrits dans l'espace des variables supposé être de dimension d, $\mathbb{R}^d$. L'ensemble des labels est de dimension m. Étant donné que le problème est celui de la régression, cet espace est décrit dans $\mathbb{R}^m$.

Notons $\mathcal{V} \in \mathbb{R}^{N,d}$ la matrice de données, dont les lignes correspondent aux individus. On note n_l le nombre d'individus labellisés au total.

Associée à cette matrice de données, la matrice de labels de régression $\mathbf{Y} \in \mathbb{R}^{N,m}$ contient une ligne par individu ; seules les lignes correspondant aux individus labellisés sont renseignées. On suppose que tous les labels sont manquants simultanément : si un individu est labellisé, les valeurs de tous les labels sont renseignés. Sinon, aucune valeur n'est renseignée. On note $\mathbf{Y_l}$ la sous-matrice dont les lignes sont les individus labellisés.

3.2 Changement d'espace

Comme pour l'algorithme *SSSL*, une première étape consiste à choisir une fonction noyau symétrique, $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, et à construire la matrice noyau entre tous les individus de l'ensemble d'apprentissage :

$$\forall i, j \in \{1, ..., N\}, \quad \mathbf{K}_{i,j} = \kappa(\mathcal{V}_{i,.}, \mathcal{V}_{j,.}) \tag{1}$$

$\mathbf{K}$ étant une matrice symétrique réelle, on peut en extraire les s valeurs propres réelles les plus élevées, pour $s \leq N$, et les vecteurs propres associés. Ces derniers forment une matrice $\mathbf{U} \in \mathbb{R}^{N,s}$, dont les colonnes sont de norme unitaire, où chaque ligne correspond à un individu de l'ensemble d'apprentissage. Étant donné que l'on n'utilise pas la matrice de labels, cette première étape est non supervisée.

Le changement d'espace s'obtient en posant $\mathbf{X} \leftarrow \mathbf{KU}$. La matrice $\mathbf{X}$ comprend donc une ligne par individu, et nous pouvons en extraire la sous-matrice correspondant aux individus labellisés, $\mathbf{X_l} \in \mathbb{R}^{n,s}$. En utilisant ce changement d'espace, notre approche proposée constitue donc une extension de l'algorithme SSSL.

3.3 Régularisation Laplacienne

Nous proposons de remplacer la régression simple utilisée dans l'algorithme *SSSL* par une régularisation Laplacienne semi-supervisée, et une régularisation multi-labels. Ainsi, en définissant le graphe des labels par la matrice d'adjacence $\mathbf{M_m} \in \mathbb{R}^{m,m}$, et le graphe des individus par la matrice d'adjacence $\mathbf{M_s} \in \mathbb{R}^{N,N}$, le terme à pénaliser devient :

$$\underset{\mathbf{W} \in \mathbb{R}^{d,m}}{\text{minimize}} \quad \|\mathbf{X_l W} - \mathbf{Y_l}\|_F^2 + \alpha \text{tr}(\mathbf{W'X'L_s XW}) + \beta \text{tr}(\mathbf{WL_m W'}) \tag{2}$$

Où α est le régulariseur semi-supervisé, β est le régulariseur multi-labels, $\mathbf{L_s}$ est la matrice Laplacienne du graphe des individus, $\mathbf{L_s} = \mathbf{D_s} - \mathbf{M_s}$ avec $\mathbf{D_s}$ la matrice de degré du graphe des individus, $\forall i \in \{1, ..., N\}, \quad \mathbf{D_s}(i,i) = \sum_{j=1}^{N} \mathbf{M_s}(i,j)$, et $\mathbf{L_m}$ est la matrice Laplacienne du graphe des labels, $\mathbf{L_m} = \mathbf{D_m} - \mathbf{M_m}$ avec $\mathbf{D_m}$ la matrice de degré du graphe des labels, $\forall k \in \{1, ..., m\}, \quad \mathbf{D_m}(k,k) = \sum_{l=1}^{m} \mathbf{M_m}(k,l)$. La matrice $\mathbf{W}$ représente le modèle linéaire de la régression dans cet espace.

Le premier terme mesure l'écart de prédiction pour les individus labellisés, vis-à-vis de tous les labels. Il est possible de l'écrire ainsi puisque les individus labellisés sont complètement annotés. Nous ne traitons pas le cas d'une annotation plus faible, pour lequel la somme doit s'effectuer uniquement pour certains couples (individu, label).

Afin de préserver l'hypothèse de la régression Laplacienne pour l'apprentissage semi-supervisé, le graphe des individus doit être constitué à partir des données de l'espace original. Il est possible d'utiliser le même noyau que pour le changement d'espace, mais l'interprétation des contraintes imposées par ce graphe d'individus, qui pénalise l'écart de prédiction entre individus similaires dans l'espace réel, invite à rendre la matrice d'adjacence $\mathbf{M_m}$ éparse afin de se concentrer sur les similarités les plus importantes. Ce n'est pas le cas pour le noyau employé pour le changement d'espace.

Dans la suite, nous appellerons la modification proposée *LSMR*, pour *Laplacian-regularized Simple algorithm for Semi-supervised multi-labels Regression*.

3.4 Algorithme d'optimisation

La fonction objectif employée dans l'algorithme *SSSL* admet une solution analytique, qui permet d'obtenir la valeur du modèle $\mathbf{W}$ grâce à une résolution d'un système linéaire de dimension $s \times s$. Malheureusement, on ne peut pas l'appliquer directement pour *LSMR*, à moins d'avoir à résoudre m^2 systèmes linéaire distincts. Cette limitation est commune à beaucoup de méthodes de régularisation multi-labels, ce qui pousse les méthodes de MALSAR [1] à adopter une méthode d'optimisation différente.

1. http://jiayuzhou.github.io/MALSAR/

La fonction objectif obtenue dans (2) présente les caractéristiques suivantes :
— la variable d'optimisation, $\mathbf{W}$, est réelle ;
— la fonction objectif est convexe, si les arêtes des matrices d'adjacence des graphes des individus et des labels sont à poids positifs ;
— la fonction objectif est lisse.

Par conséquent, le problème peut être résolu par une descente de gradient, dont le calcul est donné par (3).

$$\nabla_{\mathbf{W}} = 2\left[\mathbf{X_l}'\mathbf{X_l} + \alpha\mathbf{X}'\mathbf{L_s}\mathbf{X}\right]\mathbf{W} - 2\mathbf{X_l}'\mathbf{Y_l} + 2\beta\mathbf{W}\mathbf{L_m} \tag{3}$$

Le modèle $\mathbf{W} \in \mathbb{R}^{s,m}$ est initialisé comme la solution du problème de régression linéaire, avec un terme de régularisation Ridge utilisant une valeur de régulariseur faible.

$$\mathbf{W} \leftarrow \left[\mathbf{X_l}'\mathbf{X_l} + \epsilon\mathbf{I_s}\right]^{-1}\mathbf{X_l}'\mathbf{Y_l} \tag{4}$$

Avec ϵ faible (pour l'implémentation, nous avons retenu $\epsilon = 10^{-6}$), et $\mathbf{I_s} \in \mathbb{R}^{s,s}$ la matrice identité en dimension s.

L'algorithme de descente du gradient produit des itérations de descente de l'erreur d'apprentissage jusqu'à convergence de la variable d'optimisation. Chaque itération consiste à calculer la valeur du gradient au point donné pour la valeur courante de la variable d'optimisation, $\nabla_{\mathbf{W}}$ (selon (3)), puis poser :

$$\mathbf{W} \leftarrow \mathbf{W} - \eta\nabla_{\mathbf{W}} \tag{5}$$

où η est le *pas d'apprentissage*. Dans notre cas, le gradient (3) est une fonction Lipschitzienne : pour deux valeurs quelconques du modèle, $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{s,m}$,

$$\|\nabla_{\mathbf{W}}(\mathbf{P}) - \nabla_{\mathbf{W}}(\mathbf{Q})\|_F^2 \leq C\,\|\mathbf{P} - \mathbf{Q}\|_F^2 \tag{6}$$

Avec :

$$C = 2\left(\rho\left(\mathbf{X_l}'\mathbf{X_l} + \alpha\mathbf{X}'\mathbf{L_s}\mathbf{X}\right) + \beta\rho(\mathbf{L_m})\right) \tag{7}$$

Pour toute matrice $\mathbf{M}$ symétrique réelle, $\rho(\mathbf{M})$ désigne le rayon spectral de $\mathbf{M}$, c'est-à-dire sa plus grande valeur propre. En considérant un pas d'apprentissage $\eta = \frac{1}{C}$, la convergence de l'algorithme de descente du gradient est assurée (Nesterov, 2007).

L'algorithme de descente de gradient accélérée (Nesterov, 2007) est une amélioration de l'algorithme de descente de gradient originale qui propose une convergence plus rapide. Pour notre application, l'algorithme d'optimisation est résumé dans l'algorithme 1. La prédiction s'effectue ainsi avec l'algorithme 2, qui est simplement la version multi-labels de l'algorithme de prédiction de *SSSL*, en remplaçant le modèle de dimension s par un modèle de dimension $s \times m$.

4 Expérimentations

Afin de s'assurer de la performance de notre approche, nous proposons deux études expérimentales. Tout d'abord, puisqu'elle généralise une approche existante, nous devons nous assurer que cette extension est nécessaire. Puis nous montrerons quelques comparaisons avec d'autres régularisations pour la régression multi-labels.

Algorithm 1 LSMR : Apprentissage

Données : $\mathcal{V} \in \mathbb{R}^{N,d}$, $\mathbf{Y} \in \mathbb{R}^{n_l,m}$
Hyperparamètres : $s \in \{1, ..., N\}$, $\kappa \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
Hyperparamètre : matrice Laplacienne du graphe des individus, $\mathbf{L_s} \in \mathbb{R}^{N,N}$
Hyperparamètre : matrice Laplacienne du graphe des labels, $\mathbf{L_m} \in \mathbb{R}^{m,m}$
Hyperparamètres : $\alpha > 0$, $\beta > 0$
1. Construire la matrice $\mathbf{K} \in \mathbb{R}^{N,N} : \forall i, j, \quad \mathbf{K}_{i,j} = \kappa(\mathcal{V}_{i,.}, \mathcal{V}_{j,.})$
2. Décomposer la matrice $\mathbf{K}$ en valeurs propres et vecteurs propres, sélectionner les s vecteurs propres ayant la plus grande valeur propre associée : $\mathbf{U} \in \mathbb{R}^{N,s}$
3. Poser $\mathbf{X} := \mathbf{KU}$
4. sélectionner la sous-matrice $\mathbf{X_l} \in \mathbb{R}^{n,s}$ de $\mathbf{X}$ correspondant aux individus labellisés
5. Appliquer l'algorithme de descente de gradient accélérée, en utilisant les paramètres suivants :
5.a. Pas d'apprentissage : $\frac{1}{C}$, $C = 2\left(\rho\left(\mathbf{X_l}'\mathbf{X_l} + \alpha\mathbf{X}'\mathbf{L_s}\mathbf{X}\right) + \beta\rho(\mathbf{L_m})\right)$
5.b. Modèle initial : $\mathbf{W} \in \mathbb{R}^{s,m} = \left[\mathbf{X_l}'\mathbf{X_l} + \epsilon\mathbf{I_s}\right]^{-1}\mathbf{X_l}'\mathbf{Y_l}$
5.c. Calcul du gradient : $\nabla_{\mathbf{W}} \leftarrow 2\left[\mathbf{X_l}'\mathbf{X_l} + \alpha\mathbf{X}'\mathbf{L_s}\mathbf{X}\right]\mathbf{W} - 2\mathbf{X_l}'\mathbf{Y_l} + 2\beta\mathbf{W}\mathbf{L_m}$
Résultat : $\mathbf{U} \in \mathbb{R}^{N,s}$
Résultat : $\mathbf{W} \in \mathbb{R}^{s,m}$

Algorithm 2 LSMR : Prédiction

Donnée : $\mathcal{V}_t \in \mathbb{R}^{N,d}$
Hyperparamètre : $\kappa \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
Modèle : $\mathbf{U} \in \mathbb{R}^{N,s}$
Modèle : $\mathbf{W} \in \mathbb{R}^{s,m}$
1. Construire la matrice $\mathbf{K_b} \in \mathbb{R}^{N,N} : \forall i, j, \quad \mathbf{K_b}_{i,j} = \kappa(\mathcal{V}_{i,.}, \mathcal{V}_{t\,j,.})$
2. Poser $\mathbf{X_t} := \mathbf{K_b}'\mathbf{U}$
Résultat : $\hat{\mathbf{Y}} \leftarrow \mathbf{X_t}\mathbf{W}$

4.1 Jeux de données

Nous avons utilisé des jeux de données du projet MULAN [2], décrits dans (Spyromitros-Xioufis et al., 2016), auxquels nous avons ajouté un échantillon de 1000 individus du jeu de données SARCOS (Vijayakumar et Schaal, 2000). Les caractéristiques de ces données sont résumées dans la table 1.

Nous avons divisé les jeux de données en une partie pour l'apprentissage et une partie pour le test. Ces jeux de données étant à la base totalement labellisés, nous avons sélectionné 30% des individus de l'ensemble d'apprentissage pour en supprimer la partie supervisée, pour tous les labels simultanément.

Nous avons ensuite normalisé toutes les variables et tous les labels. Ceci garantit que l'on peut appliquer à la fois une fonction noyau entre les individus à partir des variables, et à la fois une fonction noyau entre les labels à partir des données labellisées. D'autre part, si les valeurs de l'un des labels sont négligeables devant les valeurs d'un autre, le calcul des métriques multi-labels risque de ne pas montrer la pertinence des algorithmes en tant qu'algorithmes multi-labels.

4.2 Protocole expérimental

La procédure de tuning que nous avons adoptée consiste à tirer une valeur pour chaque hyperparamètre, selon la recherche aléatoire (Bergstra et Bengio, 2012). Cette méthode permet d'éviter l'écueil de la *recherche en grille*, qui considère tous les hyperparamètres comme équitablement importants. Dans le

2. `http://mulan.sourceforge.net/datasets-mtr.html`

TAB. 1: Jeux de données utilisés pour l'étude expérimentale de notre approche, **LSMR**

Jeu de données	Nb. indiv. (avec label + sans)	Test	Variables	Labels
atp1d	262 (76 + 186)	165	411	6
atp7d	234 (67 + 167)	147	411	6
edm	121 (35 + 86)	73	16	2
enb	601 (173 + 428)	366	8	2
jura	281 (81 + 200)	173	15	3
oes10	314 (91 + 223)	198	298	16
oes97	257 (75 + 182)	163	263	16
osales	495 (144 + 351)	309	401	12
sarcossub	779 (225 + 554)	467	21	7
scpf	889 (256 + 633)	521	23	3
sf1	250 (73 + 177)	158	31	3
sf2	832 (240 + 592)	501	31	3
wq	827 (238 + 589)	487	16	14

cas où l'algorithme présente de nombreux hyperparamètres, comme par exemple **LSMR**, la recherche aléatoire est à préférer.

Les hyperparamètres donnant la meilleure métrique aRMSE (la valeur la plus faible) en validation croisée à 10 *folds* sont retenus. Pour rappel, la métrique aRMSE est définie par :

$$\text{aRMSE} = \frac{1}{m}\sum_{k=1}^{m}\sqrt{\frac{\left\|\mathbf{Y}_{.,m} - \hat{\mathbf{Y}}_{.,m}\right\|_2^2}{n_t}} \tag{8}$$

$\hat{\mathbf{Y}}_{.,m}$ désigne la prédiction sur le jeu de test pour le label m, $\mathbf{Y}_{.,m}$ la vraie valeur du label m dans le jeu de test, et n_t le nombre d'individus du jeu de test.

La validation croisée en 10 *folds* consiste à partitionner les données labellisées de l'ensemble d'apprentissage en 10 échantillons. L'apprentissage se fait sur 9 échantillons, plus toutes les données non labellisées. Le test se fait sur l'échantillon restant, en calculant la métrique aRMSE. En répétant l'opération sur les 10 échantillons pour le test, la métrique aRMSE moyenne est calculée.

Notre approche utilise 4 hyperparamètres de différentes natures : la fonction noyau, κ, le nombre de composantes principales, s, le régulariseur semi-supervisé, α, et le régulariseur multi-labels, β. Nous étudions également différentes fonction de noyau : le produit scalaire, la similarité cosinus, le noyau gaussien (défini par $\kappa(x, y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$, avec un hyperparamètre σ).

4.3 Résultats avec tuning local

En tant qu'extension de l'algorithme *SSSL*, notre approche *SSSL* doit vérifier la pertinence des hyperparamètres introduits. Nous commençons par tuner le noyau et le nombre de composantes pour minimiser l'erreur de régression de l'algorithme *SSSL*. En réutilisant ces valeurs, nous cherchons celles pour les deux autres hyperparamètres. Nous comparons avec la performance de l'algorithme *SSSL*, en calculant l'aRMSE relative comme le rapport entre la performance de *LSMR* et celle de *SSSL*.

En agrégeant les performances relatives sur tous les points de tuning, on obtient la table 2. Celle-ci montre que beaucoup de points du tuning local donnent un meilleur résultat que *SSSL*. L'agrégation des scores relatifs se fait avec la moyenne, le premier et troisième quartile, le minimum et le maximum. Sur tous les jeux de données, il existe un point où *LSMR* est bien meilleur que *SSSL* (valeur inférieure à 1), et

un point où *LSMR* est pire. En regardant les quartiles, il existe de larges zones dans lesquelles *LSMR* tuné est meilleur que *SSSL*.

TAB. 2: Tuning local : aRMSE moyenne relative de LSMR par rapport à SSSL

Jeu de données	Moyenne	Q1	Q3	Meilleur point	Pire point
sf2	**0.983**	**0.898**	1.086	**0.553**	1.245
scpf	1.191	**0.862**	1.554	**0.564**	1.684
osales	**0.989**	**0.938**	1.046	**0.635**	1.191
oes97	1.078	**0.872**	1.170	**0.662**	2.344
sf1	**0.994**	**0.922**	1.062	**0.682**	1.206
oes10	1.015	**0.908**	1.110	**0.715**	1.448
jura	1.014	**0.954**	1.065	**0.799**	1.485
atp1d	1.048	**0.958**	1.070	**0.812**	1.922
edm	**0.998**	**0.964**	1.037	**0.821**	1.190
atp7d	1.006	**0.961**	1.053	**0.833**	1.242
enb	**0.998**	**0.964**	1.027	**0.864**	1.124
wq	**0.999**	**0.992**	1.007	**0.960**	1.032

4.4 Résultats avec tuning global

Nous mettons maintenant en place une approche de tuning global de façon à comparer notre approche proposée à d'autres approches de régularisation multi-labels pour la régression. Les méthodes évoquées résolvent un problème de régression multi-labels linéaire, en optimisant une fonction objectif convexe, au moyen de l'algorithme de descente de gradient accélérée. Ces méthodes sont une ligne de base sans régularisation (notée `LSQ` pour *least squares*), une régularisation l_1 (Tibshirani (1996b), notée `MTL` pour *multi-task learning*), la régularisation Laplacienne multi-labels seule (Zhou et al. (2012), notée `SGR` pour *sparse graph regularization*), la régularisation Lasso par groupe (Argyriou et al. (2007), notée `JFS` pour *joint feature selection*), la régularisation portant sur le rang (Argyriou et al. (2008), notée `TNR` (*trace norm regularization*), le *Dirty Model* (Jalali et al. (2010), noté `DM`), et l'algorithme *CMTL* (Zhou et al. (2011)).

La table 3 montre les résultats des différentes approches. Nous remarquons que :

— notre approche, pour un tuning global, est meilleure que *SSSL*, cependant le tuning local est insuffisant :
— notre approche est souvent meilleure que la régularisation multi-labels `SGR` ;
— le tuning global pour notre approche donne très souvent de meilleurs résultats que le tuning local ;
— contrairement au *SSSL*, notre approche donne des résultats comparables à l'état de l'art sur les jeux de données considérés, pour des approches de régularisation.

4.5 Validation statistique

Afin de savoir comment se positionnent les algorithmes les uns par rapport aux autres, nous effectuons un test statistique en suivant la méthodologie indiquée dans (Demšar, 2006), qui suit la batterie de tests suivants.

4.5.1 *t-test* par paires

En considérant une paire d'algorithmes, il est possible de calculer la différence entre les performances de ces deux algorithmes sur chaque jeu de données. Lorsque l'on prend la moyenne de cette différence,

TAB. 3: Tuning global, métrique aRMSE. Notre approche proposée est la colonne **LSMR**, la colonne **local** correspond à notre algorithme dont les hyperparamètres communs avec le SSSL ont les mêmes valeurs.

Jeu	**LSMR**	**local**	*SSSL*	*CMTL*	*DM*	*JFS*	*LSQ*	*MTL*	*SGR*	*TNR*
atp1d	**0.472**	1.006	0.481	0.533	0.522	0.543	0.549	0.534	0.548	0.548
atp7d	0.888	**0.713**	0.779	0.727	0.782	0.764	0.787	0.787	0.787	0.787
edm	**0.841**	1.005	0.895	0.849	0.857	0.856	1.198	0.855	1.207	0.852
enb	**0.320**	0.541	0.339	0.332	0.333	0.332	0.332	0.332	0.332	0.332
jura	0.661	0.847	0.727	0.597	0.593	**0.591**	0.609	0.593	0.611	0.598
oes10	**0.390**	0.488	0.395	0.398	0.402	0.402	0.402	0.402	0.402	0.402
oes97	**0.445**	0.918	0.494	0.515	0.510	0.534	0.534	0.534	0.534	0.534
osales	1.012	0.957	0.938	0.882	**0.839**	0.861	0.921	0.873	0.873	0.906
$\text{sarcos}_{\text{sub}}$	0.363	0.816	0.428	0.357	0.357	**0.354**	0.357	0.354	0.357	0.357
scpf	1.111	0.926	1.253	**0.621**	0.636	0.635	0.635	0.635	0.635	0.635
sf1	1.070	**0.998**	1.092	0.999	0.999	0.999	1.292	0.999	1.285	0.999
sf2	**0.973**	1.041	1.114	1.022	1.161	1.024	1.262	1.024	1.249	1.024
wq	0.999	0.999	1.003	**0.946**	1.006	0.982	1.079	0.947	1.079	1.023

à supposer qu'elle s'applique sur un grand nombre de jeux de données, on peut effectuer un test pour savoir si cette moyenne est positive, ou pour savoir si elle est négative. Cela permettrait de savoir si l'un des algorithmes est relativement meilleur que l'autre.

Comme nous n'avons que 13 jeux de données, il n'est pas possible de considérer les performances de chaque algorithme comme un échantillon gausssien, ce qui pose une limite assez claire à cette approche.

De plus, les jeux de données contenant très peu d'individus ont une performance très variable.

4.5.2 Test de Wilcoxon

Ce test permet aussi de départager deux algorithmes, mais il ne se fonde plus sur les valeurs de la métrique envisagée mais sur les rangs des algorithmes. Plus précisément, on calcule les différences entre les deux algorithmes sur chaque jeu de données en valeur absolue, puis on trie ces différences absolues. Pour chaque algorithme, on sélectionne les différences qui sont en faveur de cet algorithme, et on somme leurs rangs. En prenant une valeur de $\alpha = 0.1$, on obtient la figure 1. Notre généralisation est meilleure que l'algorithme *SSSL*.

4.5.3 Tests de Friedman et Nemenyi

Le test de Friedman permet de rejeter l'hypothèse suivante en se basant sur les rangs des approches : tous les régresseurs sont équivalents, c'est-à-dire que leurs rangs moyens sont égaux. Cette hypothèse est rejetée dans notre cas pour un risque $\alpha = 0.05$.

Puisqu'il y a des différences de rangs moyens entre les algorithmes, le test de Nemenyi est pertinent. Il permet d'établir la *distance critique* entre ces rangs. Si des algorithmes ont une différence entre leurs rangs moyens supérieure à cette distance critique, ils ne sont pas équivalents (figure 2). On constate que notre approche, même si elle n'obtient pas le meilleur rang moyen, est dans le groupe des meilleurs algorithmes, contrairement aux deux approches sur lesquelles elle est fondée.

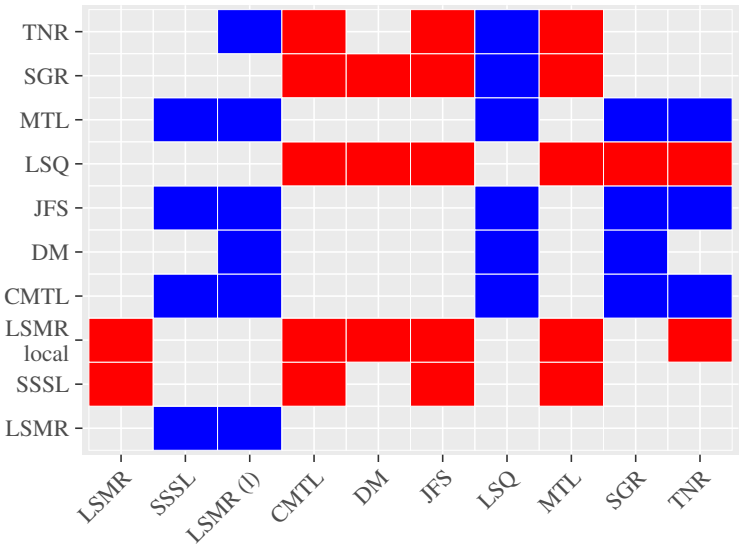


FIG. 1: Résultat du test de comparaison par paires d'algorithmes de Wilcoxon, avec un risque $\alpha = 0.1$. Une cellule bleue indique que l'algorithme en ligne bat l'algorithme en colonne. Une cellule rouge indique que l'algorithme en colonne bat l'algorithme en ligne.

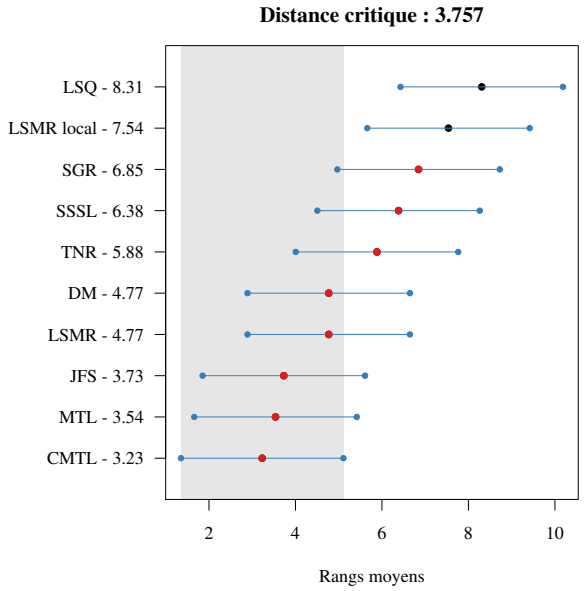


FIG. 2: Résultat du test de comparaison par paires d'algorithmes de Nemenyi, avec un risque $\alpha = 0.05$.

5 Conclusion

Nous avons proposé une extension de l'algorithme *SSSL* adapté à l'apprentissage multi-labels, en combinant une régularisation Laplacienne portant sur le graphe des individus pour l'apprentissage semi-supervisé, d'une part, et une régularisation multi-labels portant sur le graphe des labels, d'autre part. Expérimentalement, cette approche donne une erreur de régression multi-labels plus faible que *SSSL*, même en réutilisant une partie du tuning des hyperparamètres. En optimisant les hyperparamètres de façon globale, l'approche demeure compétitive avec des méthodes représentatives de l'état de l'art.

La suite des travaux vise à étudier d'autres formes de régularisation multi-labels, considérant les mauvaises performances de la régularisation Laplacienne multi-labels seule sur les jeux de données étudiés.

Références

Argyriou, A., T. Evgeniou, et M. Pontil (2007). Multi-Task Feature Learning. *Advances in neural information processing systems*, 41–48.

Argyriou, A., T. Evgeniou, et M. Pontil (2008). Convex Multi-Task Feature Learning. *Machine learning 73*(3), 243–272.

Azriel, D., L. D. Brown, M. Sklar, R. Berk, A. Buja, et L. Zhao (2016). Semi-Supervised linear regression. *arXiv preprint arXiv :1612.02391*.

Basu, S., I. Davidson, et K. Wagstaff (2008). Constrained Clustering : Advances in Algorithms, Theory, and Applications.

Belkin, M., P. Niyogi, et V. Sindhwani (2006). Manifold regularization : A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research 7*(Nov), 2399–2434.

Bennett, K. P. et A. Demiriz (1999). Semi-supervised support vector machines. In *Advances in Neural Information processing systems*, pp. 368–374.

Bergstra, J. et Y. Bengio (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13*(10), 281–305.

Blum, A. et T. Mitchell (1998). Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, New York, NY, USA, pp. 92–100. ACM.

Borchani, H., G. Varando, C. Bielza, et P. Larrañaga (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery 5*(5), 216–233.

Cai, D., X. He, et J. Han (2006). Semi-supervised regression using spectral techniques. Technical report.

Chapelle, O., B. Scholkopf, et A. Zien (Eds.) (2006). *Semi-Supervised Learning*. The MIT Press.

Chen, Y.-n. et H.-t. Lin (2012). Feature-aware Label Space Dimension Reduction for Multi-label Classification. In F. Pereira, C. J. C. Burges, L. Bottou, et K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*, pp. 1529–1537. Curran Associates, Inc.

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of machine learning research 7*, 1–30.

Gasse, M., A. Aussem, et H. Elghazel (2015). On the Optimality of Multi-Label Classification under Subset Zero-One Loss for Distributions Satisfying the Composition Property. In F. R. Bach et D. M. Blei (Eds.), *International Conference on Machine Learning*, Volume 37 of *Journal of Machine Learning Reasearch Proceedings*, Lille, France, pp. 2531–2539.

Gong, P., J. Ye, et C.-s. Zhang (2012). Multi-stage multi-task feature learning. In *Advances in neural information processing systems*, pp. 1988–1996.

Jalali, A., S. Sanghavi, C. Ruan, et P. K. Ravikumar (2010). A Dirty Model for Multi-task Learning. *Advances in neural information processing systems 23*, 9.

Ji, M., T. Yang, B. Lin, R. Jin, et J. Han (2012). A simple algorithm for semi-supervised learning with improved generalization error bound. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 835–842.

Jian, L., J. Li, K. Shu, et H. Liu (2016). Multi-Label Informed Feature Selection. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1627–1633.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *ICML*, Volume 99, pp. 200–209.

Li, Y.-F., I. W. Tsang, J. T. Kwok, et Z.-H. Zhou (2013). Convex and scalable weakly labeled svms. *The Journal of Machine Learning Research 14*(1), 2151–2188.

Liu, W., I. W. Tsang, et K.-R. Müller (2017). An easy-to-hard learning paradigm for multiple classes and multiple labels. *The Journal of Machine Learning Research 18*(1), 3300–3337.

Luo, M., L. Zhang, F. Nie, X. Chang, B. Qian, et Q. Zheng (2017). Adaptive Semi-Supervised Learning with Discriminative Least Squares Regression. pp. 2421–2427. International Joint Conferences on Artificial Intelligence Organization.

Moscovich, A., A. Jaffe, et B. Nadler (2016). Minimax-optimal semi-supervised regression on unknown manifolds : supplementary material.

Nesterov, Y. (2007). Gradient methods for minimizing composite functions. *Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers 140*.

Nigam, K., A. K. McCallum, S. Thrun, et T. Mitchell (2000). Text classification from labeled and unlabeled documents using em. *Machine learning 39*(2), 103–134.

Read, J., B. Pfahringer, G. Holmes, et E. Frank (2011). Classifier chains for multi-label classification. *Machine learning 85*(3), 333.

Ryan, K. J. et M. V. Culp (2015). On semi-supervised linear regression in covariate shift problems. *Journal of Machine Learning Research 16*, 3183–3217.

Scudder, H. (1965). Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory 11*(3), 363–371.

Shi, Y., J. Miao, Z. Wang, P. Zhang, et L. Niu (2018). Feature Selection With $\ell_{2,1-2}$ Regularization. *IEEE Trans. Neural Netw. Learning Syst. 29*(10), 4967–4982.

Spyromitros-Xioufis, E., G. Tsoumakas, W. Groves, et I. Vlahavas (2016). Multi-Target Regression via Input Space Expansion : Treating Targets as Inputs. *Machine Learning 104*(1), 55–98. arXiv : 1211.6581.

Tibshirani, R. (1996a). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological) 58*(1), 267–288.

Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological) 58*(1), 267–288.

Vijayakumar, S. et S. Schaal (2000). Locally weighted projection regression : An o(n) algorithm for incremental real time learning in high dimensional spaces. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, Volume 1, Stanford, CA, pp. 288–293. clmc.

Yu, K., V. Tresp, et A. Schwaighofer (2005). Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd international conference on Machine learning*, pp. 1012–1019. ACM.

Zhang, Z., T. W. Chow, et M. Zhao (2012). Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization. *IEEE Transactions on Knowledge and Data Engineering 25*(5), 1148–1161.

Zhao, M., T. W. Chow, Z. Wu, Z. Zhang, et B. Li (2015). Learning from normalized local and global discriminative information for semi-supervised regression and dimensionality reduction. *Information Sciences 324*, 286–309.

Zhou, J., J. Chen, et J. Ye (2011). Clustered Multi-Task Learning Via Alternating Structure Optimization. *Advances in neural information processing systems 24*, 9.

Zhou, J., J. Chen, et J. Ye (2012). Malsar : Multi-task learning via structural regularization.

Zhou, Z.-H. et M. Li (2005). Semi-Supervised Regression with Co-Training. In *Proceedings of the 19th international joint conference on Artificial intelligence*, Volume 5, pp. 908–913.

Zhu, X., Z. Ghahramani, et J. D. Lafferty (2003). Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919.

Summary

In multi-label learning, one has to find a model suitable for predicting multiple values for the same individual, based on the same features. The effectiveness of most multi- label algorithms lies in the fact that it is able to consider the correlations between the related labels. On the other hand, in many applications, a multi-label learning task incurs a high cost for the annotation of a single data point. This leads to a dataset consisting of a few labeled data points, and many more unlabeled data points. In this scenario, semi-supervised methods can take advantage of the unlabeled data points. In this article, we propose a new algorithm for multi-label semi- supervised regression, LSMR, as a multi-label extension of a semi-supervised regression algorithm. We provide experimental results on some publicly-available regression datasets showing the effectiveness of our approach.

Keywords: Multi-label learning, Semi-supervised learning, Regression, Laplacian regularization.

Apprentissage semi-supervisé transductif basé sur le transport optimal

Mourad El Hamri* **, Younès Bennani* **, Issam Falih*** **

* LIPN UMR 7030 CNRS, Université Sorbonne Paris Nord, France
name.surname@sorbonne-paris-nord.fr
**LaMSN, La Maison des Sciences Numériques, France
name.surname@lamsn.sorbonne-paris-nord.fr
***LIMOS-UMR 6158 CNRS, Université Clermont-Auvergne, France
name.surname@uca.fr

Résumé. Dans cet article, nous abordons le problème de l'apprentissage semi-supervisé transductif qui vise à prédire les labels des données non étiquetées selon le principe de Vapnik. L'approche que nous proposons est basée sur le transport optimal, une théorie mathématique qui a été utilisée avec succès pour résoudre divers problèmes d'apprentissage automatique, et qui commence à susciter un regain d'intérêt dans la communauté de l'apprentissage semi-supervisé. L'approche proposée, Optimal Transport Propagation (OTP), effectue dans un processus incrémental, la propagation des labels à travers les arêtes d'un graphe biparti pondéré, dont la matrice d'affinité est construite à partir du plan de transport optimal entre les mesures empiriques définies sur les données étiquetées et non étiquetées. OTP assure une haute qualité des prédictions en contrôlant le processus de propagation à l'aide d'un score de certitude basé sur l'entropie de Shannon. Les expérimentations montrent la supériorité de l'approche proposée par rapport à l'état de l'art. Nous mettons notre code à la disposition de la communauté. [1]
Mots-clés : Transport Optimal, Apprentissage semi-supervisé, Propagation de labels

1 Introduction

Les modèles d'apprentissage profond ont atteint des performances de pointe sur un large éventail de tâches d'apprentissage et deviennent de plus en plus populaires dans divers domaines d'application où des grandes quantités de données étiquetées sont disponibles, tels que la classification d'images et la reconnaissance vocale. Cependant, pour de nombreuses tâches, le coût de la collecte d'un grand jeu de données étiquetées de haute qualité est souvent prohibitif en raison du manque de temps, de ressources ou d'autres facteurs, alors que les données non étiquetées sont abondantes. L'apprentissage semi-supervisé (SSL) constitue une approche attrayante pour remédier au manque de massifs jeux de données étiquetées. Il cherche à alléger

1. Le code est diponible sur : `https://github.com/MouradElHamri/OTP`

largement le besoin de données labellisées en fournissant un moyen d'exploiter conjointement les instances non labellisées. Les approches basées sur les graphes constituent une des classes d'apprentissage semi-supervisé les plus utilisées, en raison de leurs performances et de la multiplication des jeux de données représentés par des graphes. La propagation de labels est une méthode populaire de ce genre d'approches qui a montré de bonnes performances dans différentes applications d'apprentissage au cours des dernières années, comme l'analyse des réseaux sociaux (Zhang et al. (2017)) et le traitement du langage naturel (Barba et al. (2020)).

La plupart des algorithmes de propagation de labels existants infèrent les étiquettes sur un graphe entièrement connecté construit en reliant les échantillons similaires. Le graphe entièrement connecté conduit généralement à un étiquetage des données non labellisées et un réétiquetage des données déjà labellisées, ce qui peut être intéressant sous l'hypothèse de labels bruités, sinon, il est nécessaire d'ajouter un terme de régularisation à la fonction objectif correspondante afin de pénaliser les étiquettes prédites qui ne correspondent pas aux vraies étiquettes (Van Engelen et Hoos (2020)). Les principales approches de propagation de labels peuvent être divisées en deux catégories : les méthodes de la première catégorie comme (Zhu et Ghahramani (2002), Zhou et al. (2003)), capturent l'information à un niveau bilatéral, et les méthodes de la seconde catégorie comme (Wang et Zhang (2007)), capturent l'information à un niveau local. Les approches de la première catégorie utilisent un noyau gaussien avec un paramètre libre σ pour calculer les relations par paires entre les données, ce qui présente certains inconvénients, car il est difficile de déterminer la valeur optimale de σ si l'on ne dispose que de très peu d'instances étiquetées (Zhou et al. (2003)), et l'étiquetage est très sensible au paramètre σ (Wang et Zhang (2007)). Au lieu des relations par paires qui ne prennent en compte que les relations bilatérales entre les instances, les approches de la deuxième catégorie utilisent une information de voisinage local qui suppose que chaque donnée peut être reconstruite de manière optimale sous forme d'une combinaison linéaire de ses voisins, ce qui présente également de nombreux inconvénients, puisque le nombre optimal d'instances constituant le voisinage linéaire doit être déterminé à l'avance, et même une petite variation de sa valeur pourrait rendre les résultats de l'étiquetage très différents, sans oublier que l'hypothèse de linéarité est principalement destinée pour faciliter les calculs (Wang et Zhang (2007)). En outre, les approches des deux catégories sont incapables de capturer la géométrie sous-jacente de l'espace d'entrée en entier ainsi que les différentes interactions qui peuvent se produire entre les données étiquetées et non étiquetées à un niveau global, et elles ont un autre inconvénient majeur, celui d'inférer simultanément tous les pseudo-labels par affectation dure (hard assignment), tout en négligeant le degré de certitude différent de chaque prédiction. Une approche efficace de propagation de labels capable de traiter tous ces points n'a pas encore été rapportée.

L'un des paradigmes utilisé pour saisir la géométrie sous-jacente des données repose sur la théorie du transport optimal (Villani (2008)). Le transport optimal fournit un moyen mathématique puissant doté de nombreuses propriétés théoriques attrayantes pour comparer les mesures de probabilité dans un cadre lagrangien, ce qui fait que de nombreux domaines de l'apprentissage automatique s'appuient sur lui pour modéliser des tâches, calculer des solutions et fournir une analyse théorique des algorithmes, comme l'adaptation du domaine (Courty et al. (2016), Redko et al. (2017)), le clustering (Laclau et al. (2017), Ben Bouazza et al. (2019), Ben Bouazza et al. (2020)) et plus récemment l'apprentissage semi-supervisé (Taherkhani et al.

(2020)).

Ce papier est organisé comme suit : dans la section 2, nous présentons un aperçu de l'apprentissage semi-supervisé transductif. La section 3 détaille le problème du transport optimal et sa version régularisée. Dans la section 4, nous présentons l'approche OTP proposée. Dans la section 5, nous fournissons des comparaisons avec les méthodes de l'état de l'art sur six jeux de données de référence.

2 L'apprentissage semi-supervisé transductif

En apprentissage automatique, une distinction a généralement été faite entre deux tâches principales : l'apprentissage supervisé (SL) et l'apprentissage non-supervisé (UL). Conceptuellement, l'apprentissage semi-supervisé (SSL) (Zhu (2005)) se situe entre les deux. Son objectif est d'utiliser la quantité abondante d'échantillons non étiquetés, ainsi qu'un ensemble généralement plus petit d'instances étiquetées, pour améliorer la performance qui peut être obtenue soit en rejetant les données non étiquetées et en effectuant la classification (SL), soit en rejetant les étiquettes disponibles et en effectuant le clustering (UL).

En fonction de son objectif, l'apprentissage semi-supervisé peut être classé en deux sous-paradigmes : l'apprentissage semi-supervisé transductif et inductif (Van Engelen et Hoos (2020)). L'apprentissage semi-supervisé transductif s'intéresse exclusivement à l'obtention des labels pour les données non étiquetées. Cependant, l'apprentissage semi-supervisé inductif cherche à inférer un bon classifieur qui peut estimer efficacement l'étiquette pour n'importe quelle instance dans l'espace d'entrée, même pour des données non-vues précédemment.

L'objectif de l'apprentissage semi-supervisé transductif est en fait par essence une illustration parfaite du principe de Vapnik : lorsqu'on essaie de résoudre un problème, on ne devrait pas résoudre un problème plus difficile comme étape intermédiaire. Ainsi, au lieu d'inférer un classifieur sur l'espace d'entrée et de l'évaluer sur les points non étiquetés, le principe de Vapnik suggère naturellement de propager l'information à travers des connexions directes entre les données, ce qui peut être réalisé en utilisant une méthode basée sur les graphes, à savoir la propagation de labels. Les approches de propagation de labels comportent généralement deux phases : une phase de construction du graphe où chaque instance est représentée par un sommet, les sommets similaires sont ensuite reliés entre eux par des arêtes, puis les arêtes sont pondérées pour indiquer le degré de similarité entre les sommets. Les poids des arêtes sont regroupés dans une matrice d'affinité. La deuxième phase est la propagation de labels, où le graphe déjà construit dans la première phase est utilisé pour diffuser les étiquettes à travers ses arrêts, des sommets étiquetés aux sommets non étiquetés.

3 Cadre formel

Le transport optimal (Villani (2008)) est la branche des mathématiques qui cherche à transformer une mesure de probabilité en une autre tout en minimisant le coût total du transport. On doit la première formulation du problème du transport optimal au mathématicien français Gaspard Monge (Monge (1781)) : soit $(\mathcal{X}, \mu)$ et $(\mathcal{Y}, \nu)$ deux espaces de probabilité, c une

fonction de coût positive sur $\mathcal{X} \times \mathcal{Y}$, qui représente le travail nécessaire pour transporter une unité de masse de $x \in \mathcal{X}$ à $y \in \mathcal{Y}$. Le problème cherche à trouver une application de transport mesurable $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ telle que :

$$(\mathcal{M}) \quad \inf_{\mathcal{T}} \{ \int_{\mathcal{X}} c(x, \mathcal{T}(x)) d\mu(x) | \mathcal{T}\#\mu = \nu \}, \tag{1}$$

où $\mathcal{T}\#\mu$ représente la mesure image de μ par $\mathcal{T}$. Le problème $(\mathcal{M})$ n'est pas symétrique, et peut ne pas admettre de solution, c'est le cas lorsque μ est une masse de Dirac et ν ne l'est pas.

Une relaxation convexe du problème original a été suggérée par le mathématicien et économiste soviétique Leonid Kantorovitch (Kantorovich (1942)), cette formulation permet le fractionnement de masse et garantit une solution sous des hypothèses très générales.

$$(\mathcal{MK}) \quad \inf_{\gamma} \{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\gamma(x,y) \,|\, \gamma \in \Pi(\mu,\nu) \}, \tag{2}$$

où $\Pi(\mu, \nu)$ est l'ensemble des plans de transport, constitué de toutes les mesures de probabilité conjointes γ sur $\mathcal{X} \times \mathcal{Y}$ ayant comme marginales μ et ν : $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) | \pi_1 \# \gamma = \mu$ et $\pi_2 \# \gamma = \nu\}$. π_1 et π_2 représentent les applications de projection :

$$\begin{array}{ccc} \pi_1 \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{X} & \text{et} & \pi_2 \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}. \\ (x, y) \mapsto x & & (x, y) \mapsto y \end{array}$$

Lorsque $\mathcal{X} = \mathcal{Y}$ est un espace métrique doté d'une distance d, il est naturel de l'utiliser comme fonction de coût, par exemple $c(x, y) = d(x, y)^p$ pour $p \in [1\,, +\infty[$. Dans ce cas, le problème $(\mathcal{MK})$ définit une métrique entre les mesures de probabilité sur $\mathcal{X}$, appelée la p-ième distance de Wasserstein, définie comme suit, $\forall \mu, \nu \in \mathcal{P}(\mathcal{X})$:

$$\mathrm{W}_p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} (\int_{\mathcal{X}^2} d^p(x, y)\, d\gamma(x, y))^{1/p}, \tag{3}$$

La distance de Wasserstein a une formulation intuitive ainsi que la capacité de capturer la géométrie sous-jacente des mesures en s'appuyant sur la métrique d. Elle métrise la convergence faible et permet de comparer des mesures de probabilité, même lorsque leurs supports ne se chevauchent pas. Ces propriétés en font un candidat idéal pour les problèmes d'apprentissage.

Dans la version discrète du problème de transport optimal, lorsque μ et ν ne sont disponibles qu'à travers des échantillons discrets $X = (x_1, ..., x_n) \subset \mathcal{X}$ et $Y = (y_1, ..., y_m) \subset \mathcal{Y}$, les distributions empiriques peuvent être considérées comme étant des mesures discrètes : $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ et $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, où $a = (a_1, ..., a_n) \in \sum_n$ et $b = (b_1, \ldots, b_m) \in \sum_m$. La fonction de coût ne doit être spécifiée que pour chaque paire $(x_i, y_j) \in X \times Y$ induisant une matrice de coûts $C \in \mathcal{M}_{n\times m}(\mathbb{R}^+)$. Le problème de transport optimal devient alors un programme linéaire, paramétré par la matrice de coûts C et le polytope de transport $U(a, b) = \{\gamma \in \mathcal{M}_{n\times m}(\mathbb{R}^+) \,|\, \gamma 1_m = a$ et $\gamma^{\mathbf{T}} 1_n = b\}$, qui correspond à l'ensemble réalisable. Ainsi, la résolution de ce programme linéaire consiste à trouver un plan γ^* qui réalise :

$$(\mathcal{D}_{\mathcal{MK}}) \quad \min_{\gamma \in U(a,b)} \langle \gamma, C \rangle_F, \tag{4}$$

où $\langle ., . \rangle_F$ est le produit scalaire de Frobenius.

Le transport optimal discret est un programme linéaire, et peut donc être résolu exactement en $\mathcal{O}(n^3 log(n))$ lors de la comparaison de deux mesures discrètes de n points avec des méthodes de points intérieurs, ce qui représente une grande complexité calculatoire. Dans (Cuturi (2013)), l'auteur a proposé d'ajouter une régularisation entropique à l'expression de $(\mathcal{D}_{\mathcal{MK}})$ qui permet un calcul très rapide du plan de transport. La version régularisée est la suivante :

$$\min_{\gamma \in U(a,b)} \langle \gamma, C \rangle_F - \varepsilon \mathcal{H}(\gamma), \tag{5}$$

où $\mathcal{H}(\gamma) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij}(\log(\gamma_{ij}) - 1)$ est l'entropie de γ.

Ce problème peut être résolu efficacement par une procédure itérative : l'algorithme de Sinkhorn-Knopp qui met à jour itérativement $u^{(l+1)} = \frac{a}{Kv^{(l)}}$, et $v^{(l+1)} = \frac{b}{K^{\mathbf{T}} u^{(l+1)}}$, initialisé avec un vecteur arbitraire positif $v^{(0)} = 1_m$..

4 Approche proposée : Optimal Transport Propagation

Dans cette section, nous présentons l'approche OTP (Optimal Transport Propagation) proposée. L'idée principale sous-jacente de OTP est d'utiliser le plan de transport optimal entre les mesures empiriques définies sur les instances étiquetées et non étiquetées afin de construire une matrice d'affinité améliorée, puis de l'utiliser pour propager les étiquettes des données étiquetées vers les données non étiquetées dans un processus incrémental garantissant la certitude des prédictions.

4.1 Formulation de l'approche

Soient $X = \{x_1, ..., x_{l+u}\}$ un ensemble de $l + u$ points dans l'espace d'entrée $\mathbb{R}^d$ et = $\mathcal{C} = \{c_1, ..., c_K\}$ un ensemble d'étiquettes discrètes composé de K classes. Les l premiers points, désignés par $X_L = \{x_1, \ldots, x_l\}$ sont étiquetés selon $Y_L = \{y_1, ..., y_l\}$, où $y_i \in \mathcal{C}$ pour chaque $i \in \{1, ..., l\}$, et les échantillons restants, désignés par $X_U = \{x_{l+1}, ..., x_{l+u}\}$, ne sont pas étiquetés. Habituellement, $l \ll u$. Le but des algorithmes de propagation de labels est d'inférer les étiquettes inconnues Y_U en utilisant toutes les données dans $X = X_L \cup X_U$ et les étiquettes Y_L.

Pour utiliser une formulation appropriée au paradigme du transport optimal, la distribution empirique de X_L et X_U doivent être exprimées respectivement à l'aide de mesures discrètes comme suit :

$$\mu = \sum_{i=1}^{l} a_i \delta_{x_i} \quad \text{et} \quad \nu = \sum_{j=l+1}^{l+u} b_j \delta_{x_j}, \tag{6}$$

Si l'on suppose que X_L et X_U sont une collection de points indépendants et identiquement distribués, les poids de toutes les instances de chaque échantillon sont naturellement égaux :

$$a_i = \frac{1}{l}, \forall i \in \{1, ..., l\} \text{ et } b_j = \frac{1}{u}, \forall j \in \{l+1, ..., l+u\}, \tag{7}$$

Les algorithmes de propagation de labels sont souvent des méthodes basées sur les graphes, qui se composent généralement de deux phases : la première est la construction du graphe et la seconde est la phase de propagation de labels, où nous diffusons les labels des sommets étiquetés du graphe déjà construit dans la première phase, vers les sommets non étiquetés.

4.2 Construction du graphe

L'idée principale de la première phase est l'utilisation d'un graphe complet biparti pondéré $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, où $\mathcal{V} = X$ est l'ensemble des sommets, qui peut être divisé en deux ensembles disjoints et indépendants $\mathcal{L} = X_L$ et $\mathcal{U} = X_U$, $\mathcal{E} \subset \{\mathcal{L} \times \mathcal{U}\}$ est l'ensemble des arêtes, et $\mathcal{W} \in \mathcal{M}_{l,u}(\mathbb{R}^+)$ est la matrice d'affinité pour désigner les poids des arêtes. Le poids $w_{i,j}$ sur l'arête $e_{i,j} \in \mathcal{E}$ reflète le degré de similarité entre $x_i \in \mathcal{L}$ et $x_j \in \mathcal{U}$. Nous proposons d'inférer la matrice d'affinité à partir du plan de transport optimal entre les mesures μ et ν. L'intuition derrière l'utilisation de ce type de graphe peu commun dans l'apprentissage semi-supervisé, est d'exploiter sa capacité d'attribuer les étiquettes uniquement pour les instances dans $\mathcal{U}$, sans modifier les étiquettes des échantillons dans $\mathcal{L}$, ce qui le rend très attrayant pour les tâches de propagation de labels en dehors de l'hypothèse de labels bruités, c'est-à-dire qu'il n'est pas nécessaire d'ajouter un terme de régularisation à la fonction objectif de l'algorithme afin de pénaliser les étiquettes prédites qui ne correspondent pas aux vraies étiquettes dans $\mathcal{L}$.

Pour mesurer quantitativement la similarité entre les sommets, nous devons utiliser une certaine distance sur l'espace d'entrée $\mathbb{R}^d$. Soit $C \in \mathcal{M}_{l,u}(\mathbb{R}^+)$ la matrice des distances euclidiennes au carré entre les sommets dans $\mathcal{L}$ et $\mathcal{U}$, définie comme suit :

$$c_{i,j} = \|x_i - x_j\|^2, \ \ \forall (x_i, x_j) \in \mathcal{L} \times \mathcal{U}, \tag{8}$$

Afin de construire une matrice d'affinité $\mathcal{W}$ qui capture la géométrie sous-jacente de la totalité des données X dans l'espace d'entrée et toutes les interactions entre les données étiquetées et non étiquetées dans une vision globale, au lieu des relations par paire qui performent à un niveau bilatéral ou de l'information de voisinage local, et pour éviter l'utilisation du noyau gaussien, un choix naturel est de s'appuyer sur la théorie du transport optimal. Puisque le transport optimal souffre d'une grande complexité calculatoire, nous pouvons surmonter ce problème en utilisant sa version régularisée, de la manière suivante :

$$\gamma_\varepsilon^* = \underset{\gamma \in U(a,b)}{argmin} \ \langle \gamma, C \rangle_F - \varepsilon \mathcal{H}(\gamma), \tag{9}$$

Le plan de transport optimal γ_ε^* nous fournit les poids des associations entre les sommets de $\mathcal{L}$ et $\mathcal{U}$, ainsi, γ_ε^* peut être interprétée dans notre contexte de propagation de labels comme une matrice de similarité entre les deux parts $\mathcal{L}$ et $\mathcal{U}$ du graphe $\mathcal{G}$: les sommets étiquetés et non étiquetés similaires correspondent à des valeurs élevées dans γ_ε^*.

Pour avoir une interprétation probabiliste, on normalise les colonnes de la matrice γ_ε^*, cela donnera une matrice d'affinité rectangulaire stochastique à gauche (left stochastic matrix) $\mathcal{W}$, définie comme suit :

$$w_{i,j} = \frac{\gamma^*_{\varepsilon_{i,j}}}{\sum_i \gamma^*_{\varepsilon_{i,j}}}, \ \ \forall i, j \in \{1, ..., l\} \times \{l+1, ..., l+u\}, \tag{10}$$

où $w_{i,j},\ \forall i,j \in \{1,...,l\} \times \{l+1,...,l+u\}$ est alors, la probabilité de sauter du sommet $x_i \in \mathcal{L}$ au sommet $x_j \in \mathcal{U}$.

4.3 Algorithme de propagation de labels

Notre intuition est d'utiliser la matrice d'affinité $\mathcal{W}$ dans la deuxième phase pour identifier les données étiquetées qui devraient propager leurs étiquettes à des instances non étiquetées similaires. Nous suggérons d'utiliser un processus incrémental pour étiqueter les points dans $\mathcal{U}$. Nous suggérons également de fournir avec chaque pseudo-label un score de certitude qui mesure la certaineté des prédictions, et de l'utiliser afin de contrôler le processus de propagation de labels incrémental.

Tout d'abord, nous devons construire une matrice d'étiquetage $U \in \mathcal{M}_{u,K}(\mathbb{R}^+)$ afin de désigner la probabilité de chaque donnée non étiqueté x_j, $j \in \{l+1,...,l+u\}$ d'appartenir à une classe c_k, $k \in \{1,...,K\}$. Pour une construction harmonieuse de la matrice d'étiquetage U avec l'information provenante du plan de transport optimal γ_ε^*, nous proposons de définir la probabilité d'appartenance d'une donnée non étiquetée x_j à une classe c_k comme étant la somme de ses similarités avec les représentants de cette classe :

$$u_{j,k} = \sum_{i/x_i \in c_k} w_{i,j}, \forall j,k \in \{l+1,...,l+u\} \times \{1,...,K\}, \tag{11}$$

La matrice U est une matrice rectangulaire stochastique à droite (right stochastic matrix) , et peut être interprétée comme une fonction à valeur vectorielle $U : X_U \to \sum_K$, qui attribue un vecteur stochastique $U_j \in \sum_K$ à chaque donnée non étiquetée x_j, $j \in \{l+1,...,l+u\}$.

Les approches de propagation de labels traditionnelles infèrent simultanément tous les pseudo-étiquettes par affectation dure, sans se soucier du fait que ces prédictions d'étiquettes n'ont pas le même degré de certitude. Ce problème, comme mentionné par (Iscen et al. (2019)), peut dégrader significativement la performance des approches de propagation de labels. Pour éviter cela, nous proposons d'associer un score de certitude s_j à la prédiction d'étiquette de chaque $x_j, j \in \{l+1,...,l+u\}$. Le score de certitude proposé s_j est défini de la manière suivante :

$$s_j = 1 - \frac{H(Z_j)}{log_2(K)}, \quad \forall j \in \{l+1,...,l+u\}, \tag{12}$$

où $Z_j : \mathcal{C} \to \mathbb{R}$ est une variable aléatoire à valeur réelle, définie par $Z_j(c_k) = k$, pour associer une valeur numérique k au pseudo-label potentiel c_k. La distribution de probabilité de la variable aléatoire Z_j est codée dans le vecteur stochastique U_j :

$$\mathbb{P}(Z_j = c_k) = u_{j,k}, \quad \forall j,k \in \{l+1,...,l+u\} \times \{1,...,K\}$$

H est l'entropie de Shannon, définie par : $H(Z_j) = -\sum_k u_{j,k} \log_2(u_{j,k})$,
nous divisons H par $log_2(K)$ pour la normaliser entre 0 et 1.

Pour contrôler la certitude du processus de propagation, nous définissons un seuil de confiance $\alpha \in [0,1]$, et pour chaque donnée non étiquetée x_j, nous effectuons une comparaison entre α

et s_j. Si le score s_j est supérieur à α, nous attribuons à x_j un pseudo-label $\hat{y_j}$, de la manière suivante :

$$\hat{y_j} = \underset{c_k \in \mathcal{C}}{argmax}\, u_{j,k},\ \forall j \in \{l+1, ..., l+u\}, \tag{13}$$

Ainsi, l'instance non étiquetée x_j appartiendra à la classe c_k avec la probabilité de classe la plus élevée $u_{j,k}$, en d'autres termes, à la classe dont les représentants possèdent la plus grande similarité avec lui. Sinon, nous ne donnons aucune étiquette au point x_j.

Le processus ci-dessus correspond à une itération de l'approche incrémentale proposée. A chacune de ses itérations, X_L est enrichi avec de nouvelles instances, et le nombre de données dans X_U est réduit. Cette modification de X_L, Y_L et X_U résultant de chaque itération de l'approche incrémentale est d'une importance majeure dans le contexte de propagation de labels, puisque l'efficacité de ce genre d'algorithme dépend de la quantité d'information préalable, ainsi, l'augmentation de la taille de X_L à chaque itération, augmentera de façon similaire la performance de l'approche proposée, et permettra d'étiqueter les points encore dans X_U avec un haut degré de certitude aux itérations suivantes. Nous répétons la même procédure à chaque itération jusqu'à convergence, la convergence signifie ici que toutes les données initialement dans X_U seront étiquetées au cours de ce processus incrémental. L'algorithme proposé, nommé OTP, est formellement résumé en Algorithme 1 :

Algorithme 1 : OTP

Parameters : ε, α
Input : X_L, X_U, Y_L
while *not converged* **do**
  Calculer la matrice de coût C par Eq(8)
  Résoudre le problème du transport optimal dans Eq(9)
  Calculer la matrice d'affinité $\mathcal{W}$ par Eq(10)
  Obtenir la matrice d'étiquetage U par Eq(11)
  for $x_j \in X_U$ **do**
    Calculer le score de certitude s_j par Eq(12)
    if $s_j > \alpha$ **then**
      Obtenir le pseudo-label $\hat{y_j}$ par Eq(13)
      Injecter x_j dans X_L
      Injecter $\hat{y_j}$ dans Y_L
    else
      Maintenir x_j dans X_U
    end
  end
end
return Y_U

5 Validation expérimentale

5.1 Protocole

Les expériences ont été conçue pour évaluer l'approche proposée sur six jeux de données de référence. [2] Afin d'évaluer la performance de notre approche, deux mesures d'évaluation ont été utilisées : l'information mutuelle normalisée (NMI), et l'indice de Rand ajusté (ARI). L'algorithme proposé a été comparé avec trois approches de propagation de labels, notamment LP (Zhou et al. (2003)) et LS (Zhu et Ghahramani (2002)), qui sont les algorithmes classiques de propagation de labels, LNP (Wang et Zhang (2007)), qui est un autre algorithme de propagation de label avec une matrice d'affinité améliorée, ainsi qu'à CNMF (Liu et Wu (2010)), qui est une méthode de clustering sous contraintes basée sur NMF, et PLCC (Liu et al. (2017)), qui est une méthode de clustering sous contraintes basée sur k-means. Pour comparer les six approches, leurs paramètres connexes ont été spécifiés comme suit : chacun des algorithmes comparés LP, LS et NLP, nécessite une fonction gaussienne avec un paramètre libre σ à déterminer afin de construire leur matrice d'affinité. Dans les comparaisons, chacun de ces trois algorithmes a été testé avec différentes valeurs de σ, et sa meilleure valeur correspondant aux valeurs NMI et ARI les plus élevées sur chaque jeu de données a été sélectionnée. Le nombre de clusters k a été fixé au nombre réel de classes sur chaque jeu de données pour CNMF et PLCC. La performance d'une approche de propagation de labels dépend de la quantité d'information préalable disponible. Ainsi, dans les expériences, la quantité de données labéllisées a été fixée à 15, 25 et 35 pour cent du nombre total d'échantillons dans les jeux de données. La performance d'une approche de propagation de labels dépend également de la qualité de l'information préalable. Par conséquent, dans les expériences, étant donné la quantité d'information préalable, les six algorithmes comparés ont été exécutés avec 10 ensembles différents d'information préalable afin de calculer les résultats moyens pour NMI et ARI sur chaque jeu de données. La comparaison présente également les performances moyennes de chaque approche sur tous les jeux de données.

5.2 Résultats

Les tableaux 1 et 2 présentent les performances des six approches sur l'ensemble des jeux de données. Les expériences confirment que l'information préalable est capable d'améliorer l'efficacité de l'étiquetage, en fait, étant donné un jeu de données, tous les algorithmes de propagation de labels et de clustering sous contraints montrent une croissance de leur performance par rapport à NMI et ARI, en parallèle avec l'augmentation de la quantité d'information préalable. En outre, les tableaux montrent que l'approche proposée est clairement plus performante que LP, LS, NLP, CNMF et PLCC sur tous les jeux de données testés. Les tableaux présentent également les résultats moyens de chaque algorithme, qui confirment que l'approche de propagation de labels basée sur le transport optimal proposée surpasse les autres méthodes sur tous les jeux de données, suivie par LS, LP, NLP, PLCC puis CNMF, dans cet ordre.

Ces résultats sont principalement attribués à la capacité de OTP à capturer beaucoup plus d'informations que les autres méthodes grâce à la matrice d'affinité améliorée construite par

2. Les jeux de données sont disponible sur : `https://archive.ics.uci.edu/`

le transport optimal. Il convient également de noter que la performance de OTP réside dans le fait que le processus incrémental tire parti de la dépendance des algorithmes de propagation de labels à la quantité d'information préalable, donc l'enrichissement de l'ensemble étiqueté à chaque itération avec de nouvelles instances, permet aux échantillons non étiquetés d'être labellisées avec un degré de certitude élevé lors des itérations suivantes. Nous pouvons également expliquer l'amélioration apportée par notre approche par sa capacité à contrôler la certitude des prédictions d'étiquettes grâce au score de certitude utilisé, qui ne permet pas aux données d'être étiquetées sauf si elles ont un haut degré de certitude de prédiction.

TAB. 1 – *Les valeurs de NMI pour les méthodes d'apprentissage semi-supervisé transductif*

Jeux de données	pourcentage	LP	LS	LNP	CNMF	PLCC	OTP
Iris	15%	0.8412	0.8442	0.7534	0.5274	0.5835	**0.8447**
	25%	0.8584	0.8621	0.8269	0.5717	0.6489	**0.8667**
	35%	0.8621	0.8649	0.8314	0.6198	0.7067	**0.8852**
Ionosphere	15%	0.3502	0.3535	0.3256	0.2278	0.3007	**0.4676**
	25%	0.3848	0.3911	0.3572	0.2605	0.3356	**0.5000**
	35%	0.3972	0.4014	0.3725	0.2892	0.3529	**0.5383**
Dermatology	15%	0.8770	0.8779	0.8349	0.5531	0.6991	**0.8935**
	25%	0.8932	0.8932	0.8692	0.6238	0.7201	**0.9033**
	35%	0.9128	0.9128	0.8959	0.6703	0.7732	**0.9164**
Waveform	15%	0.4950	0.5009	0.4628	0.2453	0.3191	**0.5256**
	25%	0.5124	0.5192	0.4763	0.2619	0.3307	**0.5319**
	35%	0.5192	0.5229	0.4807	0.2792	0.3391	**0.5421**
Digits	15%	0.9150	0.9150	0.8891	0.1617	0.7412	**0.9290**
	25%	0.9443	0.9443	0.9268	0.2435	0.7801	**0.9489**
	35%	0.9570	0.9570	0.9318	0.3174	0.7956	**0.9607**
MNIST	15%	0.8019	0.8028	0.7759	0.2452	0.6329	**0.8177**
	25%	0.8389	0.8367	0.7931	0.2912	0.6506	**0.8442**
	35%	0.8542	0.8599	0.8136	0.3201	0.6711	**0.8730**
Tous les données	Moyenne	0.7721	0.7742	0.7346	0.3180	0.5015	**0.7994**

TAB. 2 – *Les valeurs de ARI pour les méthodes d'apprentissage semi-supervisé transductif*

Jeux de données	Pourcentage	LP	LS	LNP	CNMF	PLCC	OTP
Iris	15%	0.8453	0.8492	0.7861	0.4986	0.5403	**0.8621**
	25%	0.8680	0.8704	0.8321	0.5215	0.6029	**0.8884**
	35%	0.8754	0.8783	0.8424	0.5791	0.6471	**0.9027**
Ionosphere	15%	0.4221	0.4248	0.3998	0.2541	0.3491	**0.5723**
	25%	0.4606	0.4673	0.4324	0.3005	0.3491	**0.5927**
	35%	0.4650	0.4702	0.4418	0.3217	0.3902	**0.6281**
Dermatology	15%	0.8807	0.8813	0.8438	0.5725	0.7174	**0.8996**
	25%	0.8972	0.8972	0.8751	0.6401	0.7486	**0.9093**
	35%	0.9146	0.9146	0.9007	0.6935	0.7910	**0.9218**
Waveform	15%	0.5639	0.5678	0.5163	0.2819	0.3486	**0.5945**
	25%	0.5819	0.5864	0.5279	0.3059	0.3618	**0.6031**
	35%	0.5870	0.5880	0.5342	0.3242	0.3745	**0.6182**
Digits	15%	0.9126	0.9127	0.8993	0.1174	0.6931	**0.9306**
	25%	0.9432	0.9432	0.9287	0.1834	0.7306	**0.9508**
	35%	0.9567	0.9567	0.9407	0.2587	0.7294	**0.9621**
MNIST	15%	0.7930	0.7944	0.7697	0.0970	0.5692	**0.8393**
	25%	0.8487	0.8466	0.8152	0.1193	0.5927	**0.8685**
	35%	0.8721	0.8777	0.8438	0.1452	0.6201	**0.8935**
Tous les données	Moyenne	0.7604	0.7625	0.7460	0.2674	0.4963	**0.8017**

6 Conclusion

Dans cet article, nous avons proposé OTP, une nouvelle méthode traitant l'apprentissage semi-supervisé transductif à travers une propagation de labels. Notre méthode est différente des approches classiques et consiste à inférer une matrice d'affinité améliorée à partir du plan de transport optimal entre les instances étiquetées et non étiquetées. Une procédure incrémentale a été utilisée pour tirer profit de la dépendance des méthodes de propagation de labels à la quantité d'information préalable, et un score de certitude a été incorporé pour assurer la certitude des prédictions pendant le processus de propagation de labels. Les expériences ont montré que l'approche OTP surpasse les méthodes actuelles de l'état de l'art. À l'avenir, nous prévoyons d'étendre OTP à un contexte inductif et d'utiliser les pseudo-étiquettes inférées par OTP en conjonction avec les données étiquetées pour entraîner un modèle CNN pour résoudre des tâches de vision par ordinateur. Nous avons également l'intention de faire évoluer notre approche vers un transport optimal dans des espaces de représentation réduits. En particulier, par des projections non supervisées (Cabanes et Bennani (2007)) ou supervisées (Bennani (1992)), ou en utilisant la sélection de caractéristiques (Cakmakov et Bennani (2002)).

Références

Barba, E., L. Procopio, N. Campolungo, T. Pasini, et R. Navigli (2020). Mulan : Multilingual label propagation for word sense disambiguation. In *Proc. of IJCAI*, pp. 3837–3844.

Ben Bouazza, F. E., Y. Bennani, G. Cabanes, et A. Touzani (2020). Collaborative clustering through optimal transport. In *International Conference on Artificial Neural Networks*.

Ben Bouazza, F. E., Y. Bennani, M. El Hamri, G. Cabanes, B. Matei, et A. Touzani (2019). Multi-view clustering through optimal transport. *Aust. J. Intell. Inf. Process. Syst. 15*(3).

Bennani, Y. (1992). Text-independent talker identification system combining connectionist and conventional models. In *Neural Networks for Signal Processing*, Volume 2.

Cabanes, G. et Y. Bennani (2007). A simultaneous two-level clustering algorithm for automatic model selection. In *Sixth International Conference on Machine Learning and Applications*.

Cakmakov, D. et Y. Bennani (2002). *Feature selection for pattern recognition*. Informa.

Courty, N., R. Flamary, D. Tuia, et A. Rakotomamonjy (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence 39*(9).

Cuturi, M. (2013). Sinkhorn distances : Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300.

Iscen, A., G. Tolias, Y. Avrithis, et O. Chum (2019). Label propagation for deep semi-supervised learning.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, Volume 37, pp. 199–201.

Laclau, C., I. Redko, B. Matei, Y. Bennani, et V. Brault (2017). Co-clustering through optimal transport. *arXiv preprint arXiv :1705.06189*.

Liu, H., Z. Tao, et Y. Fu (2017). Partition level constrained clustering. *IEEE transactions on pattern analysis and machine intelligence 40*(10), 2469–2483.

Liu, H. et Z. Wu (2010). Non-negative matrix factorization with constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 24.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Redko, I., A. Habrard, et M. Sebban (2017). Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer.

Taherkhani, F., A. Dabouei, S. Soleymani, J. Dawson, et N. M. Nasrabadi (2020). Transporting labels via hierarchical optimal transport for semi-supervised learning. In *European Conference on Computer Vision*, pp. 509–526. Springer.

Van Engelen, J. E. et H. H. Hoos (2020). A survey on semi-supervised learning. *Machine Learning 109*(2), 373–440.

Villani, C. (2008). *Optimal transport : old and new*, Volume 338. Springer Science & Business Media.

Wang, F. et C. Zhang (2007). Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering 20*(1), 55–67.

Zhang, X.-K., J. Ren, C. Song, J. Jia, et Q. Zhang (2017). Label propagation algorithm for community detection based on node importance and label influence. *Physics Letters A*.

Zhou, D., O. Bousquet, T. Lal, J. Weston, et B. Schölkopf (2003). Learning with local and global consistency. *Advances in neural information processing systems 16*, 321–328.

Zhu, X. et Z. Ghahramani (2002). Learning from labeled and unlabeled data with label propagation.

Zhu, X. J. (2005). Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Summary

In this paper, we tackle the transductive semi-supervised learning problem that aims to obtain label predictions for the given unlabeled data points according to Vapnik's principle. Our proposed approach is based on optimal transport, a mathematical theory that has been successfully used to address various machine learning problems, and is starting to attract renewed interest in semi-supervised learning community. The proposed approach, Optimal Transport Propagation (OTP), performs in an incremental process, label propagation through the edges of a complete bipartite edge-weighted graph, whose affinity matrix is constructed from the optimal transport plan between empirical measures defined on labeled and unlabeled data. OTP ensures a high degree of predictions certitude by controlling the propagation process using a certainty score based on Shannon's entropy. Experiments task show the superiority of the proposed approach over the state-of-the-art. We make our code publicly available. [3]

Keywords: Optimal Transport, Semi-supervised learning, Label Propagation

3. Code is available at: `https://github.com/MouradElHamri/OTP`

Techniques de génération de population initiale d'algorithmes génétiques pour la sélection de caractéristiques*

Marc Chevallier[1], Nicoleta Rogovschi[1], Faouzi Boufarès[1], Nistor Grozavu[1], and Charly Clairmont

[1] LIPN Laboratory, Sorbonne Paris Nord University, Villetaneuse, France
mchevallier@lipn.univ-paris13.fr
[2] nicoleta.rogovschi@lipn.univ-paris13.fr
[3] faouzi.boufares@lipn.univ-paris13.fr
[4] Nistor.Grozavu@lipn.univ-paris13.fr
[5] charly.clairmont@synaltic.fr

Abstract. Le processus de sélection des caractéristiques est un problème difficile. L'objectif est de sélectionner le meilleur sous ensemble d'un ensemble de caractéristiques pour un algorithme d'apprentissage artificiel donné. Le meilleur étant, celui qui pour un algorithme d'apprentissage artificiel donné permet d'obtenir le meilleur taux de bonne reconnaissance. Les algorithmes génétiques (AG) sont faciles à implémenter et leurs résultats sont aisément explicables. Cependant, il n'est pas garanti qu'ils trouvent la meilleure solution. Dans le but d'améliorer la performance des AG, nous avons introduit deux méthodes de création de la population initiale de l'AG. Les deux techniques ont été utilisées sur deux AG utilisant un classifieur Bayésien pour évaluer le taux de bonne reconnaissance. Les tests ont été effectués sur cinq jeux de données et nos méthodes ont été comparées à d'autres algorithmes de réduction dimensionnelle. Nos résultats indiquent une amélioration des performances lorsque nos techniques sont utilisées.

Keywords: Machine learning · Meta-heuristic · Random Forest · Bayesian networks · Features selection · Genetic Algorithm

1 Introduction

Dans un monde où de plus en plus de données sont collectées, assurer la qualité des données devient une problématique majeure. De nombreuses études ont été menées ces dernières années sur la qualité des données [9], particulièrement sur les données dupliquées [9], ainsi que sur les anomalies causées par le non respect des dépendances fonctionnelles qui peuvent exister entre les colonnes [1]. Une manière d'analyser la qualité des données structurées est de découvrir les dépendances fonctionnelles qu'elles contiennent (les relations entre les colonnes).

* Soutenu par l'entreprise Synaltic

Dans ce domaine, nous explorons une nouvelle méthode pour découvrir ces dépendances. Notre algorithme actuel s'appuie sur l'apprentissage artificiel et l'extraction de nombreuses caractéristiques sur les données. Cependant, nous avons besoin d'une méthode pour sélectionner les meilleures caractéristiques.

La sélection de caractéristiques est un problème classique en apprentissage artificiel. L'objectif est de sélectionner le meilleur sous ensemble de caractéristiques pour un problème d'apprentissage artificiel donné. Le domaine se compose de trois types de méthode : embedded, wrappers et filters methods [13].

Les méthodes Wrappers et embedded nécessitent l'usage d'un classifieur pour sélectionner les meilleures caractéristiques, les algorithmes les plus communs dans ce domaine sont : LASSO [17], random selection [3], Recursive feature Elimination [6], Algorithmes génétiques [15]. En revanche, les méthodes filters ne nécessitent pas l'utilisation d'un classifieur, des exemples de ce type d'algorithmes sont: Correlation-based Feature Selection, Information Gain, ReliefF [7].

Nous avons choisi les algorithmes génétiques pour leur simplicité d'implémentation ainsi que pour la possibilité d'optimiser les résultats pour un classifieur spécifique.

2 Algorithmes génétiques

2.1 Généralités sur les Algorithmes génétiques

Les algorithmes génétiques sont un sous-ensemble des algorithmes de métaheuristique inspirés par la génétique. L'objectif de ces algorithmes est de déterminer les meilleurs paramètres pour optimiser une fonction d'évaluation. Une population initiale est générée, chaque individu qui compose cette population contient un ensemble de paramètres appelé chromosome. Chaque paramètre dans un chromosome est appelé gène, ces gènes peuvent être encodés de différentes manières. Ces algorithmes s'appuient sur deux mécanismes, la mutation et le croisement. La mutation modifie aléatoirement des gènes dans certains individus et le croisement a lieu dans une phase de l'algorithme appelée reproduction. Dans cette phase deux individus échangent une partie de leur chromosomes afin de créer de nouveaux individus appelés progéniture. Après chaque génération, tous les individus sont évalués à l'aide d'une fonction d'évaluation et les meilleurs sont conservés pour la génération suivante. Les AG sont utilisés dans de nombreux problèmes. Dans certains cas, initialiser la population de départ avec des individus non aléatoires permet d'obtenir de meilleurs résultats [14,11].

2.2 Éclectique GA (EGA)

Nous allons tout d'abord introduire l'algorithme génétique appelé Éclectique GA (EGA). EGA utilise un croisement annulaire, des mutations uniformes ainsi qu'un élitisme total [12]. Si l'on note G le nombre de générations, n le nombre d'individus, *B2M* le nombre de gènes à muter, *I(n)* le *n-ième* individu, L la longueur d'un chromosome, P_c la probabilité de croisement et P_m la probabilité de mutation.

EGA [12], est décrit dans l'algorithme 1.

Algorithm 1 EGA

Étape 0 $B2M \leftarrow [nLxP_m]$
Étape 1 $i \leftarrow 1$
Étape 2 Générer une population aléatoire
Étape 3 Évaluer la population
Étape 4 Dupliquer la population
for j = 1 à n **do**
  I(n+j)←I(j)
  fitness(n+j)←fitness(j)
end for
Étape 5 Croisement annulaire déterministe
for j = 1 à n/2 **do**
  Générer un nombre aléatoire uniforme $0 \leq \rho$ 1
  if If $\rho \leq P_c$ **then**
    Générer un nombre aléatoire $1 \leq \rho$ L/2
    Échanger le demi-anneau en commençant à l'emplacement ρ entre I(j) et I(n-j-1)
  end if
end for
Étape 6 Mutation
for j à B2M **do**
  Générer un nombre aléatoire uniforme 0 ρ_1,ρ_2 1 Muter Bit [ρ_2L] de I([ρ_1n])
end for
Étape 7 Sélection
Classer les 2n individus par leur fitness, de manière ascendante
Étape 8 $i \leftarrow i + 1$
if If i = G **then**
  retourner I(1) et arrêt
else
  Aller à Étape 3
end if

2.3 GAAM

L'algorithme génétique avec des mutations agressives (GAAM) est un AG avec des particularités [16,10]. Si l'on note L la dimension de l'ensemble de caractéristiques, K le nombre d'individus, V le nombre de gènes par chromosome et G le nombre de générations. GAAM utilise un système d'encodage par entier des gènes, chaque chromosome contient une liste de valeurs entre *0* et L. GAAM est décrit dans l'algorithme 2.

3 Expériences

3.1 Conditions d'expérience

Tous nos tests ont été effectués sur des instances google colab, équipées de processeurs Intel Xeon 2.30GHz à 4 coeurs et 12Go de RAM.

Pour nos expériences, nous avons choisi cinq jeux de données décrits dans le tableau 1.

Algorithm 2 GAAM

INPUT : *G, K, V ,P*
g = 0
1 Construire *K* individus avec *V* gènes choisis aléatoirement dans {0,1,2..L} dans le but de créer la population initiale *Ip*
2 Agressive mutation : Créer *Mp* la population mutée
for j=1 à K **do**
  for x=1 à V **do**
    choisir une valeur aléatoire *m* dans {*0,1,2..L*}
    Assigner à *D* une copie de **Ip(k)**
    D(x)=m
    ajouter *D* à *Mp*
  end for
end for
3 Croisement : appliquer un classique croisement de holland à chaque individu
4 Créer *Tp=Ip+Mp+Cp* évaluer chaque individu de la population totale et les classer par leur fitness.
5 Supprimer les *K+K*V* individus avec la fitness la plus faible de *Tp* et remplacer *Ip* par les individus restants *g+=1* Si *g = G* retourner *Ip(0)* Sinon retour à l'Étape 2

Table 1: Description des jeux de données

Jeux de données	Nombre de caractéristiques	Nombre de lignes	Nombre de classes
Madelon	500	2600	2
Semeion	256	1593	10
Har	561	10299	6
Parkinson	753	754	2
Hill-Valley	100	1212	2

Nous avons concentré notre analyse sur deux jeux de données Semeion et Madelon. Nous avons comparé différentes proportions d'initialisation non aléatoire de la population initiale: 0%,10%,50% et 100%. Pour EGA, nous avons sélectionné une population initiale de 100 individus, et choisi comme maximum d'itérations 100. Les probabilités de mutation et de croisement sont calculées dynamiquement en suivant la stratégie ILM/DHC [8] :

On note *Gi* le numéro de la génération courante, *Gm* le nombre maximum de générations.

$$P_c = 1 - \frac{Gi}{Gm} \tag{1}$$

$$P_m = \frac{Gi}{Gm} \tag{2}$$

Le classifieur utilisé est un classifieur bayésien naïf [19]. Les résultats sont mesurés à partir de la moyenne du taux de bonne reconnaissance calculé sur cinq validations croisées et est noté *ma*. La fonction d'évaluation (fitness) est définie par 1 - *ma*. Pour l'algorithme GAAM, nous avons décidé de conserver

20 caractéristiques, d'utiliser une population de 20 individus, d'utiliser la même fonction d'évaluation ainsi que de nous limiter à 100 itérations.

4 Techniques de création de la population initiale

Pour l'initialisation de la population, nous avons développé deux méthodes notées SSM (Standard Seeding Method) et ESM (Elitism Seeding Method). Notre modèle pour la création des chromosomes nécessaires à l'initialisation s'appuie sur l'utilisation d'une forêt d'arbres décisionnels [2]. Nous entraînons tout d'abord une forêt d'arbres décisionnels sur le jeu de données traité avec les paramètres suivants : 100 estimateurs, profondeur maximale 8. L'importance de chaque caractéristique est évaluée par la mesure d'impureté gini et normalisée (la somme du score de toutes les caractéristiques est égale à 1). Ensuite, nous utilisons cette évaluation comme probabilité de sélection pour chaque caractéristique dans notre algorithme.

Pour SSM, nous utilisons les probabilités préalablement définies pour sélectionner aléatoirement le nombre de caractéristiques pour chaque individu.

Pour ESM nous procédons de la même manière mais en générant cinq fois plus d'individus que nécessaire puis nous évaluons ces individus à l'aide d'un classifieur bayésien naïf et nous ne gardons que les N premiers individus dont nous avons besoin.

Les chromosomes générés de cette façon peuvent avoir des valeurs dupliquées. GAAM gère nativement ce problème en ne gardant qu'une des valeurs. Pour EGA, lors de la conversion en binaire des gènes, les gènes dupliqués sont supprimés.

Pour s'assurer de la validité de la méthode, nous avons généré 1000 exemples (en utilisant le jeu de données Semeion, avec un classifieur bayéien, le taux de reconnaissance étant calculé à l'aide de la moyenne de cinq validations croisées) pour chaque méthode, les résultats sont présentés dans la Fig 1.

Le taux de reconnaissance médian est plus élevé dans les groupes générés à l'aide de SSM et de ESM. Ces résultats indiquent que les deux méthodes génèrent des individus initiaux meilleurs que ceux générés aléatoirement. De plus les individus générés présentent beaucoup moins de variabilité au niveau des résultats tout en gardant une assez grande diversité.

5 Résultats

Du fait de la nature probabiliste des algorithmes génétiques, chaque simulation est effectuée 10 fois et la moyenne des 10 simulations est conservée. Nous pouvons tirer quelques conclusions de ces résultats.

Sur les 60 expériences que nous avons menées en utilisant nos techniques d'initialisation, 50 présentent de meilleurs résultats que celles sans initialisation. Pour les 10 résultats restants, ils sont quasiment identiques à ceux sans technique d'initialisation. Cela confirme l'utilité de l'approche. Ces résultats sont logiques

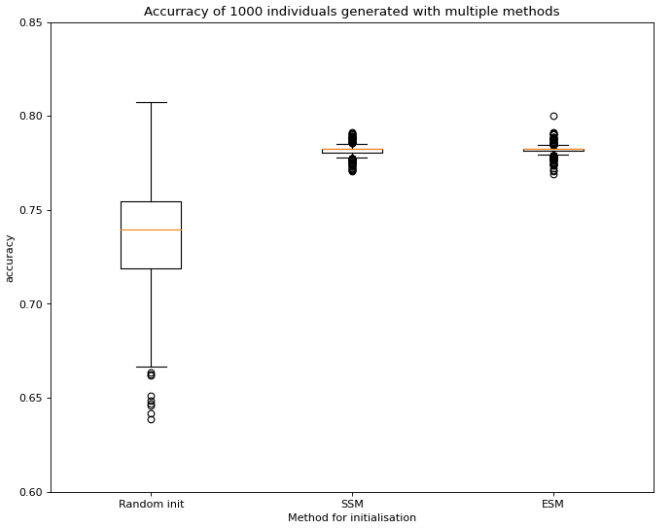


Fig. 1: Boite à moustache du taux de reconnaissance pour 1000 individus générés avec chaque méthode

car les individus qui sont injectés dans la population initiale via nos méthodes sont déjà passés à travers un processus d'optimisation avec le démarrage de l'AG.

Dans la section suivante, nous allons analyser plus profondément les résultats sur Semeion et Madelon.

5.1 Résultats avec l'algorithme EGA

Pour EGA, les résultats sont particulièrement intéressants car l'initialisation par nos techniques mène majoritairement à un meilleur taux de reconnaissance et une meilleure stabilité des résultats. Les Fig.2 et 3 présentent les résultats de EGA en utilisant la technique SSM.

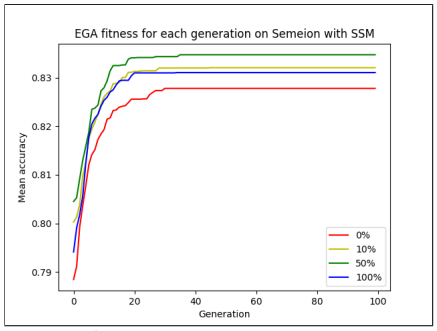


Fig. 2: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant SMM sur Semeion

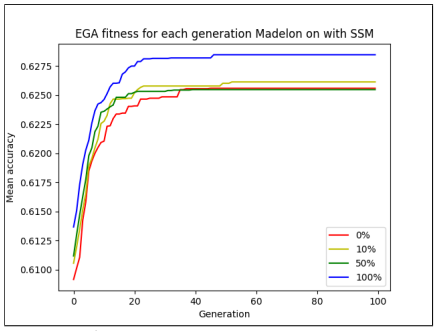


Fig. 3: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant SMM sur Madelon

Les résultats indiquent que notre technique permet d'obtenir des résultats meilleurs qu'une initialisation aléatoire. De plus dans le pire scénario les résultats

sont équivalents à ceux réalisés par la méthode d'initialisation aléatoire. La population issue de notre technique étant déjà plus optimale que la population issue d'une initialisation aléatoire, elle présente toujours de meilleurs résultats jusqu'à la vingtième génération. La vitesse de convergence n'est pas affectée par notre technique. De plus, les résultats d'EGA se stabilisent toujours après environ 50 générations. Notre technique nous permet d'atteindre des taux de reconnaissance plus élevés.

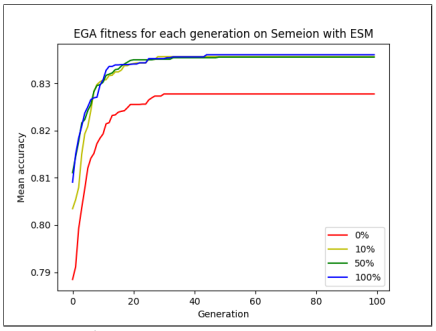


Fig. 4: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant ESM sur Semeion

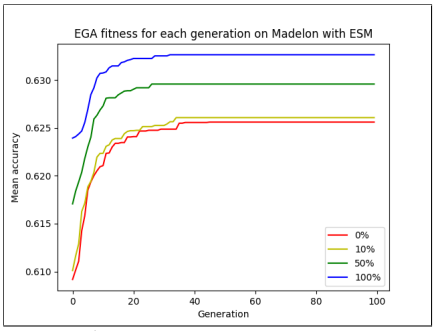


Fig. 5: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant ESM sur Madelon

Avec la technique ESM (Fig 4 et 5), notre technique performe toujours mieux qu'une initialisation aléatoire. Sur Semeion qui ne contient que 256 caractéristiques, qu'importe la proportion d'individus optimisés dans la population initiale; l'algorithme converge toujours vers le même résultat. En revanche, sur Madelon, plus le pourcentage de population optimisée dans la population initiale augmente, plus les résultats s'améliorent. On peut en conclure que ESM renvoie toujours un groupe d'individus initiaux qui sont très bons ce qui conduit l'algorithme à converger vers un meilleur résultat final.

5.2 Résultats avec l'algorithme GAAM

Avec l'algorithme GAAM (Fig 6,7,8 et 9), la taille de l'espace de recherche est fortement réduite car le nombre de caractéristiques à sélectionner est fixé dès le départ. Cette réduction mène à moins de possibilités d'optimisation. Comme prévu les résultats sont meilleurs sur Madelon que sur Semeion car l'espace de recherche est plus grand. Avec SSM, une proportion de 10% d'individus optimisés dans la population initiale mène aux meilleurs résultats dans tous les cas. Ce résultat peut être expliqué par le fonctionnement de GAAM: durant la phase de mutation agressive chaque gène de chaque chromosome est muté. S'il y a trop de similitudes dans la population initiale à cause de notre technique, les bons gènes demeurent, mais une trop grande proximité entre les individus initiaux bride l'espace de recherche.

Les résultats de toutes nos expériences sont résumés dans la Table 2 et la Table 3. Nous avons ajouté l'écart type calculé sur les 10 simulations pour évaluer la dispersion des résultats pour chaque scénario.

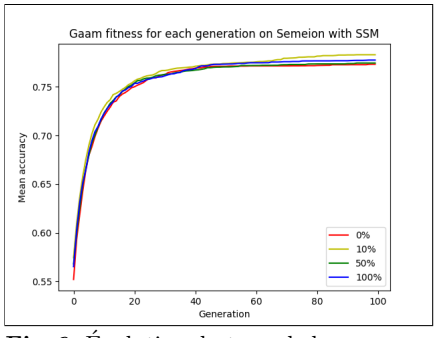


Fig. 6: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant SMM sur Semeion

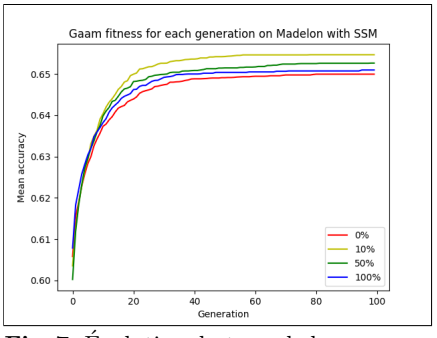


Fig. 7: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant SMM sur Madelon

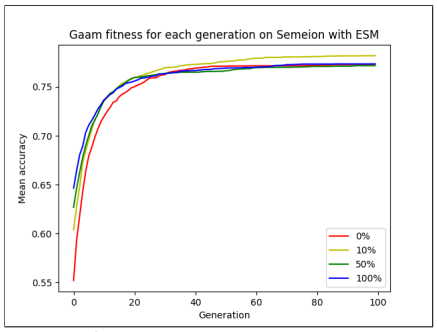


Fig. 8: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant ESM sur Semeion

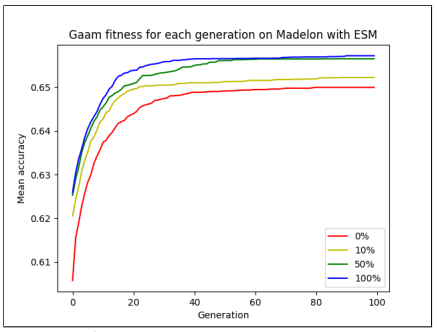


Fig. 9: Évolution du taux de bonne reconnaissance moyen calculé sur 10 simulations en utilisant ESM sur Madelon

Table 2: Résultats expérimentaux pour le jeu de données Madelon

Madelon	Initialisation standard	SSM			ESM		
% of seeding	0%	10%	50%	100%	10%	50%	100%
GAAM accuracy	0.649	0.655	0.653	0.651	0.652	0.656	0.657
std	0.007	0.004	0.005	0.002	0.007	0.008	0.004
Ega accuracy	0.625	0.626	0.625	0.628	0.626	0.630	0.632
std	0.002	0.001	0.001	0.001	0.002	0.0018	0.001

Table 3: Résultats expérimentaux pour le jeu de données semeion

Semeion	Initialisation standard	SSM			ESM		
% of seeding	0%	10%	50%	100%	10%	50%	100%
GAAM accuracy	0.773	0.783	0.774	0.777	0.781	0.771	0.773
std	0.002	0.007	0.009	0.008	0.004	0.009	0.006
Ega accuracy	0.827	0.832	0.834	0.831	0.835	0.835	0.836
std	0.0028	0.0027	0.0024	0.0026	0.0041	0.0027	0.0013

5.3 Comparaison avec d'autres méthodes de réduction dimensionnelle

Nous avons comparé les résultats de nos méthodes avec deux méthodes classiques de réduction dimensionnelle ACP [5] et Auto-encodeur [18] sur cinq jeux de données. Ces tests sont effectués avec deux paramétrages différents. Le premier est une réduction dimensionnelle à 20 dimensions afin de pouvoir comparer les résultats avec ceux de GAAM. Le deuxième est une réduction dimensionnelle à une taille égale à la moitié du nombre de caractéristiques initiales afin de pouvoir comparer avec EGA (les algorithmes génétiques classiques utilisant comme critère d'évaluation le taux de bonne reconnaissance tendent à converger vers ce résultat [10]. Les paramètres de l'auto-encodeur sont les suivants: 22 couches utilisant Selu et Sigmoïde comme fonction d'activation.

Les résultats (dans les tableaux 4 et 5) indiquent que les AG utilisant notre technique réalisent de meilleurs résultats que les autres types de réductions dimensionnelles; sur quatre jeux de données sur cinq, les résultats sont particulièrement meilleurs si le jeu de données contient plus de 500 caractéristiques.

Table 4: Résultats avec 20 caractéristiques

Methods	Madelon	Semeion	Hill-Valley	Har	Parkinson
PCA 20 caractéristiques	0.627	**0.867**	0.511	0.794	0.725
Auto-encoder 20 caractéristiques	0.550	0.784	0.511	0.779	0.737
GAAM SSM meilleur résultat	0.654	0.783	0.518	**0.927**	0.871
GAAM ESM meilleur résultat	**0.657**	0.781	**0.519**	0.927	**0.873**

Table 5: Résultats sur la moitié du nombre de caractéristique

Methods	Madelon	Semeion	Hill-Valley	Har	Parkinson
PCA	0.582	**0.883**	0.504	0.675	0.681
Auto-encodeur	0.545	0.819	0.510	0.779	0.7369
EGA SSM meilleur résultat	0.628	0.835	0.518	0.878	0.805
EGA ESM meilleur résultat	**0.632**	0.836	**0.519**	**0.881**	**0.807**

6 Conclusion

Dans cette étude, nous avons étudié l'influence de techniques d'initialisation pour les algorithmes génétiques dans le cadre de la sélection de caractéristiques [4]. Nos expériences ont été réalisées avec deux algorithmes génétiques, deux méthodes d'initialisation et cinq jeux de données. Dans chaque scénario, trois taux d'initialisation non aléatoire ont été comparés. Nos techniques ont aussi été comparées à deux algorithmes classiques de réduction dimensionnelle. Nos résultats indiquent qu'initialiser de manière non aléatoire la population de départ est une méthode utile pour améliorer les résultats des algorithmes génétiques dans le cadre de la sélection de caractéristiques, surtout quand les jeux de données sont de grandes dimensions. Ces résultats sont prometteurs et nous inspirent des solutions pour le cadre de l'étude de la qualité des données. Cependant, d'autres études doivent être menées afin de trouver une méthode pour identifier de manière fiable la quantité d'individus optimisés à injecter dans la population initiale afin d'obtenir le meilleur résultat.

References

1. Berti-Équille, L., Harmouch, H., Naumann, F., Novelli, N., Thirumuruganathan, S.: Discovery of genuine functional dependencies from relational data with missing values. Proc. VLDB Endow. **11**(8), 880–892 (Apr 2018). https://doi.org/10.14778/3204028.3204032, https://doi.org/10.14778/3204028.3204032
2. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (Oct 2001). https://doi.org/10.1023/A:1010933404324, https://doi.org/10.1023/A:1010933404324
3. Burduk, R.: Recognition task with feature selection and weighted majority voting based on interval-valued fuzzy sets. In: Nguyen, N.T., Hoang, K., Jdrzejowicz, P. (eds.) Computational Collective Intelligence. Technologies and Applications. pp. 204–209. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
4. Chevallier, M., Rogovschi, N., Boufarès, F., Grozavu, N., Clairmont, C.: Seeding initial population, in genetic algorithm for features selection. In: Abraham, A., Ohsawa, Y., Gandhi, N., Jabbar, M., Haqiq, A., McLoone, S., Issac, B. (eds.) Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020). pp. 572–582. Springer International Publishing, Cham (2021)
5. F.R.S., K.P.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901). https://doi.org/10.1080/14786440109462720, https://doi.org/10.1080/14786440109462720
6. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46**(1), 389–422 (Jan 2002). https://doi.org/10.1023/A:1012487302797, https://doi.org/10.1023/A:1012487302797
7. Hall, M.A.: Correlation-based feature selection of discrete and numeric class machine learning. Tech. rep. (2000), https://hdl.handle.net/10289/1024, working Paper

8. Hassanat, A., Almohammadi, K., Alkafaween, E., Abunawas, E., Hammouri, A., Prasath, V.B.S.: Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach. Information **10**(12) (2019). https://doi.org/10.3390/info10120390, https://www.mdpi.com/2078-2489/10/12/390
9. Ilyas, I.F., Chu, X.: Data Cleaning. Association for Computing Machinery, New York, NY, USA (2019)
10. Izabela RejerIzabela, R.L.: Classic genetic algorithm vs. genetic algorithm with aggressive mutation for feature selection for a brain-computer interface (2015)
11. Julstrom, B.A.: Seeding the population: Improved performance in a genetic algorithm for the rectilinear steiner problem. In: Proceedings of the 1994 ACM Symposium on Applied Computing. p. 222–226. SAC '94, Association for Computing Machinery, New York, NY, USA (1994). https://doi.org/10.1145/326619.326728, https://doi.org/10.1145/326619.326728
12. Kuri-Morales, A., Aldana-Bobadilla, E.: The best genetic algorithm i. In: Castro, F., Gelbukh, A., González, M. (eds.) Advances in Soft Computing and Its Applications. pp. 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
13. Liu, H., Motoda, H.: Feature selection for knowledge discovery and data mining, vol. 454. Springer Science & Business Media (2012)
14. Osaba, E., Carballedo, R., Diaz, F., Onieva, E., Lopez, P., Perallos, A.: On the influence of using initialization functions on genetic algorithms solving combinatorial optimization problems: A first study on the tsp. In: 2014 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). pp. 1–6 (2014). https://doi.org/10.1109/EAIS.2014.6867465
15. Reeves, C.R.: Genetic algorithms. In: Handbook of metaheuristics, pp. 109–139. Springer (2010)
16. Rejer, I.: Genetic algorithm with aggressive mutation for feature selection in bci feature space. Pattern Analysis and Applications **18**(3), 485–492 (Aug 2015). https://doi.org/10.1007/s10044-014-0425-3, https://doi.org/10.1007/s10044-014-0425-3
17. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996). https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x, https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x
18. Wang, Y., Yao, H., Zhao, S.: Auto-encoder based dimensionality reduction. Neurocomputing **184**, 232–242 (2016). https://doi.org/https://doi.org/10.1016/j.neucom.2015.08.104, https://www.sciencedirect.com/science/article/pii/S0925231215017671, roLoD: Robust Local Descriptors for Computer Vision 2014
19. Zhang, H.: The optimality of naive bayes (2004)

Clustering quantique à base de prototypes

Kaoutar Benlamine [*,**], Younès Bennani[*,**], Ahmed Zaiou[*,**,***]
Mohamed Hibti [***], Basarab Matei [*,**], Nistor Grozavu [*,**]

[*] LIPN UMR 7030 CNRS, Université Sorbonne Paris Nord, France
name.surname@sorbonne-paris-nord.fr
[**]LaMSN, La Maison des Sciences Numériques, France
name.surname@lamsn.sorbonne-paris-nord.fr
[***] EDF Lab Saclay, PERICLES, Gaspard Monge Street, Palaiseau, France
name.surname@edf.fr

Résumé. L'apprentissage quantique est un nouveau domaine de recherche avec de récents travaux sur les algorithmes quantiques supervisés et non supervisés. Ces dernières années, de nombreux algorithmes d'apprentissage quantique permettant une accélération des algorithmes classiques ont été proposés. Dans cet article, nous proposons une analyse et une comparaison de trois distances pour les techniques quantiques de classification basées sur des prototypes. Comme application de ce travail, nous présentons la version quantique de l'algorithme K-means qui donne une bonne classification tout comme sa version classique, la différence réside dans la complexité : alors que la version classique de K-means prend un temps polynomial, sa version quantique ne prend qu'un temps logarithmique, en particulier pour les grands ensembles de données. Enfin, des expérimentations ont été faites pour valider l'approche proposée.

Mots-clés : Apprentissage artificiel quantique, K-means, Apprentissage non supervisé.

1 Introduction

Comme la quantité de données générées dans notre société augmente considérablement, il est nécessaire de disposer de moyens plus puissants de traitement de l'information. C'est pour cela les études et les applications récentes se concentrent sur le problème de l'apprentissage automatique à grande échelle. Beaucoup de travaux ont été consacrés à l'apprentissage quantique. Par exemple, l'élaboration de procédures quantiques pour l'algèbre linéaire comme : multiplication matricielle, vecteurs propres et valeurs propres de matrices et estimation de distances entre états quantiques. Des efforts ont également été faits pour résoudre des problèmes de reconnaissance des formes (Schützhold, 2003) et pour développer des versions quantiques des réseaux de neurones artificiels (Gupta et Zia, 2001) largement utilisés en apprentissage automatique. Dans cet article, nous nous intéressons à l'estimation de la distance pour la classification non supervisée quantique basée sur des prototypes, car la tâche principale de l'algorithme d'apprentissage automatique consiste à analyser la similarité entre les vecteurs à travers

l'évaluation de la distance et du produit scalaire. La notion de distance dans le cas des algorithmes de classification quantique diffère de celle classique dans le sens qu'elle peut varier en fonction des effets probabilistes dûs à la nature quantique des états.

Le reste du papier est organisé comme suit. La section 2 présente l'estimation de la distance. La section 3 serait la description de l'algorithme proposé à savoir K-means quantique. La section 4 est consacrée aux résultats expérimentaux. Enfin, la conclusion résume notre travail et ses avantages et propose quelques directions de réflexion.

2 Estimation de la distance entre une donnée et un centroïde

La mesure de distance est nécessaire pour un algorithme de classification non supervisée (clustering). La question qui se pose c'est comment pouvons-nous mesurer la distance d'une manière facile et efficace dans un contexte quantique ? Pour le cas conventionnel, calculer les distances euclidiennes, par exemple, est facile, mais le faire de la même manière dans le cas quantique serait beaucoup plus compliqué et nécessiterait plus de qubits que nous ne pouvons nous permettre. D'autre part, la nature probabiliste des qubits facilite la mesure des différences de phase et des amplitudes de probabilité.

Pour les algorithmes de clustering, les distances sont nécessaires simplement pour affecter des points de données à différents clusters. L'objectif est alors de savoir quel cluster est le plus proche d'un point de données. Cette mesure n'a pas besoin d'être proportionnelle à la distance réelle, mais seulement en corrélation positive avec celle-ci. En effet, nous n'avons besoin que du plus proche centroïde et non des valeurs exactes des distances réelles.

En informatique quantique, de nombreuses mesures de type distance sont disponibles lorsque nous traitons des qubits, comme le produit scalaire entre deux vecteurs (normalisés) et les probabilités de mesurer un qubit dans les états $|0\rangle$ ou $|1\rangle$ (Wiebe et al., 2018), (Lloyd et al., 2013), (Anagolum, 2019).

2.1 Fidelité comme mesure de similarité

La fidélité est une mesure de similarité entre deux états quantiques, définie dans le cas de deux états purs $|\psi\rangle$ et $|\phi\rangle$ par $Fid(|\psi\rangle, |\phi\rangle) = |\langle\psi|\phi\rangle|^2$. La fidélité varie entre 0 si les états sont orthogonaux (c'est-à-dire parfaitement distinguables) et 1 si les états sont identiques.

La fidélité $|\langle\psi|\phi\rangle|$ (Aïmeur et al., 2006) de deux états quantiques $|\psi\rangle$ et $|\phi\rangle$ peut être obtenue par le circuit swap test présenté dans la figure 1.

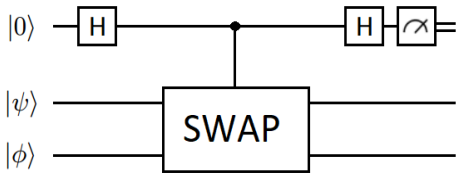


FIG. 1: Circuit swap test

Le circuit swap test permet de comparer deux états quantiques. Il est composé de deux portes de Hadamard et d'une porte Control swap.

La mesure fournie par le circuit, donne la probabilité $P(|0_A\rangle) = \frac{1}{2} + \frac{1}{2}|\langle\psi|\phi\rangle|^2$. Une probabilité de $1/2$ montre que les deux états quantiques $|\psi\rangle$ et $|\phi\rangle$ sont orthogonaux, alors qu'une probabilité de 1 indique qu'ils sont identiques.

2.2 Construction des états

Plusieurs travaux ont été réalisés pour calculer la fidélité entre deux états quantiques. Tout ces travaux utilisent le circuit du swap test pour obtenir la mesure de similarité, mais une préparation et une construction de données différentes. Dans cette section, nous présentons la phase de préparation des données ainsi que celle de la construction des états $|\psi\rangle$ et $|\phi\rangle$ pour chaque méthode.

Considérons que nous voulons calculer la distance entre les deux états quantiques $|x\rangle$ et $|w\rangle$.

2.2.1 Approche Wiebe et al.

1. *Préparation des données*
   Étant donné $N = 2^n$ vecteurs complexes $\vec{x}$ et $\vec{w}$ avec des composantes $x_j = |x_j|e^{-i\alpha_j}$ et $w_j = |w_j|e^{-i\beta_j}$ respectivement. Supposons que $\{|x_j|, \alpha_j\}$ et $\{|w_j|, \beta_j\}$ sont stockés sous forme de nombres réelles.
2. *Construction des états*
   Wiebe, Kapoor et Svore (Wiebe et al., 2018) suggèrent une représentation des états qui vise à écrire les paramètres en amplitudes des états quantiques.

$$|\psi\rangle = \frac{1}{\sqrt{d}} \sum_j |j\rangle \left(\sqrt{1 - \frac{|x_j|^2}{r_{max}^2}} e^{-i\alpha_j} |0\rangle + \frac{x_j}{r_{max}} |1\rangle \right) |1\rangle$$

$$|\phi\rangle = \frac{1}{\sqrt{d}} \sum_j |j\rangle |1\rangle \left(\sqrt{1 - \frac{|w_j|^2}{r_{max}^2}} e^{-i\beta_j} |0\rangle + \frac{w_j}{r_{max}} |1\rangle \right)$$

   Avec $j = \{1, ..., n\}$ et r_{max} est la limite supérieure de la valeur maximale de toute entité de l'ensemble des données. Les vecteurs d'entrée sont d-sparse, c'est-à-dire ne contiennent pas plus de d entrées autres que zéro.

En utilisant le circuit de Swap test, le produit scalaire est évalué par :

$$d_{q_1}(|x\rangle, |w\rangle) = d^2 r_{max}^4 (2P(|0\rangle) - 1) \tag{1}$$

2.2.2 Approche Lloyd et al.

1. *Préparation des données*
   Afin d'utiliser les forces de la mécanique quantique sans se limiter aux idées classiques d'encodage des données, Lloyd, Mohseni et Rebentrost (Lloyd et al., 2013) ont proposé un moyen de coder le vecteur classique en un état quantique.
   Considérons $N = 2^n$ vecteurs complexes $\vec{x}$ et $\vec{w}$, on a :

$$|x\rangle = \frac{\vec{x}}{|\vec{x}|} \ , \ |w\rangle = \frac{\vec{w}}{|\vec{w}|}$$

2. *Construction des états*
   Seth Lloyd et ses collègues ont proposé un moyen pour construire les états $|\psi\rangle$ et $|\phi\rangle$. L'idée est de joindre un bit auxiliaire aux états créant un état intriqué $|\psi\rangle$. Plus la différence entre les états $|x\rangle$ et $|w\rangle$ est grande, plus l'état résultant est intriqué (Cai et al., 2015).

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle\,|x\rangle + |1\rangle\,|w\rangle)$$

$$|\phi\rangle = \frac{1}{\sqrt{Z}}(|\vec{x}|\,|0\rangle - |\vec{w}|\,|1\rangle)$$

   Où $Z = |\vec{x}|^2 + |\vec{w}|^2$

Après avoir appliqué le circuit de Swap test, la distance est évaluée par :

$$d_{q_2}(|x\rangle\,, |w\rangle) = 2Z(2P(|0\rangle) - 1) \tag{2}$$

2.2.3 Approche Anagolum

1. *Préparation des données*
   Pour simplifier, supposons que nous sommes dans un espace à 2 dimensions. Considérons que nous avons deux vecteurs $\vec{x}(x_0, x_1)$ et $\vec{w}(w_0, w_1)$.
   Nous pouvons mapper les valeurs des données sur les valeurs θ et α en utilisant les équations ci-dessous. Pour x nous obtenons :

$$\alpha_0 = (x_0 + 1)\frac{\pi}{2} \ , \ \theta_0 = (x_1 + 1)\frac{\pi}{2} \tag{3}$$

   De même pour w nous obtenons :

$$\alpha_1 = (w_0 + 1)\frac{\pi}{2} \ , \ \theta_1 = (w_1 + 1)\frac{\pi}{2} \tag{4}$$

2. *Construction des états*
   Pour construire les deux états $|\psi\rangle$ et $|\phi\rangle$ Anagolum (Anagolum, 2019) a proposé d'utiliser la porte U comme suit :

$$|\psi\rangle = U(\theta_0, \alpha_0, 0)|0\rangle \tag{5}$$

$$|\phi\rangle = U(\theta_1, \alpha_1, 0)|0\rangle \tag{6}$$

   En effet, la porte U implémente les rotations que nous devons effectuer pour encoder nos points de données.

$$U(\theta,\alpha,\lambda) = \begin{pmatrix} \cos\frac{\theta}{2} & -e^{i\lambda}\sin\frac{\theta}{2} \\ e^{i\alpha}sin\frac{\theta}{2} & e^{i\lambda+i\alpha}\cos\frac{\theta}{2} \end{pmatrix}$$

Cette instruction entraînerait le qubit à déplacer θ radians de l'axe z positif, et α radians de l'axe x positif.

En utilisant le circuit de Swap test, la distance est évaluée par :

$$d_{q_3}(|x\rangle, |w\rangle) = P(|1\rangle) \tag{7}$$

3 Méthode de classification quantique proposée

3.1 Concept général

K-means (MacQueen et al., 1967) est un type d'apprentissage automatique non supervisé qui vise à rechercher des groupes homogènes dans les données en divisant le jeu de données en K groupes (clusters). Dans cette section, nous présentons l'algorithme de K-means quantique (QK-means) qui est adaptable aux trois types de distances présentées précédemment.

Supposons que nous avons un ensemble d'états quantiques $|X\rangle = \{|x_n\rangle \in \mathbb{C}^M, n = 1, ..., N\}$ et un ensemble de K clusters C_k, $|C_k|$ est le nombre de vecteurs dans le cluster k. La classification K-means a pour but de partitionner les N observations en K clusters C_k avec $|W\rangle = |w_1\rangle, |w_2\rangle, ..., |w_K\rangle$ centroids, afin de minimiser la variance intra-cluster. Formellement, l'objectif est de trouver :

$$\underset{W}{\operatorname{argmin}} D(|x\rangle, |w\rangle) = \underset{C}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{|x_n\rangle \in C_k} d^2_{q_i}(|x_n\rangle, |w_k\rangle) \tag{8}$$

Nous calculons la distance entre chaque état d'entraînement et chaque centre de cluster à l'aide du circuit swap test figure 1. Ensuite, nous affectons chaque état au plus proche centroïde en utilisant l'algorithme de Grover. Cet algorithme permet de rechercher un ou plusieurs éléments dans une base de données non triée avec N entrées dans un temps $O(\sqrt{N})$.

La deuxième étape de QK-means consiste à mettre à jour le centroïde de chaque cluster. La mise à jour du centroïde de chaque cluster est donnée par :

$$|w_k^{(t+1)}\rangle = |(Y_k^{(t)})^T X\rangle$$

avec

$$|Y_k^{(t)}\rangle = \frac{1}{\sqrt{|C_k|}} \sum_{n=1}^{N} y_{nk} |n\rangle$$

et

$$y_{nk} = \begin{cases} 1 & \text{if } x_n \in C_k \\ 0 & \text{else} \end{cases}$$

3.2 Algorithme d'apprentissage quantique

Nous donnons ci-dessous les principales étapes de l'algorithme proposé. La distance $d_{q_i}(|x_n\rangle, |w_k\rangle)$ est au choix de l'utilisateur. Dans notre cas, nous optons pour la distance d_{q_1} car elle donne les meilleurs résultats (voir l'équation 1).

Algorithme 1 : Algorithme K-means quantique QK-means

Input : $|X\rangle = \{|x_n\rangle \in \mathbb{C}^M, n = 1, ..., N\}$, K nombre de clusters C_k, centroïdes initiaux aux clusters à $t = 0$: $|w_1^{(0)}\rangle, |w_2^{(0)}\rangle, ..., |w_K^{(0)}\rangle$.

Output : K clusters C_k.

repeat

Assignment step (clustering) : Chaque donnée est affectée au cluster qui a le plus proche centre en utilisant l'algorithme de Grover :

$$C_k^{(t)} \longleftarrow \{|x_n\rangle : d_{q_i}^2(|x_n\rangle, |w_k^{(t)}\rangle) \leq d_{q_i}^2(|x_n\rangle, |w_j^{(t)}\rangle), \forall j, 1 \leq j \leq K\}$$

où chaque $|x_n\rangle \in |X\rangle$ est affecté à un seul $C_k^{(t)}$.

Update step : Le centre de chaque cluster C_k est recalculé comme étant la moyenne de toutes les données appartenant à ce cluster :

$$|w_k^{(t+1)}\rangle \longleftarrow |(Y_k^{(t)})^T X\rangle$$

until Convergence est atteinte

3.3 Critère de validation

Comme critère de validation, nous utilisons l'indice de Davies-Bouldin (DB) et l'indice quantique de Davies-Bouldin (QDB) (Benlamine et al., 2019). Vu que l'estimation de la distance dans la version classique n'est pas identique à la version quantique.

L'indice de Davies Bouldin (Davies et Bouldin, 1979) peut être calculé avec l'expression suivante :

$$DB = \frac{1}{K} \sum_{k=1}^{K} \max_{k \neq k'} \frac{d_n(w_k) + d_n(w_{k'})}{d(w_k, w_{k'})} \tag{9}$$

où K est le nombre de clusters, d_n est la distance moyenne de tous les éléments du cluster C_k à leur centre de cluster w_k, $d(w_k, w_{k'})$ est la distance entre les centres de clusters w_k et $w_{k'}$. Plus la valeur de DB est faible, plus le clustering est meilleur.

Comme nous l'avons déjà mentionné, la notion de la distance dans les approches quantiques est différente du cas classique. La distance quantique n'a pas besoin d'être proportionnelle à la distance réelle, elle doit juste être en corrélation positive avec elle. Nous n'avons besoin que du plus proche centroïde et non de la valeur exacte de la distance réelle. Pour évaluer la qualité du clustering quantique avec un type d'indice de Davies-Bouldin basé sur les

distances intra et inter-cluster, nous proposons de l'adapter au cas quantique. Pour ce faire, nous définirons l'indice de qualité Quantum Davies-Bouldin (QDB) comme suit :

$$QDB = \frac{1}{K}\sum_{k=1}^{K}\max_{k\neq k'}\frac{\delta_n(w_k)+\delta_n(w_{k'})}{\delta(w_k,w_{k'})} \tag{10}$$

avec

$$\delta_n(w_k) = \frac{1}{|C_k|}\sum_{i=1}^{|C_k|} d_{q_1}(|x_i\rangle_{x_i\in C_k}, |w_k\rangle)$$

et

$$\delta(w_k,w_{k'}) = d_{q_1}(|w_k\rangle, |w_{k'}\rangle)$$

4 Résultats expérimentaux

4.1 Jeux de données

Les versions classique et quantique de K-means ont été testées sur trois jeux de données réelles disponibles pour un usage public dans le référentiel UCI Machine Learning (Dua et Graff, 2017). Le jeu de données Iris contient 3 classes de 50 instances chacune, chaque classe faisant référence à un type de la fleur iris. Wine est un ensemble de données lié à une analyse chimique de vins cultivés dans la même région en Italie, mais issus de cultivars différents. Les données Breast cancer contiennent 569 instances avec 32 variables (ID, diagnostic, 30 variables d'entrée réelles). Chaque observation de données est étiquetée comme bénigne (357) ou maligne (212).

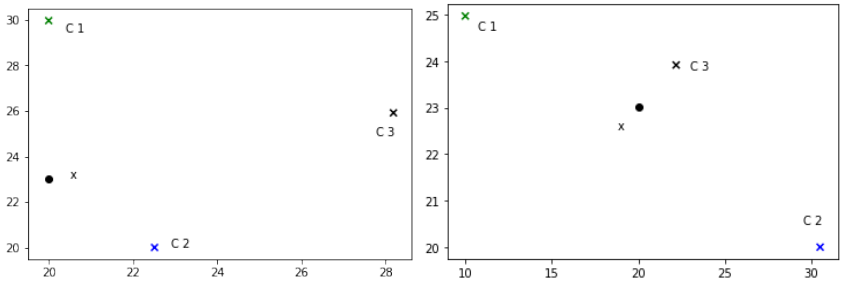
4.2 Comparaison entre les différentes distances quantiques

4.2.1 Evaluation de la distance quantique en terme de stabilité des valeurs du centre le plus proche

En raison de la nature probabiliste des qubits, la distance entre deux états est difficile à calculer car le résultat sera probabiliste ; la distance est instable. Cependant, il est plus facile d'attribuer des points de données à différents groupes car nous n'avons pas besoin des distances exactes pour chacun d'eux, mais seulement le plus proche du centroid. Ainsi, nous pouvons simplement mettre le nouveau point de données dans le cluster associé à la plus petite valeur prise par notre paramètre. Pour illustrer davantage notre idée, nous avons donné un exemple de deux distributions. La figure 2 représente deux distributions où le point noir X représente les données de test et les trois croix sont les centres ($C1$, $C2$, $C3$).

Vu l'instabilité de la distance quantique, on a effectué plusieurs itération dans chacune des méthodes afin de calculer la distance. Le résultat était différent d'une itération à l'autre mais l'affectation au plus proche centroide était correcte vu que c'était celle qui a donné la probabilité la plus élevée.

Le tableau 1 illustre bien que la distance change d'une itération à l'autre, mais l'affectation au plus proche centroïde est correcte.



(a) Distribution 1 (b) Distribution 2

FIG. 2: Deux Distributions avec X représente les données de test et les trois croix sont les centres

Dis dq_i	D	Vert (fois)	Bleu (fois)	Noir (fois)	Probabilité de succès $(\%)$	Temps d'exécution (sec.)
d_{q_1}	1	665	9322	13	$[92.71, 93.70]$	1644.26
	2	0	0	10000	$[99.96, 100]$	
d_{q_2}	1	2190	7620	190	$[75.35, 77.02]$	1993.33
	2	0	515	9485	$[94.40, 95.26]$	
d_{q_3}	1	2722	6530	748	$[64.36, 66.22]$	1808.72
	2	2486	2175	5339	$[52.41, 54.36]$	

TAB. 1: Comparison des différentes distances

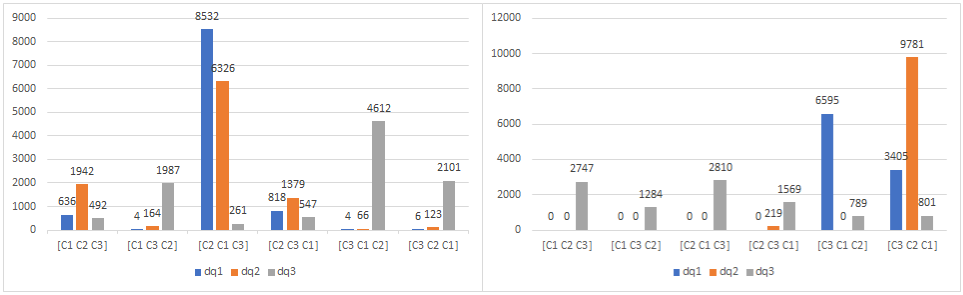
L'approche Wiebe et al. donne un intervalle de confiance plus élevé dans un temps inférieur aux autres approches.

Où : D est la distribution. Dis est la distance avec d_{q_1} est la distance de la première approche Wiebe et al, d_{q_2} est la distance de la deuxième approche Lloyd et al, et d_{q_3} est la distance de la troisième approche Anagolum. Le temps d'exécution est en secondes.

4.2.2 Evaluation de la distance quantique en termes de stabilité de l'ordre du centre le plus proche

Après avoir analysé les performances des différentes distances quantiques en termes de stabilité des valeurs permettant le choix du bon centre, nous étudierons le comportement de ces distances quantiques, mais cette fois en terme de stabilité de l'ordre des centres les plus proches. En d'autres termes, à quelle distance est-il possible de trouver les centres les plus proches dans le bon ordre quelle que soit l'itération ? Pour ce faire, nous avons effectué 10 000 recherches pour les centres les plus proches pour les deux distributions étudiées. Nous avons analysé la stabilité de l'ordre des centres les plus proches trouvés pour chaque distance quantique. Les résultats montrent que la distance d_{q_1} est la meilleure qui offre une très bonne stabilité dans l'ordre des centres les plus proches dans le cas des deux distributions étudiées.

Comme le montre le tableau 1, la distance d_{q_1} présente une très bonne stabilité dans l'ordre des centres les plus proches par rapport aux deux autres distances quantiques. Pour la distribution 1, le bon ordre des centres les plus proches est [$C2$ $C1$ $C3$]. La distance d_{q_1} trouve cet ordre avec une probabilité de 85,32% (8532 fois sur 10 000 recherches), tandis que la distance d_{q_2} et d_{q_3} trouvent cet ordre avec une probabilité de 53,26% (5326/10 000) et 26,10% (261/10 000) respectivement. Dans le cas de la distribution 2, la situation est plus compliquée car le point de test est presque à mi-chemin entre deux centres. Cette situation est confirmée par les résultats obtenus à la figure 3. En effet, la distance d_{q_1} trouve toujours le bon ordre des centres les plus proches [$C3$ $C1$ $C2$]. Néanmoins, cette distance continue de fournir la bonne solution, mais l'ordre change considérablement [$C3$ $C2$ $C1$]. Comparée aux deux autres distances quantiques, la distance d_{q_1} semble beaucoup plus stable dans l'ordre des centres les plus proches. Comme on peut le voir dans la figure 3, les deux autres distances quantiques d_{q_2} et d_{q_3} changent assez souvent l'ordre par rapport à la distance d_{q_1}. La stabilité des commandes est une information très pertinente sur le comportement des distances quantiques.



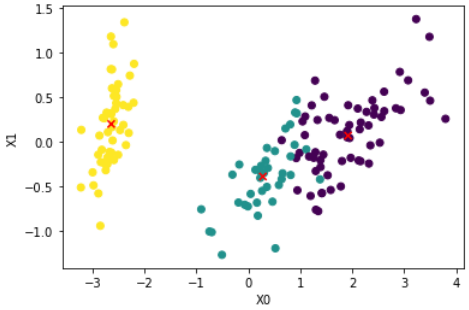
(a) Stabilité de la distribution 1 (b) Stabilité de la distribution 2

FIG. 3: Stabilité des trois distances d_{q_1}, d_{q_2} et d_{q_3}

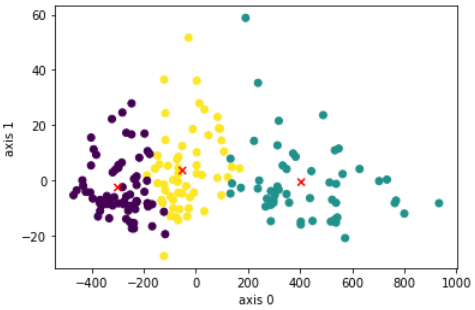
4.3 K-means quantique

Nous avons utilisé trois différents jeux de données pour montrer les résultats expérimentaux de QK-means. La figure 4 représentent la projection de jeux de données sur Iris, Wine et Breast cancer respectivement, en utilisant l'analyse en composantes principales. Nous pouvons noter que l'algorithme QK-means a identifié les différents groupes qui sont significativement différents (distants) les uns des autres. Par conséquent, le K-means quantique donne une bonne classification, tout comme sa version classique, mais l'avantage de la version quantique est qu'elle peut traiter des espaces de grandes dimensions dans un temps beaucoup plus rapide que la version classique, ce qui est crucial de nos jours.

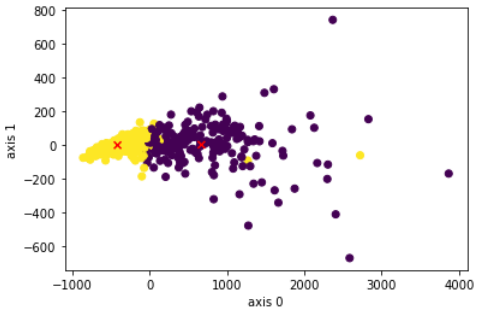
Pour chaque jeu de données, nous comparons l'indice de Davies-Bouldin (DB) pour la version classique et quantique de K-means. Ces résultats sont représentés dans le tableau 2. Les indices DB et QDB ne sont pas calculés aux mêmes distances. La comparaison directe est donc difficile, mais nous pouvons voir que QDB présente un comportement décroissant au cours des différentes itérations du processus d'apprentissage, ce qui indique une amélioration



(a) QK-means clustering de Iris data



(b) QK-means clustering de Wine data



(c) QK-means clustering de Breast Cancer data

FIG. 4: QK-means clustering

de la qualité de la classification quantique. Nous pouvons donc considérer que la QDB est un indicateur de bonne qualité pour le clustering quantique.

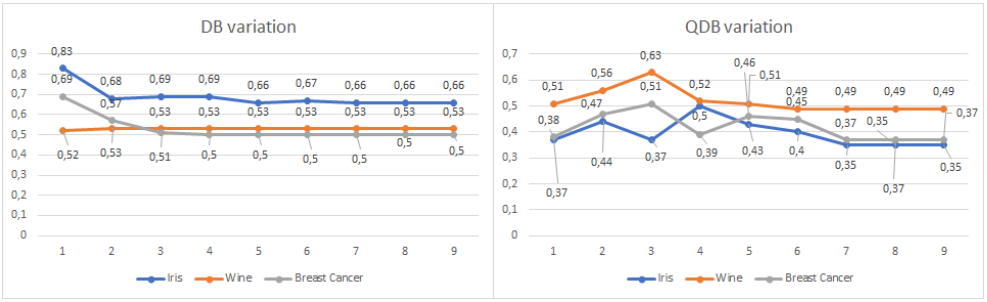
5 Conclusion

Dans cet article, nous avons implémenté un nouvel algorithme quantique de clustering à savoir K-means quantique. Nous avons analysé trois différentes méthodes pour estimer la distance pour des algorithmes de classification basés sur des prototypes quantiques. A travers cette analyse, nous avons constaté que la notion de distance en informatique quantique est différente de celle du classique, car ce qui compte dans le calcul quantique, c'est plutôt la corrélation et non pas la distance exacte. Cette analyse est très cruciale car elle peut résoudre n'importe quel algorithme de classification basé sur des prototypes.

Pour mesurer la qualité de la classification, nous avons adapté un critère classique au cas quantique. Cette version quantique de K-means a donné une bonne classification, tout comme sa version classique, la seule différence est sa complexité ; alors que la version classique de K-means prend un temps polynomial, la version quantique ne prend qu'un temps logarith-

Dataset	K-means (DB)	QK-means (QDB)
Iris	0.66	$[0.37, 0.56]$
wine	0.53	$[0.40, 0.59]$
Breast Cancer	0.50	$[0.38, 0.57]$

TAB. 2: K-means & QK-means en utilisant l'indice de DB



(a) (b)

FIG. 5: Variation de QDavies-Bouldin

mique, en particulier dans les grandes dimensions. Comme extension de cette proposition, nous travaillons sur une version quantique d'un autre algorithme de clustering topologique à deux niveaux incluant une sélection de modèle, proposé par (Cabanes et Bennani, 2007).

Références

Aïmeur, E., G. Brassard, et S. Gambs (2006). Machine learning in a quantum world. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pp. 431–442. Springer.

Anagolum, S. (2019). Quantum machine learning : distance estimation for k-means clustering. *https ://towardsdatascience.com/quantum-machine-learning-distance-estimation-for-k-means-clustering-26bccfbfcc76*.

Benlamine, K., Y. Bennani, A. Zaiou, M. Hibti, B. Matei, et N. Grozavu (2019). Distance estimation for quantum prototypes based clustering. In *International Conference on Neural Information Processing*, pp. 561–572. Springer.

Cabanes, G. et Y. Bennani (2007). A simultaneous two-level clustering algorithm for automatic model selection. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pp. 316–321.

Cai, X.-D., D. Wu, Z.-E. Su, M.-C. Chen, X.-L. Wang, L. Li, N.-L. Liu, C.-Y. Lu, et J.-W. Pan (2015). Entanglement-based machine learning on a quantum computer. *Physical review letters 114*(11), 110504.

Davies, D. L. et D. W. Bouldin (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* (2), 224–227.

Dua, D. et C. Graff (2017). UCI machine learning repository.

Gupta, S. et R. Zia (2001). Quantum neural networks. *Journal of Computer and System Sciences 63*(3), 355–383.

Lloyd, S., M. Mohseni, et P. Rebentrost (2013). Quantum algorithms for supervised and unsupervised machine learning. *arXiv preprint arXiv :1307.0411*.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Schützhold, R. (2003). Pattern recognition on a quantum computer. *Physical Review A 67*(6), 062311.

Wiebe, N., A. Kapoor, et K. M. Svore (2018). Quantum nearest-neighbor algorithms for machine learning. *Quantum Information and Computation 15*.

Summary

Quantum machine learning is a new area of research with the recent work on quantum versions of supervised and unsupervised algorithms. In recent years, many quantum machine learning algorithms have been proposed providing a speed-up over the classical algorithms. In this paper, we propose an analysis and a comparison of three quantum distances for protoptypes-based clustering techniques. As an application of this work, we present a quantum K-means version which gives a good classification just like its classical version, the difference resides in the complexity: while the classical version of K-means takes polynomial time, the quantum version takes only logarithmic time especially in large datasets. Finally, we validate the benefits of the proposed approach by performing a series of empirical evaluations regarding the quantum distance estimation and its behavior versus the stability of finding the nearest centers in the right order.

Keywords: Quantum machine learning, K-means, Unsupervised learning, Prototypes based Clustering.

Une méthode de classification ascendante hiérarchique par compromis : hclustcompro

Lise Bellanger *, Arthur Coulon**, Philippe Husi**

* Université de Nantes Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, 2 rue de la Houssinière
BP 92208, 44322 Nantes Cedex 03, France
lise.bellanger@univ-nantes.fr
http://www.math.sciences.univ-nantes.fr/~bellanger/
** CNRS/Université de Tours, UMR 7324 CITERES, Laboratoire Archéologie et Territoires,
40 rue James Watt, ActiCampus 1, 37200 Tours, France

Résumé. Les méthodes d'apprentissage semi-supervisé permettent d'utiliser des connaissances a priori pour guider l'algorithme de classification dans la découverte de groupes. Dans ce travail, nous proposons un nouvel algorithme de classification de type ascendante hiérarchique (CAH) prenant en compte deux sources d'information associées aux mêmes objets. Cette méthode appelée CAH par compromis (hclustcompro), permet un compromis entre les hiérarchies obtenues à partir de chaque source prise séparément. Une combinaison convexe des dissimilarités associées à chacune des sources est utilisée pour modifier la mesure de dissimilarité dans l'algorithme CAH classique. Le choix du paramètre de mélange est le point clé de la méthode. Nous proposons une fonction objectif à minimiser basée sur la différence absolue des corrélations entre dissimilarités initiales et distances cophénétiques, ainsi qu'une procédure de rééchantillonnage pour assurer la robustesse du choix du paramètre de mélange. Nous illustrons notre méthode avec des données archéologiques provenant du site d'Angkor Thom au Cambodge.

Mots-clés : classification ascendante hiérarchique, apprentissage semi-supervisé, compromis, distance cophénétique, archéologie.

1. Introduction

Les problèmes de classification ou clustering peuvent être abordés à l'aide d'une grande variété de méthodes, toutes nécessitant des techniques dédiées pour la phase de prétraitement des données. Il existe une littérature abondante sur le sujet de la classification, voir par exemple (Aggarwal et Reddy, 2014 ; Everitt et al., 2001 ; Kaufman et Rousseeuw, 2005 ; Nakache et Confais, 2005). Les deux algorithmes les plus connus sont la classification par partition et la classification hiérarchique. Dans cet article, nous nous concentrons sur la classification

ascendante hiérarchique (CAH) dont le principe fondamental consiste à construire une structure de données basée sur un arbre binaire appelé le dendrogramme. Quand il existe une information a priori sur les relations entre les individus (e.g. relation de voisinage spatial, relation d'ordre (en génomique), ...), l'utilisation de méthodes de classification avec contraintes (Ferligoj et al., 1982 ; Legendre et al., 1985 ; Chavent et al., 2018) ou avec prior knowledge (Ma et Dhavala, 2018) permet d'en tenir compte pour construire les groupes. Quand plusieurs sources de données existent, l'utilisation de méthodes de classification par consensus permet d'agréger l'information (Hulot et al, 2020).

A l'origine de la méthode présentée dans cet article, se trouve un problème récurrent chez les archéologues abordé ici dans le cadre d'un projet interdisciplinaire "ModAThom" (projet ANR, 2018-2022) Modèle explicatif de la fabrique d'Angkor Thom : archéologie d'une ville capitale disparue. Il s'agit de disposer d'un outil statistique d'aide à la périodisation d'ensembles stratigraphiques (niveaux d'occupation, dépotoirs, destruction d'un bâtiment…) pouvant provenir du même site ou de plusieurs sites d'une même ville, parfois spatialement distant donc sans connexion les uns avec les autres. Le résultat prend la forme d'un diagramme des ensembles périodisés jalonnant l'histoire du site en fonction de la proximité temporelle des ensembles. La méthode de classification développée avait donc initialement été appelée perioclust (Bellanger et al., 2021a). Mais après discussion avec Gilbert Saporta (PR Emérite CNAM Paris), il s'est avéré plus juste de parler de CAH par compromis et non avec contraintes. De plus, cette méthode a depuis été mise en œuvre sur des données médicales (Bellanger et al., 2021b). Nous avons donc décidé de changer son nom et opté pour un nom plus générique d'où hclustcompro. Cette CAH par compromis est une procédure d'apprentissage semi-supervisé, conçue pour prendre en compte deux sources d'information associées aux mêmes observations et potentiellement sujettes à des erreurs. Une approche basée sur la distance est adoptée pour modifier la mesure de distance dans l'algorithme CAH classique en utilisant une combinaison convexe des deux matrices de dissimilarités initiales. Le choix du paramètre de mélange est donc le point-clé. Nous définissons un critère de sélection de ce paramètre basé sur les distances cophénétiques, ainsi qu'une procédure de rééchantillonnage pour décider du choix du paramètre de mélange dans la méthode de classification proposée.

Cet article est organisé comme suit. Dans la section 2, nous décrivons les méthodes existantes tenant compte d'informations a priori. Dans la section 3, nous présentons l'algorithme proposé. Dans la section 4, nous illustrons notre approche sur un jeu de données archéologiques provenant du site d'Angkor Thom au Cambodge.

2. Bref aperçu des méthodes CAH tenant compte d'informations a priori

Les méthodes d'apprentissage semi-supervisé permettent d'utiliser des connaissances a priori pour guider l'algorithme de classification dans la découverte de groupes. Quand il existe une information sur les relations entre les objets (e.g. relation de voisinage spatial, relation d'ordre (en génomique), ...), l'utilisation de méthodes de classification *avec contraintes* permet d'en tenir compte pour construire les groupes. Dans le cas de plusieurs sources de données, l'utilisation de méthodes de classification par *consensus* permet d'agréger l'information.

En français, une contrainte définit une règle qui impose un certain comportement. De la même manière, la classification avec contraintes est une classe d'algorithmes d'apprentissage semi-supervisé qui diffère de son homologue sans contraintes en ce sens que les seuls groupes admissibles sont ceux qui respectent plus ou moins strictement la(es) relation(s). Nous n'évoquons ici que celles appelées "Instance Level contraints" (IL) (Davidson et Basu, 2007 ; Struyf, J. et S. Džeroski,2007), spécifiant des règles sur les objets qui peuvent ou non appartenir au même groupe. Les contraintes de type IL ont été incorporées avec succès à l'algorithme CAH (Davidson et Ravi, 2005). Il existe deux grandes approches : (i) celles dans lesquelles l'algorithme de classification est modifié pour intégrer les contraintes, (ii) celles dans lesquelles seule la dissimilarité est modifiée dans l'algorithme de classification.

Dans l'approche basée sur l'intégration des contraintes dans l'algorithme, les méthodes CAH basée sur la formule de Lance et Williams (Lance et Williams, 1967) sont facilement modifiables pour intégrer la contrainte. Les algorithmes de classification avec contrainte temporelle (ou spatiale) doivent indiquer sans ambiguïté quels sont les objets voisins. La solution la plus courante pour la classification avec contrainte de contiguïté est d'utiliser des schémas de connexion simples (voir par exemple Legendre et Legendre, 2012, Ferligo et Batagelj, 1982). Cette approche présente quelques inconvénients : (i) elle peut occasionnellement produire des inversions dans le dendrogramme, sauf dans le cas du critère du diamètre (Ferligoj et Batagelj, 1982), (ii) elle ne considère généralement que les dissimilarités entre objets liés, ce qui peut être trop restrictif dans certains domaines comme l'archéologie, comme nous le verrons plus loin. Une approche basée sur la dissimilarité, adaptée aux contraintes de proximité géographique, proposée par Chavent et al. (2018), consiste à modifier la dissimilarité dans l'algorithme CAH. Les contraintes géographiques sont intégrées à travers deux matrices de dissimilarités et un paramètre de mélange. Cette procédure a l'avantage d'être basée sur la dissimilarité et un critère d'hétérogénéité à minimiser à chaque étape pour construire le dendrogramme. Cependant, elle se fonde sur la stratégie d'agrégation de Ward qui ne convient pas à tous les types d'objets et le choix du paramètre de mélange n'est pas toujours évident.

Une autre approche (Ma et Dhavala, 2018), basée sur la dissimilarité, consiste à intégrer les connaissances a priori en combinant deux dissimilarités (celle associée aux données originelles et une distance ultramétrique relative aux connaissances a priori). Pour un nombre de groupes fixé, le paramètre de mélange peut être obtenu en maximisant une mesure de stabilité de la partition telles que l'indice de Davies-Bouldin ou l'indice de Dunn. Cette approche présente les inconvénients suivants : (i) les auteurs ne la présentent que pour la CAH avec lien simple et le cas de l'intégration de connaissance ontologique, (ii) le choix du paramètre et du nombre de groupes se fait simultanément à l'aide d'un critère de stabilité de la partition.

Enfin, la CAH par consensus (Hulot et al., 2020) conduit à regrouper un ensemble d'arbres ayant les mêmes feuilles pour créer un arbre consensus. Dans l'arbre consensus, un groupe à la hauteur h contient les objets qui sont dans le même groupe pour tous les arbres à la hauteur h. Le principal avantage de cette méthode est de pouvoir travailler avec plus de deux sources d'information. Cependant, par définition d'un consensus, la méthode construit un arbre correspondant à un accord ou consentement du plus grand nombre ; ce qui n'est pas toujours le but recherché.

Dans ce travail, nous proposons une approche CAH, appelée CAH par compromis, pour tenir compte des informations disponibles pour deux sources. Notre approche reprend, en

l'adaptant, l'idée présente dans certaines méthodes citées précédemment (Chavent et al., 2018; Ma et Dhavala, 2018) de déterminer une combinaison convexe associée à chacune des sources pour construire le dendrogramme.

3. CAH par compromis : hclustcompro

Tout d'abord, rappelons qu'en français, le compromis se définit comme une action qui implique des concessions réciproques. En ce sens, la méthode de classification semi-supervisée hclustcompro peut être vue comme un compromis entre deux CAH obtenues à l'aide de deux sources d'information.

3.1 Une approche basée sur la dissimilarité

Considérons un ensemble de n objets et notons $\mathbf{D}_1$ (resp. $\mathbf{D}_2$) la matrice de dissimilarités normalisée[1] $n \times n$ associée à la première (resp. deuxième) source d'information. Comme décrit dans Bellanger et al. (2020), le principe de hclustcompro est d'appliquer une méthode CAH à la combinaison convexe suivante :

$$\mathbf{D}_\alpha = \alpha \mathbf{D}_1 + (1-\alpha)\mathbf{D}_2 \quad (1)$$

où $\alpha \in [0;1]$ est un paramètre fixé qui pondère chaque matrice de dissimilarité (Eq. 1).

Lorsque $\alpha = 0$ (resp. $\alpha = 1$), les dissimilarités obtenues à partir de la matrice de dissimilarités $\mathbf{D}_1$ (resp. $\mathbf{D}_2$) ne sont pas prises en compte dans le processus de classification hiérarchique. Une fois α fixé, le dendrogramme de la CAH peut être construit à l'aide d'une des stratégies d'agrégation satisfaisant la formulation de Lance et Williams. Ainsi, le point clé de cette approche est le choix de α. La détermination de α dépend d'un critère conçu dans l'esprit de la corrélation cophénétique proposée par Sokal et Rohlf (1962). La corrélation cophénétique fait appel à la notion de matrice cophénétique, matrice dont les éléments sont les niveaux de dissimilarité auxquels les objets deviennent membres du même groupe dans le dendrogramme. La corrélation cophénétique correspond à la corrélation linéaire de Pearson entre la matrice de dissimilarité de départ et la matrice cophénétique issue du dendrogramme. Elle permet de mesurer la fidélité avec laquelle un dendrogramme préserve les dissimilarités initiales (voir Sokal et Rohlf, 1962 ; Everitt et al., 2001). La détermination de α est basée sur l'optimisation de la fonction objectif suivante qui "équilibre" le poids de $\mathbf{D}_1$ et $\mathbf{D}_2$ dans la classification finale :

$$CorCrit_\alpha = \left| Cor\left(\mathbf{D}_\alpha^{coph}, \mathbf{D}_1\right) - Cor\left(\mathbf{D}_\alpha^{coph}, \mathbf{D}_2\right) \right| \quad (2)$$

où $\mathbf{D}_\alpha^{coph}$ est la matrice cophénétique obtenue à partir du dendrogramme issu de la CAH obtenue avec $\mathbf{D}_\alpha$, α fixé dans (Eq. 1). Le critère $CorCrit_\alpha$ dans (Eq. 2) représente donc la différence en valeur absolue entre deux corrélations, chacune mesurant la fidélité avec laquelle le dendrogramme obtenu avec $\mathbf{D}_\alpha^{coph}$ préserve les dissimilarités par paire entre les objets initiaux mesurées avec $\mathbf{D}_1$ (resp. $\mathbf{D}_2$).

[1] Les valeurs de dissimilarité sont comprises entre 0 et 1.

La valeur de α est déterminée à l'aide de la formule ci-après :

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}\, CorCrit_{\alpha} \quad (3)$$

$\hat{\alpha}$ dans (Eq. 3) peut s'interpréter comme celui conduisant à un dendrogramme représentant la CAH avec $\mathbf{D}_{\alpha}$ défini dans (Eq. 1) dans lequel la position relative des objets est un compromis entre les dissimilarités $\mathbf{D}_1$ et $\mathbf{D}_2$. La figure 1 ci-dessous illustre le processus d'estimation de α.

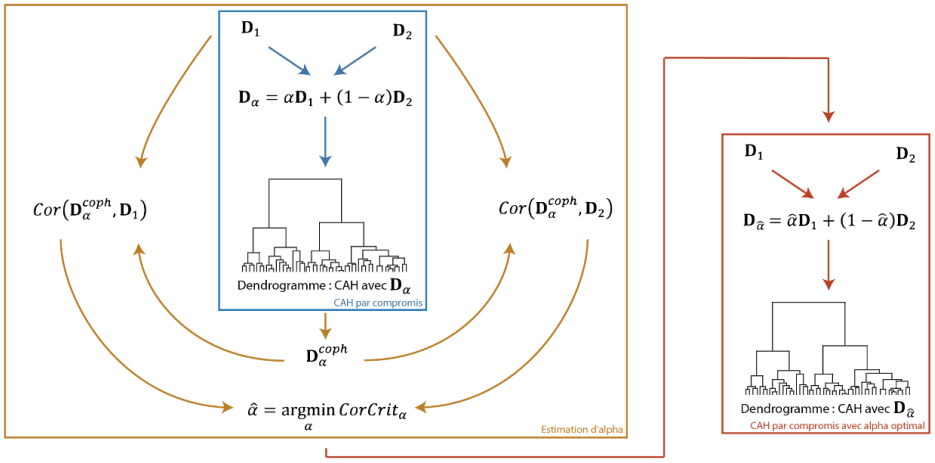


FIG. 1 – *Schéma illustratif de processus d'estimation de* α.

Cependant, $\hat{\alpha}$ est une estimation ponctuelle qui ne tient pas compte des erreurs potentielles dans le corpus de données, nous avons proposé une procédure de rééchantillonnage qui permet d'obtenir un intervalle de confiance pour α et d'étudier sa variabilité.

3.2 Stratégie de rééchantillonnage

La stratégie de rééchantillonnage proposée appelée *Add One In* est conçue dans le même esprit que celle du *Leave One Out* (Efron et Tibshirani, 1993) ; mais elle est basée sur l'ajout d'un "clone" aux objets existants plutôt que sur la suppression d'un objet comme dans le *Leave One Out*. Le principe de la méthode est présenté figure 2. Un clone c de l'observation $i \in \{1, \ldots, n\}$ est formé d'une copie de l'observation i de $\mathbf{D}_1$ et d'une copie de l'observation i' ($i' \neq i$) de $\mathbf{D}_2$ soit $n-1$ possibilités. Un total de $n(n-1)$ clones peuvent alors être créés. Les matrices de dissimarités $\mathbf{D}_1^{(c)}$ et $\mathbf{D}_2^{(c)}$ de dimension $(n+1) \times (n+1)$ représentent les dissimilarités pour un ensemble d'objets composé des objets originaux et du clone c fixé.

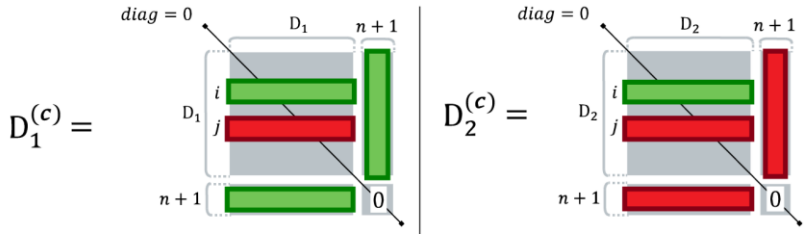


FIG. 2 – *Génération du clone c pour former* $\mathbf{D}_1^{(c)}$ *et* $\mathbf{D}_2^{(c)}$.

α est estimé à partir de l'éq. (2) en remplaçant $\mathbf{D}_1$ et $\mathbf{D}_2$ par $\mathbf{D}_1^{(c)}$ et $\mathbf{D}_2^{(c)}$ respectivement. Le paramètre $\hat{\alpha}$ peut être calculé pour chaque clone possible c afin d'estimer l'écart-type et d'obtenir un intervalle de confiance basé sur la méthode des percentiles pour α (voir algorithme 1). Une CAH peut alors être réalisée en utilisant $\mathbf{D}_{\hat{\alpha}}$ ou $\mathbf{D}_{\tilde{\alpha}}$ où $\tilde{\alpha}$ est au voisinage de $\hat{\alpha}$ dans le respect de l'intervalle de confiance $IC_{95\%}(\alpha)$.

ALGORITHME. 1 – *Add one in* procédure – Ecart-type estimé et $\boldsymbol{IC_{95\%}}$ par la méthode des percentiles pour α

Pour $c \in \{1, \dots, n(n-1)\}$ **faire**

- **Générer** un clone et créer $\mathbf{D}_{\mathbf{1}}^{(\boldsymbol{c})}$ et $\mathbf{D}_{\mathbf{2}}^{(\boldsymbol{c})}$
- **Définir** $\mathbf{D}_{\boldsymbol{\alpha}}^{(\boldsymbol{c})}$ et $CorCrit_{\alpha}^{(c)}$ où
  - $\mathbf{D}_{\boldsymbol{\alpha}}^{(\boldsymbol{c})} = (1-\alpha)\mathbf{D}_{\mathbf{1}}^{(\boldsymbol{c})} + \alpha\mathbf{D}_{\mathbf{2}}^{(\boldsymbol{c})}$ et
  - $CorCrit_{\alpha}^{(c)} = \left|Cor(\mathbf{D}_{\boldsymbol{\alpha}}^{\boldsymbol{coph}\,(\boldsymbol{c})}, \mathbf{D}_{\mathbf{1}}^{(\boldsymbol{c})}) - Cor(\mathbf{D}_{\boldsymbol{\alpha}}^{\boldsymbol{coph}\,(\boldsymbol{c})}, \mathbf{D}_{\mathbf{2}}^{(\boldsymbol{c})})\right|$
- **Evaluer** $\hat{\alpha}^{(c)} = \underset{\alpha\in[0;1]}{min}\ CorCrit_{\alpha}^{(c)}$; réplicat de $\hat{\alpha}$ pour chaque clone c

Fin

Obtenir:

- $\hat{\alpha}^* = \frac{1}{n(n-1)}\sum_{c=1}^{n(n-1)} \hat{\alpha}^{(c)}$, estimation ponctuelle de α à partir des des $\hat{\alpha}^{(c)}$;
- $\widehat{se}^* = \sqrt{\frac{\sum_{c=1}^{n(n-1)}\left(\hat{\alpha}^{(c)}-\hat{\alpha}^*\right)^2}{n(n-1)-1}}$, écart-type estimé de $se(\hat{\alpha}^*)$;
- Un intervalle de confiance (méthode des percentiles) $IC_{95\%}(\alpha)$ basé sur les réplicats.

Dans le cas où le nombre d'objets n est grand, la possibilité de ne pas calculer l'intervalle de confiance sur l'ensemble des $n(n-1)$ clones a été développée afin de réduire le temps de calcul. Il est donc possible de choisir un nombre x (fixé plus petit que $n-1$) de clones calculés pour l'observation i. Cette option permet de (i) réduire le nombre de possibilités de $n(n-1)$ clones à $n \times x$ clones, (ii) diminuer le temps de calcul à des temps raisonnables tout en conservant un nombre de clones suffisant.

Une CAH peut alors être réalisée en utilisant $\mathbf{D}_{\hat{\alpha}}$ ou $\mathbf{D}_{\tilde{\alpha}}$ où $\tilde{\alpha}$ est proche de $\hat{\alpha}$ et dans l'intervalle de confiance.

L'algorithme de CAH par compromis est implémenté dans la fonction hclustcompro du package R SPARTAAS (Coulon et al., 2021). Ce package accompagne la méthode avec un ensemble de fonctions permettant de sélectionner un α optimal, de couper l'arbre ou encore de subdiviser un cluster. De plus, une version Shiny est également disponible pour les utilisateurs occasionnels du logiciel R[2].

[2] http://www.r-project.org.

4. Résultats de la CAH par compromis sur des données archéologiques

Dans cette section, nous présentons les résultats obtenus avec la CAH par compromis (hclustcompro) sur des données issues de différents sites fouillés à Angkor Thom (Cambodge), capitale de l'empire khmer entre le IXe et le XVe s. (Gaucher, 2004). Nous comparons également les résultats à ceux obtenus avec une CAH classique pour interpréter l'apport de notre méthode dans le cas des données étudiées.

4.1 Les données archéologiques

L'un des objectifs majeurs ici est de préciser la périodisation de la ville, notamment à partir (i) du diagramme de sériation ou stratigraphique (Fig. 3) autrement appelé "matrice de Harris" (Harris, 1989) illustrant les relations physiques "sur/sous" donc chronologiques "avant/après" entre ensembles provenant de 3 sites archéologiques séparés (ii) des assemblages céramiques qui leurs sont associés (quantité de tessons par catégorie de céramique et par ensemble stratigraphique). La céramique (vaisselle en terre cuite) a l'avantage d'être indestructible, omniprésente dans les fouilles, avec des changements typologiques rapides dans le temps, ce qui en fait une des meilleures sources de datation en archéologie.

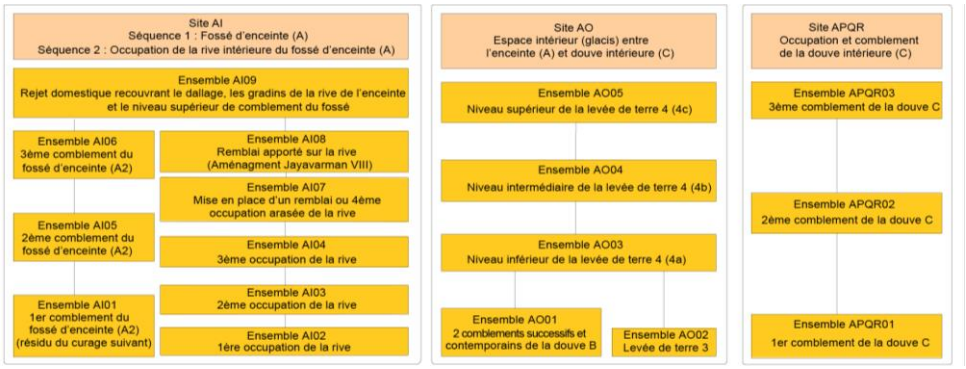


FIG. 3 – *Angkor : diagramme stratigraphique de trois sites archéologiques en relation avec le système d'enceinte de la ville.*

À partir du diagramme de sériation, il est donc possible de construire $\mathbf{S}_2$, la matrice symétrique d'adjacence définie comme une matrice binaire de connectivité, puis $\mathbf{D}_2 = \mathbf{1}_{17\times17} - \mathbf{S}_2$ associée aux 17 ensembles stratigraphiques (voir § 3.1). Les informations sur les céramiques sont contenues dans une table de contingence $\mathbf{N}$ de taille 17×12 où les lignes correspondent aux ensembles et les colonnes aux catégories céramiques. Comme très souvent sur ce type de données (Bellanger et Husi, 2012), l'analyse factorielle des correspondances (AFC) (Greenacre, 2016) sur $\mathbf{N}$ permet d'observer dans le plan factoriel 1-2 une forme parabolique dite "en fer à cheval" (effet Guttman) des projections des profils-lignes et colonnes (Fig. 4). Cette forme est révélatrice ici d'une évolution chronologique : l'ordre dans lequel se répartissent les catégories et les ensembles présente une séquence évolutive, mais d'autres facteurs peuvent entrer en ligne de compte. La CAH couplée à l'AFC est très souvent utilisée

pour définir des groupes d'ensembles. Cependant ces groupes ne tiennent pas compte de l'information sur la stratigraphe.

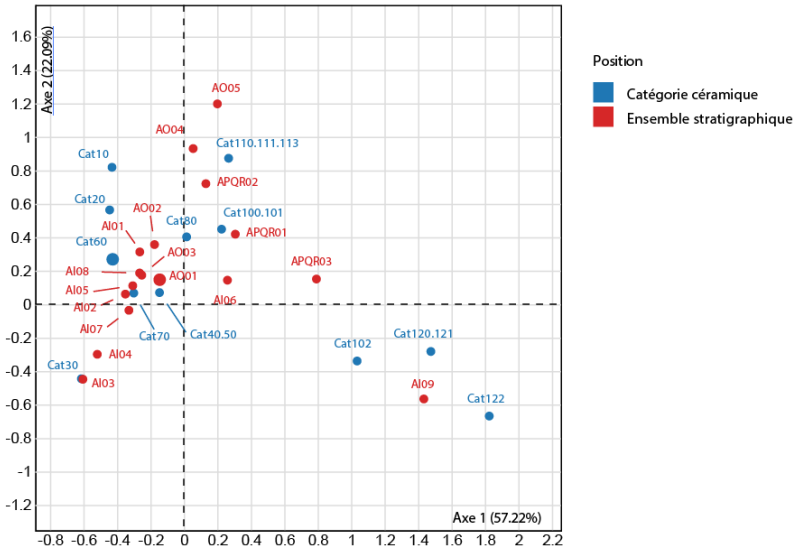


FIG. 4 – *Plan 1-2 de l'analyse factorielle des correspondances.*

D'où l'idée d'enrichir la construction et l'interprétation globale de la chronologie des 3 sites en combinant l'information sur la céramique à celle découlant du diagramme stratigraphique (Fig. 3) à l'aide d'une méthode de classification adaptée telle que hclustcompro. Les distances euclidiennes entre les ensembles sont calculées à partir de toutes les composantes des profils-lignes de l'AFC sur **N**. Les CAH sur $\mathbf{D}_1$ représentant la céramique et $\mathbf{D}_2$ représentant la stratigraphie conduisent aux valeurs les plus élevées du coefficient d'agglomération (Kaufman et Rousseeuw, 2005) pour le critère de Ward. Elle peut être considérée comme la meilleure stratégie d'agrégation à adopter pour ces données. Les dendrogrammes obtenus séparément à l'aide d'une CAH avec critère de Ward peuvent être comparés à l'aide du coefficient d'entanglement qui prend des valeurs comprises entre 0 et 1 ; une valeur faible traduisant de très grandes similitudes entre les 2 dendrogrammes. Dans notre cas, l'entanglement vaut 0.39 : les dendrogrammes sont relativement similaires, mais pas identiques. Cela confirme que les informations fournies par la céramique et la stratigraphie doivent être considérées simultanément pour résoudre le problème de classification.

4.2 Obtention d'une partition avec hclustcompro

Pour appliquer hclustcompro, nous définissons $\mathbf{D}_\alpha$ à partir de (Eq. 1) et déterminons un α optimal en utilisant (Eq. 3) avec un intervalle de confiance issu de la stratégie de rééchantillonnage (voir Sect. 3.2). Nous obtenons $IC_{95\%}(\alpha) = [0.55; 0.80]$ et choisissons $\hat{\alpha} = 0.7$ (Fig 5). Cette valeur indique que pour les données d'Angkor Thom, le poids de chaque source d'information est réparti comme suit : 70% pour la céramique et 30% pour la stratigraphie. Ce déséquilibre peut résulter d'une stratigraphie dont les limites ne sont pas

toujours clairement définies, conséquence de perturbations liées à l'importance des moussons, donc de l'eau au cours du temps.

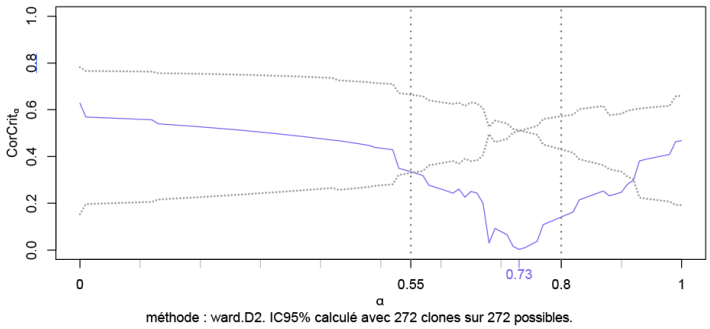


FIG. 5 – *$CorCrit_\alpha$ en fonction de α. $\hat{\alpha} = 0.73 \in [0.55: 0.8]$; $\tilde{\alpha} = 0.7$.*

Une CAH de Ward est effectué avec $\mathbf{D}_{0.7}$ comme défini dans (Eq. 1). Le nombre de groupes à retenir a été choisi en fonction de l'examen de l'échelle des indices d'agrégation associés au dendrogramme (Fig. 6) ; mais aussi en fonction de la connaissance du site par l'archéologue. En effet, le choix de 4 groupes avec le groupe D divisé en 3 sous-groupes (Fig. 6) semble le mieux adapté aux rythmes chronologiques de la ville.

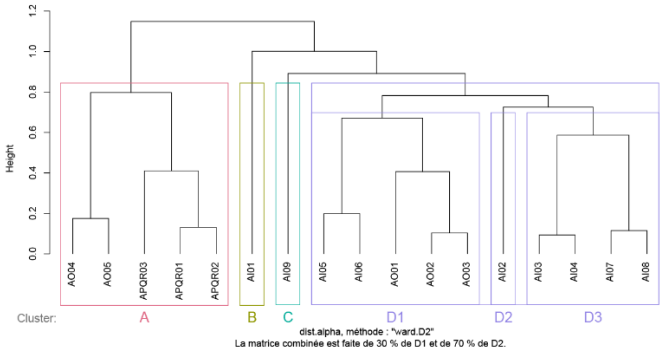


FIG. 6 – *Dendrogramme hclustcompro ($\alpha = 0.7$) critère de Ward : 4 groupes dont un subdivisé en 3.*

4.3 Comparaison entre CAH par compromis et CAH

La Fig.7 présente le dendrogramme issu de la CAH avec critère de Ward sur les données céramiques. L'absence de prise en compte de la stratigraphie rend ces résultats difficilement interprétables archéologiquement. En effet, des ensembles en relation physique (Fig. 3) situés entre deux autres peuvent se retrouver de manière incohérente dans un autre groupe, comme ici APQR02 isolé de APQR (01 et 03), cas de figure qui n'existe pas avec hclustcompro.

Une méthode de classification ascendante hiérarchique par compromis : hclustcompro

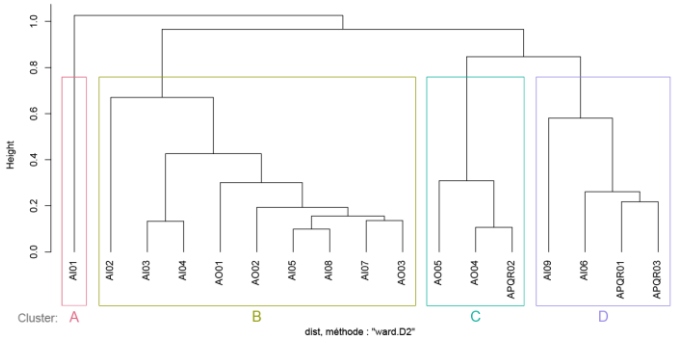


FIG. 7 – *Dendrogramme CAH critère de Ward.*

L'indice de Rand ajusté (Rand, 1971) montre une relative similitude (0.69) entre les deux partitions en quatre groupes. On observe malgré tout, des différences intéressantes (voir tableau 1).

Partitions		hclustcompro			
		A	B	C	D
hclust	A	0	1	0	0
	B	0	0	0	9
	C	3	0	0	0
	D	2	0	1	1

TAB. 1 – *Matrice de confusion entre les partitions CAH par compromis et CAH classique.*

Archéologiquement, trois ensembles sont pour certains bien plus anciens (AI01 et AI02) pour d'autres bien plus récents (AI09) que les autres caractérisant les principaux états de construction et d'occupation de l'enceinte. Dans la partition CAH classique, seul l'ensemble AI01 ressort clairement comme isolé des autres. Celle issue de la CAH par compromis traduisant mieux la réalité chronologique de l'histoire du site : elle permet d'identifier AI01, dans une moindre mesure AI02 mais surtout AI09 comme isolé des autres ensembles sachant que ce dernier est bien plus récent que ceux qui le précèdent. La CAH par compromis permet d'intégrer la source d'information stratigraphique pour construire une partition, là où d'autres méthodes de type CAH traiteraient cette information comme une contrainte trop forte.

5. Conclusions

Dans ce travail, nous avons présenté une nouvelle méthode de CAH basée sur un compromis entre deux sources d'information disponibles. Cette approche fondée sur une modification de la dissimilarité dans l'algorithme de CAH classique est simple à mettre en œuvre. La matrice de dissimilarités modifiée dans la CAH est une combinaison de deux matrices de dissimilarités, donc par construction tous les critères d'agrégation existants peuvent être utilisés. Les problèmes du choix et de l'interprétation du paramètre de mélange, points clés pour ce type de méthodes de classification, sont résolus. Le paramètre de mélange définit l'importance donnée à chaque source dans la procédure de classification. Bien que hclustcompro ait été conçue à l'origine pour répondre à un problème archéologique, cette

méthode présente un intérêt dans de nombreux autres domaines d'application comme par exemple la santé (Bellanger et al., 2021b).

La CAH par compromis trouve une résonnance toute particulière avec les méthodes factorielles d'analyse conjointe de plusieurs tableaux de données telle que STATIS qui recherche un tableau compromis le plus représentatif selon certains critères. Partant de ce constat, la perspective méthodologique naturelle est d'étendre notre méthode au cas de plus de deux sources d'information croisant les mêmes objets.

Remerciements Cette recherche a été soutenue en partie par le projet ANR ModAThom coordonné par Philippe Husi et Jacques Gaucher (EFEO). Les auteurs tiennent à remercier Jacques Gaucher pour ses commentaires et son expertise des données et Gilbert Saporta pour nous avoir suggéré la dénomination "CAH par compromis".

Références

Aggarwal, C. et C. Reddy (2014). *Data Clustering: Algorithms and Applications*. Boca Raton: Chapman and Hall/CRC.

Bellanger, L. et P. Husi (2012). Statistical Tool for Dating and interpreting archaeological contexts using pottery. *Journal of Archaeology Science* 39(4), 777-790.

Bellanger L., A. Coulon et P. Husi (2021a) PerioClust: A Simple Hierarchical Agglomerative Clustering Approach Including Constraints. In: Chadjipadelis T., Lausen B., Markos A., Lee T.R., Montanari A., Nugent R. (eds) *Data Analysis and Rationality in a Complex World*. IFCS 2019. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham. https://doi.org/10.1007/978-3-030-60104-1_1

Bellanger, L., L. Chevreuil, P. Drouin, D.A. Laplaud et A. Stamm (2021b). Peut-on détecter des troubles de la marche avant qu'ils ne soient perceptibles ? *Revue Tangente*, Hors-série Bib73 de la "Bibliothèque Tangente" sur Maths et emploi en entreprise.

Chavent, M., V. Kuentz-Simonet, A. Labenne et J. Saracco (2018). ClustGeo: an R package for hierarchical clustering with spatial constraints. *Computational Statistics* 33(4), 1799-1822.

Coulon, A., L. Bellanger et P. Husi (2021). SPARTAAS: Statistical Pattern Recognition and daTing using Archaeological Artefacts assemblageS. R package version 1.0.0. https://CRAN.R-project.org/package=SPARTAAS.

Davidson, I. et S. Ravi (2005). Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In: *9th European Conf. on Principles and Practice of Knowledge Discovery in Databases*, 59–70.

Davidson, I. et S. Basu (2007). A Survey of Clustering with Instance Level. *ACM Transactions on Knowledge Discovery from Data*, 1-41.

Efron, B. et R. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Everitt, B., S. Landau et L. Morven (2001). *Cluster Analysis*. 4th ed. ed. Oxford: Oxford University Press Inc.

Ferligoj, A. et V. Batagelj (1982). Clustering with relational constraint. *Psychometrika* 47(4), 413-426.

Gaucher, J. (2004). Angkor Thom, une utopie réalisée ? Structuration de l'espace et modèle indien d'urbanisme dans le Cambodge ancien. *Arts Asiatiques* 59, 58-86.

Greenacre, M. (2016). *Correspondence Analysis in Practice*. Boca Raton: Chapman & Hall/CRC.

Harris, E. C. (1989). *Principles of Archaeological Stratigraphy*. 2nd ed. ed. London and San Diego: Academic Press.

Hulot, A., J. Chiquet, F. Jaffrézic et al. (2020). Fast tree aggregation for consensus hierarchical clustering. *BMC Bioinformatics* 21, 120. https://doi.org/10.1186/s12859-020-3453-6

Kaufman, L. et P. Rousseeuw (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley-Interscience.

Lance, G. N. et W. T. Williams (1967). A general theory of classificatory sorting strategies ii.clustering systems. *The computer journal* 10(3), 271–277.

Legendre, P. et L. Legendre (2012). *Numerical ecology*. 3rd ed. Amsterdam: Elsevier Sc. BV.

Ma, X. et S. Dhavala (2018). Hierarchical clustering with prior knowledge. *arXiv:1806.0343*.

Nakache, -J.-P. et J. Confais (2005). *Approche pragmatique de la Classification*. Ed. Technip, Paris.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66 (336), 846–850. https://doi.org/10.2307/22842399

Sokal, R. R. et F.J. Rohlf (1962). The comparison of dendrograms by objective methods. *Taxon* XI (2), 33-40.

Struyf, J. et S. Džeroski (2007). Clustering Trees with Instance Level Constraints. In: Kok J.N., Koronacki J., Mantaras R.L.., Matwin S., Mladenič D., Skowron A. (eds) *Machine Learning: ECML 2007. Lecture Notes in Computer Science*, vol 4701. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74958-5_34

Summary

Semi-supervised learning methods allow using a priori knowledge to guide the classification algorithm in group discovery. In this work, we propose a new hierarchical agglomerative clustering algorithm (HAC) that takes into account two sources of information associated with the same objects. This method, called compromise HAC (hclustcompro), allows a compromise between the hierarchies obtained from each source taken separately. A convex combination of the dissimilarities associated with each of the sources is used to modify the dissimilarity measure in the classical HAC algorithm. The choice of the mixing parameter is the key point of the method. We propose an objective function to be minimized based on the absolute difference of correlations between initial dissimilarities and cophenetics distances, as well as a resampling procedure to ensure the robustness of the choice of the mixing parameter. We illustrate our method with archaeological data from the Angkor site in Cambodia.

Keywords: hierarchical agglomerative clustering, semi-supervised learning, compromise, cophenetic distance, archaeology.

Clustering collaboratif à partir de données et d'informations privilégiées

Yohan Foucade, Younès Bennani

LIPN UMR 7030 CNRS, Université Sorbonne Paris Nord, France
LaMSN, La Maison des Sciences Numériques, France
name.surname@sorbonne-paris-nord.fr

Résumé. L'objectif du clustering collaboratif est d'améliorer la performance de plusieurs algorithmes en combinant leurs résultats. Le but est que chaque algorithme, éventuellement appliqué à des jeux de données distincts, bénéficie des connaissances des autres collaborateurs. Cet article est consacré au clustering collaboratif basé sur le paradigme de l'apprentissage à partir d'informations privilégiées. Les algorithmes locaux pondèrent les informations entrantes au niveau de chaque observation, en fonction du niveau de confiance de la classification de cette observation. Une comparaison entre notre algorithme et les implémentations de l'état de l'art montre une amélioration des performances de l'ensemble des collaborateurs en utilisant l'approche proposée.

Mots-clés : apprentissage non-supervisé, clustering collaboratif, apprentissage distribué, apprentissage multi-modèles.

1 Introduction

Le clustering est une tâche commune d'analyse de données. Son objectif est le partitionnement d'observations en groupes homogènes (faible variabilité intra-groupe) et distincts (grande variabilité inter-groupe). Il existe une grande variété d'algorithmes de clustering et la plupart nécessitent que toutes les données soient disponibles au même endroit et au même moment (Jain et al., 1999; Cabanes et Bennani, 2007). Cependant, des raisons techniques, légales, de confidentialité ou de performance notamment, peuvent faire obstacle au rassemblement des données ou aux échanges d'informations sensibles.

Dans le cadre du clustering collaboratif, on dispose d'un jeu de données distribuée sur différents nœuds d'un réseau. L'objectif est d'obtenir sur chaque nœud un clustering des données locales en utilisant les résultats obtenus sur les nœuds distants, sans échanger les données elles-mêmes (Pedrycz, 2002; Cornuéjols et al., 2018). Cette opération s'effectue souvent en deux phases. Tout d'abord, la phase locale, au cours de laquelle un algorithme de clustering est appliqué localement et indépendamment sur chaque site. Puis la phase de collaboration, au cours de laquelle les sites échangent leurs connaissances afin d'essayer d'améliorer leurs propres résultats.

Le type d'informations que les algorithmes partagent diffère selon la façon dont les données sont distribuées. Lorsque les objets de différents sites se trouvent dans le même espace,

c'est-à-dire qu'ils sont décrits par les mêmes variables, alors les algorithmes ont la possibilité de partager des informations qui peuvent être représentées au sein de l'espace lui-même. Cette tâche est connue sous le nom de clustering collaboratif *vertical*. À l'inverse, dans le cas du clustering collaboratif *horizontal*, les sites s'accordent sur l'identité des objets, mais ces derniers sont décrits par des variables différentes. Ainsi, ils ne peuvent plus partager le même type d'information que dans le cas précédent. Ils conservent néanmoins la possibilité de comparer les clusters qu'ils ont créés, en utilisant par exemple leurs matrices de partition, cf. section 3.

Dans cet article consacré au clustering collaboratif horizontal, nos principales contributions sont les suivantes :

— La mise en œuvre d'une idée intuitive pour la collaboration : la pondération des résultats distants dépend du niveau de certitude sur les résultats locaux.

— Une collaboration simultanée entre tous les experts, qui permet de s'affranchir de la nécessité de choisir avec quel site collaborer à chaque étape.

— Une collaboration au niveau de chaque observation, permettant aux algorithmes d'affiner la façon dont ils collaborent avec les autres.

— La détection des exemples difficiles à classifier.

— L'amélioration des performances prédictives, en combinant différents systèmes d'apprentissage ayant chacun un biais inductif différent.

Notre travail est inspiré du paradigme d'apprentissage proposé par Vapnik *et al.*, appelé Learning Using Privileged Information (LUPI) (Vapnik et Vashist, 2009; Vapnik et Izmailov, 2015). Constatant qu'un professeur joue un rôle essentiel dans l'apprentissage humain (en plus des exemples, celui-ci fournit des explications, des commentaires, des comparaisons, etc.), Vapnik propose un paradigme d'apprentissage supervisé dans lequel, lors de la phase d'apprentissage, les apprenants reçoivent des informations supplémentaires x^* portant sur les données d'entraînement x.

Ce document est organisé comme suit : la section 2 est un tour d'horizon du domaine. Notre principale contribution est ensuite présentée dans la section 3. Enfin, nous discutons des résultats expérimentaux dans la section 5.

2 État de l'art

La recherche sur le clustering collaboratif a été introduite par Pedrycz en 2002 (Pedrycz, 2002). L'auteur a proposé une version collaborative de l'algorithme Fuzzy C-Means (Bezdek, 1981). La fonction objectif de l'algorithme Fuzzy C-Means y est étendue avec un second terme qui pénalise une trop grande différence entre les partitions.

La méthode SAMARAH introduite par (Wemmert et al., 2000) peut traiter des algorithmes de clustering "dur" hétérogènes. Cette méthode vise à augmenter la similarité entre les clusterings par l'évaluation et la résolution des conflits. Afin de comparer des partitions avec des nombres de clusters éventuellement différents, une matrice de confusion est calculée entre chaque paire d'algorithmes. On aboutit ensuite à un consensus en appliquant un algorithme de vote à ces résultats.

En 2015, Sublime *et al.* ont proposé une approche levant certaines limites des travaux précédents : ils introduisent un cadre pour la collaboration avec des algorithmes hétérogènes. Cette

méthode a l'avantage de ne pas nécessiter de coefficient de confiance global, ainsi l'impact d'un algorithme distant sur les résultats locaux peut être différent pour chaque observation. Cependant, l'utilisation des résultats de clustering dur pour la construction des matrices de confusion entraîne une certaine perte d'information. De plus, chaque algorithme a le même poids dans ce processus. En conséquence, puisqu'il n'est fait aucune différence entre un site de données avec des clusters bien séparés, et un autre avec des clusters qui se chevauchent, cette méthode laisse la possibilité aux algorithmes peu performants d'entraver les résultats des autres.

Après avoir proposé des versions collaboratives des Self Organizing Maps (SOM) et des Generative Topographic Mapping (GTM), (Ghassany et al., 2013) ont combiné les algorithmes Fuzzy C-Means et GTM pour obtenir un algorithme de clustering collaboratif.

Dans (Sublime et al., 2017), les auteurs décrivent un cadre de collaboration pour les algorithmes de clustering basés sur des modèles. Ils définissent une fonction de vraisemblance globale et tentent d'optimiser un proxy pour cette fonction. Ce processus nécessite de fixer un paramètre de poids entre les informations locales et externes.

Plus récemment, une méthode automatisée pour optimiser la confiance des échanges a été proposée par Sublime *et al.*, et la collaboration concernait quatre vues différentes. Les auteurs ont obtenu des résultats probants dans la détection des vues bruyantes, mais la méthode a tendance à favoriser la collaboration entre des vues très similaires.

Pour un état de l'art plus exhaustif sur le clustering collaboratif, le lecteur peut se référer à (Cornuéjols et al., 2018).

3 Cadre théorique et algorithme

Dans cette section, nous décrivons le cadre général dans lequel s'applique notre méthode, ainsi que notre algorithme.

3.1 Formalisation

L'objectif de l'apprentissage collaboratif est d'apprendre à partir de données locales et d'un ensemble d'apprenants dont les données sont stockées sur des sites distincts.

Plus formellement, supposons que l'ensemble de données globales X est réparti sur P sites :
$X^{[1]}, \ldots, X^{[p]}, \ldots, X^{[P]}$, où $X^{[p]} = \{x_i^{[p]}\}_{i=1}^{N}$ est un ensemble de N objets et chaque objet $x_i^{[p]} \in \mathbb{R}^d$ est un vecteur à d dimensions.

Notre étude est centrée sur le problème du clustering collaboratif horizontal. Ainsi, chaque site a accès à différentes caractéristiques $d^{[p]}$ décrivant les mêmes individus. Sur chaque site, un algorithme $\mathcal{A}^{[p]}$, $p = 1, \ldots, P$ proposera une partition des données locales $X^{[p]}$. Le processus comporte deux étapes : une étape locale et une étape collaborative.

3.2 Cadre fondamental

Étape locale Au cours de l'étape locale, chaque algorithme $\mathcal{A}^{[p]}$ travaille sur son propre jeu de données $X^{[p]}$ et ajuste ses paramètres comme il le ferait dans un cadre non collaboratif. Afin de pouvoir échanger leurs résultats, les algorithmes doivent disposer d'un type d'information en commun. Dans notre cas, la partition calculée par chaque algorithme sera représentée sous la

forme d'une matrice de responsabilité $R^{[p]}$. Cette matrice contient les contributions de chaque composante à chaque observation, $R^{[p]}_{i,k} = \mathbb{P}(Z_i = k|X_i, \theta)$, où K est le nombre de clusters, $Z_i \in \{1, \ldots, K\}$ sont les composantes du modèle et les éléments de θ sont ses paramètres.

Une fois que chaque modèle a été entraîné localement, nous voulons donc qu'ils échangent ces matrices de partition entre eux afin d'améliorer leurs performances. Cet échange se fait lors de la phase de collaboration.

Étape collaborative Lors de la phase de collaboration, les différents algorithmes $\mathcal{A}^{[p]}$ vont échanger des informations dans l'optique d'améliorer leur classification respective. Le clustering ainsi obtenu sera non seulement basé sur les exemples $X^{[p]}$, mais aussi sur les informations supplémentaires $X^{[p]^*}$, qui sont en fait les matrices de partition distantes. L'apprenant local a donc accès à $X^{[p]}$ et $X^{[p]^*}$, au lieu de seulement $X^{[p]}$, pour effectuer son clustering. Dans notre cas, pour chaque algorithme $\mathcal{A}^{[p]}$, $X^{[p]^*} \equiv R^{-[p]}$, où :

$$R^{-[p]} = \{R^{[q]} \ : \ q \in \{1, \ldots, P\} \setminus \{p\}\} \tag{1}$$

L'ensemble $R^{-[p]}$ contient toutes les matrices de partition distantes, du point de vue du site p. Dans ce cas, comment pouvons-nous utiliser $X^{[p]}$ et $X^{[p]^*}$ pour améliorer chaque apprenant local $\mathcal{A}^{[p]}$?

Par souci de simplicité, nous considérons deux apprenants (sites) : $P = 2$. A l'étape locale, ils produisent deux matrices de partition : $R^{[1]}(t)$ et $R^{[2]}(t)$. En particulier, pour l'échantillon x_i, nous avons $R^{[1]}_{i,\cdot}$ et $R^{[2]}_{i,\cdot}$. Ainsi, la règle de mise à jour de l'apprenant 1 pour l'échantillon x_i est :

$$R^{[1]}_{i,\cdot}(t+1) \longleftarrow f\left(R^{[1]}_{i,\cdot}(t), R^{[2]}_{i,\cdot}(t)\right) \tag{2}$$

Nous souhaitons que cette règle de mise à jour dépende du niveau de certitude du site 1 à propos de sa propre classification . Plus ce niveau de certitude est élevé, plus la différence entre $R^{[1]}_{i,\cdot}(t+1)$ et $R^{[1]}_{i,\cdot}(t)$ doit être faible. Inversement, plus ce niveau de certitude est faible, plus on accordera de poids à $R^{[2]}_{i,\cdot}(t)$. De même, on tiendra compte du niveau de certitude du site 2 à propos de sa classification : plus la certitude du site 2 est élevée (resp. faible), plus (resp. moins) elle influencera $R^{[1]}_{i,\cdot}(t+1)$.

Il s'avère que le cadre probabiliste est doté d'une mesure qui peut être interprétée comme la quantité d'incertitude dans une distribution, à savoir l'entropie. Elle est définie comme suit :

$$H(X) = -\sum_{i=1}^{d} p(x_i) \log_2 p(x_i) \tag{3}$$

Où X est une variable aléatoire avec des valeurs possibles $\{x_1, \ldots, x_d\}$. Il s'ensuit que l'incertitude pour une distribution R_i est :

$$H(R_{i,\cdot}) = -\sum_{k=1}^{K} R_{i,k} \log_2 R_{i,k} \tag{4}$$

Et sa version normalisée, en utilisant le fait qu'elle est positive et maximale lorsque X est uniformément distribué :

$$\mathcal{H}(R_{i,.}) = \frac{H(R_{i,.})}{log_2(K)} \tag{5}$$

Alors $0 \leq \mathcal{H}(R_{i,.}) \leq 1$, et

— $\mathcal{H}(R_{i,.}) \approx 0 \Rightarrow R_{i,.} \approx \mathbb{1}_{Z_i=k}$ pour $k \in 1, \ldots, K$, cela représente un niveau de confiance élevé dans la classification de l'observation i.

— $\mathcal{H}(R_{i,.}) \approx 1 \Rightarrow R_{i,.} \approx \frac{1}{K} \; \forall k \in 1, \ldots, K$, cela traduit une forte incertitude sur la classification de l'observation i.

L'équation 6 donne la règle de mise à jour de $R^{[ii]}$ lorsque P algorithmes sont impliqués dans la collaboration.

$$R^{[p]}(t+1) \longleftarrow \alpha^{[p]} \cdot R^{[p]}(t) + \sum_{R^{[q]}(t) \in R^{-[p]}} \beta^{[p]}_{[q]} \cdot R^{[q]}(t) \tag{6}$$

où

$$\begin{cases} \alpha^{[p]} = \left(\frac{1}{P-1} \sum_{R^{[q]}(t) \in R^{-[p]}} \mathcal{H}\big(R^{[q]}(t)\big) \right) \cdot \big(1 - \mathcal{H}\big(R^{[p]}(t)\big)\big) \\ \\ \beta^{[p]}_{[q]} = \mathcal{H}\left(R^{[p]}(t)\right) \cdot \big(1 - \mathcal{H}\left(R^{[q]}(t)\right)\big) \end{cases} \tag{7}$$

Dans ces équations, $\alpha^{[p]}$ est un vecteur de taille N. Chaque élément i de $\alpha^{[p]}$ est un poids associé à la classification locale de la i-ième observation. Ce poids dépend négativement de la quantité d'incertitude portée par la classification locale, et positivement de l'incertitude moyenne de la classification distante.

De même, les $\beta^{[p]}_{[q]}$ sont des vecteurs de taille N. Chaque élément i de $\beta^{[p]}_{[q]}$ est un poids associé à la classification de la i-ème observation par l'algorithme $\mathcal{A}^{[q]}$. Ce poids dépend positivement de la quantité d'incertitude portée par la classification locale, et négativement de l'incertitude de la classification à distance.

Notez que les sites de données ne partagent que leurs matrices de partition. Par conséquent, quels que soient les algorithmes sous-jacents, à condition qu'ils recherchent le même nombre de composantes, l'algorithme de collaboration reste pertinent. Cependant, les sites de données doivent s'accorder sur l'identité des clusters. Notre implémentation utilise l'algorithme hongrois pour réorganiser les clusters sur chaque site de données. Dans le reste de cette section, nous décrivons l'algorithme CoLUPI, puis nous montrons comment les valeurs de α et β peuvent être utilisées pour visualiser le flux d'information dans le processus de collaboration.

Algorithme d'apprentissage collaboratif Sur la base du formalisme théorique développé et présenté dans la section précédente, nous pouvons concevoir un algorithme d'apprentissage pour établir des échanges entre les différents sites à travers un processus collaboratif. Cet algorithme utilise donc l'équation 6 pour mettre à jour les paramètres des sites en interaction collaborative.

Algorithme d'apprentissage CoLUPI

Etape locale :
for all algorithmes $\mathcal{A}^{[p]}$ **do**
  entraîner $\mathcal{A}^{[p]}$ à partir de $X^{[p]}$
  obtenir les paramètres locaux $\theta^{[p]}$ et matrice de partition $R^{[p]}$
end for
Etape collaborative :
repeat
  for all algorithmes $\mathcal{A}^{[p]}$ **do**
    $R^{[p]}(t+1) \leftarrow f\left(R^{[p]}(t), R^{-[p]}(t)\right)$
    entraîner le modèle à partir de la nouvelle partition $R^{[p]}(t+1)$ et de $X^{[p]}$
    obtenir $\theta^{[p]}(t+1)$ et $R^{[p]}(t+1)$
    if qualité $\left(R^{[p]}(t+1)\right)$ est meilleure que qualité$\left(R^{[p]}(t)\right)$ **then**
      Accepter la collaboration
    else
      $\theta^{[p]}(t+1) \leftarrow \theta^{[p]}(t)$
      $R^{[p]}(t+1) \leftarrow R^{[p]}(t)$
    end if
  end for
until aucun algorithme ne s'améliore pour son critère

3.3 Visualisation du processus de collaboration

Dans cette section, nous nous intéressons à la visualisation du flux d'informations pendant la collaboration. Comme il a été mentionné ci-dessus, chaque collaborateur attribue un poids à chaque site source (y compris lui-même), pour chaque observation. La moyenne de ces poids pour chaque site distant nous donne le poids moyen attribué à ce site et peut être considérée comme un coefficient de confiance, bien qu'il ne soit pas uniforme entre les observations. S'il y a P sites, cela nous donne une matrice $P \times P$ telle que définie dans l'équation 8 qui peut être considérée comme une matrice de confiance. Nous représentons cette matrice sous forme de heatmap pour l'ensemble de données Wdbc dans la section 4.

$$C_{(P\times P)} = \begin{pmatrix} \overline{\alpha^{[1]}} & \overline{\beta^{[1]}_{[2]}} & \cdots & \overline{\beta^{[1]}_{[P]}} \\ \overline{\beta^{[2]}_{[1]}} & \overline{\alpha^{[2]}} & \cdots & \overline{\beta^{[2]}_{[P]}} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\beta^{[P]}_{[1]}} & \overline{\beta^{[P]}_{[2]}} & \cdots & \overline{\alpha^{[P]}} \end{pmatrix} \tag{8}$$

Où

$$\overline{\alpha^{[p]}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_i^{[p]}, \tag{9}$$

et

$$\overline{\beta_{[q]}^{[p]}} = \frac{1}{N}\sum_{i=1}^{N} \beta_i{}_{[q]}^{[p]}. \tag{10}$$

4 Validation expérimentale

Dans cette section, nous présentons les résultats obtenus après avoir exécuté notre algorithme sur plusieurs jeux de données. Tout d'abord, nous décrivons les jeux de données qui ont été utilisés dans les expériences, puis nous présentons deux types de résultats : l'évaluation des performances et la visualisation du processus de collaboration.

4.1 Jeux de données

— Le jeu de données Breast Cancer Wisconsin (Diagnostic) (Wdbc) se compose de 569 images numérisées d'une masse mammaire. Il comporte 30 variables à valeurs réelles décrivant les noyaux cellulaires présents dans chaque image. Chaque observation est étiquetée comme bénigne ou maligne.

— Le jeu de données Spambase est composé de 57 attributs décrivant une collection de 4601 courriels spam et non-spam.

— Le jeu de données Battalia3 est un ensemble de données artificielles décrivant 2000 exoplanètes générées avec 27 attributs numériques.

— Le jeu de données MV2 comprend 2000 entrées, chacune décrite par 6 caractéristiques. Elles ont été générés aléatoirement à partir du mélange d'un bruit et de quatre composantes gaussiennes.

— Le jeu de données Isolet (isolated letters) comporte 617 variables décrivant 7797 enregistrements vocaux d'individus ayant prononcé le nom de chaque lettre de l'alphabet.

— Le jeu de données Madelon est un jeu de données artificiel contenant 4400 lignes composant 32 clusters placés sur les sommets d'un hypercube à cinq dimensions. Ensuite, 15 caractéristiques redondantes et 480 caractéristiques inutiles (sondes aléatoires) ont été ajoutées, pour un total de 500 attributs.

4.2 Protocole experimental

Nous avons divisé les jeux de données afin d'obtenir un cadre de clustering collaboratif horizontal - c'est-à-dire que chaque site de données a accès à différentes variables portant sur le même ensemble d'observations.

4.3 Résultats et analyse

4.3.1 Co-LUPI basé sur les modèles GTM

L'algorithme Co-LUPI a été appliqué à chacun des 6 jeux de données mentionnés dans la section 4.1 en utilisant des cartes topographiques génératives (GTM).

TAB. 1: Résultats expérimentaux, indice de Davies-Bouldin - Co-EM : (Sublime et al., 2017), Co-MV : (Ghassany et al., 2013), Co-GTM : (Ghassany et al., 2012), Co-SOM : (Grozavu et Bennani, 2010), Co-LUPI, RCo-LUPI

Name	Co-EM	Co-MV	Co-GTM	Co-SOM	Co-LUPI	RCo-LUPI
Wdbc	0.85	0.97	0.9	0.84	0.78	**0.69**
Spam Base	0.94	1.27	0.92	0.87	**0.42**	0.59
Battalia3	2.43	2.83	2.68	2.51	1.47	**1.37**
MV2	1.34	1.34	1.61	1.44	0.86	**0.85**
Isolet	–	–	–	–	1.33	**1.31**
Madelon	–	–	–	–	0.87	**0.82**

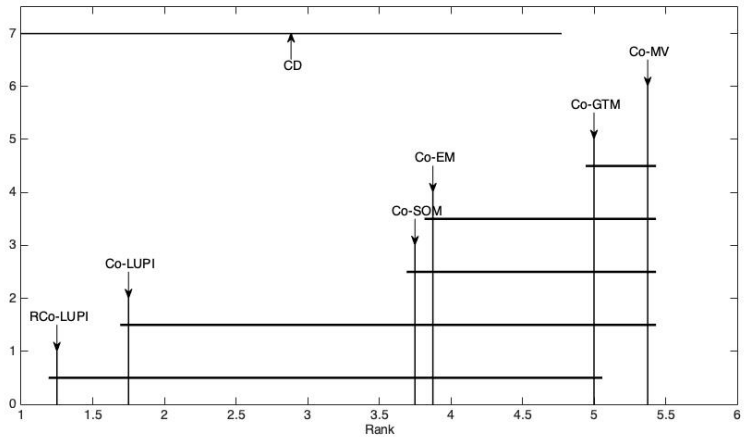
Le critère d'acceptation de la collaboration était l'amélioration de l'indice de Davies-Bouldin. Il s'agit d'un indice interne, qui ne nécessite donc pas de connaissances préalables sur la structure des données. Une deuxième version de l'algorithme, RCo-LUPI (Regulazed Collaborative Learning Using Privileged Information), a été implémentée. Elle comporte une nouvelle initialisation, aléatoire, de la matrice de responsabilité à chaque étape, ainsi que de la matrice de collaboration. Cette technique est souvent utilisée dans l'apprentissage non supervisé, elle est destinée à réduire la dépendance aux paramètres initiaux. Le tableau 1 montre que dans la plupart des cas, RCo-LUPI a donné des résultats légèrement meilleurs que Co-LUPI.

Afin de visualiser la dynamique de ce processus, nous pouvons examiner les matrices de confiance successives de l'étape de collaboration. La figure 2 représente de telles données. L'algorithme Co-LUPI a été appliqué à l'ensemble de données Wdbc, réparti sur 18 sites de données. On constate que tous les algorithmes n'ont pas bénéficié de la collaboration à chaque étape. En particulier, l'algorithme exécuté sur le site de données numéro 1 n'a pas amélioré ses résultats avant la cinquième itération de l'étape de collaboration. En outre, à l'itération numéro 7, seul le deuxième algorithme a amélioré ses résultats. Bien que cela puisse être interprété comme le signe d'une fin imminente du processus, trois autres algorithmes ont bénéficié de ces nouvelles découvertes lors de l'itération suivante. Le processus n'a donc pris fin qu'après quatre itérations supplémentaires.

4.4 Comparison avec d'autres approches collaboratives

Les algorithmes Co-LUPI et RCo-LUPI ont été comparés empiriquement à quatre implémentations récentes d'algorithmes de clustering collaboratif.

FIG. 1: Test de Friedman et Nemenyi pour comparer plusieurs approches sur plusieurs ensembles de données : Les approches sont classées de gauche (la meilleure) à droite (la plus mauvaise)



Le processus d'optimisation de l'algorithme Co-EM est basé sur l'EM variationnel. Il optimise un terme de collaboration qui est équivalent à l'entropie (Sublime et al., 2017). Le même principe est utilisé dans le Co-EM à la différence qu'il est basé sur des prototypes, alors que le Co-EM est basé sur des partitions (Ghassany et al., 2013). Dans les méthodes Co-SOM et Co-GTM, les fonctions de perte SOM et GTM ont été modifiées afin de pénaliser la différence entre les paramètres locaux et distants (Grozavu et Bennani, 2010; Ghassany et al., 2012).

Afin d'évaluer la performance de nos approches, nous utilisons le test de Friedman et le test de Nemenyi recommandés dans (Demšar, 2006). Tout d'abord, les algorithmes sont classés en fonction de leurs performances sur chaque jeu de données. Il y a alors autant de classements que de jeux de données. Ensuite, le test de Friedman est effectué pour tester l'hypothèse nulle selon laquelle toutes les approches sont équivalentes, ce qui suppose que leurs classements moyens soient égaux. Si l'hypothèse nulle est rejetée, le test de Nemenyi est alors effectué. Si les rangs moyens de deux approches diffèrent d'au moins la différence critique (CD), alors on peut conclure que leurs performances sont significativement différentes. Dans le test de Friedman, nous fixons le seuil de signification $\alpha = 0.05$. La figure 1 montre un diagramme critique représentant les rangs moyens des algorithmes. Les méthodes sont ordonnées de gauche (les meilleures) à droite (les moins bonnes) et une ligne horizontale relie les groupes d'algorithmes qui ne sont pas significativement différents (pour le niveau de signification $\alpha = 5\%$). Comme le montre la figure 1, Co-LUPI et RCo-LUPI semblent obtenir une certaine amélioration par rapport aux autres techniques proposées. Mais les résultats ne sont pas suffisants pour conclure à une amélioration statistiquement significative. Cela peut s'expliquer par le petit nombre de jeux de données et par le fait que le test de Nemenyi ne prend en compte les performances des algorithmes qu'à travers leur rang, sans tenir compte de la valeur réelle de l'indice de performance.

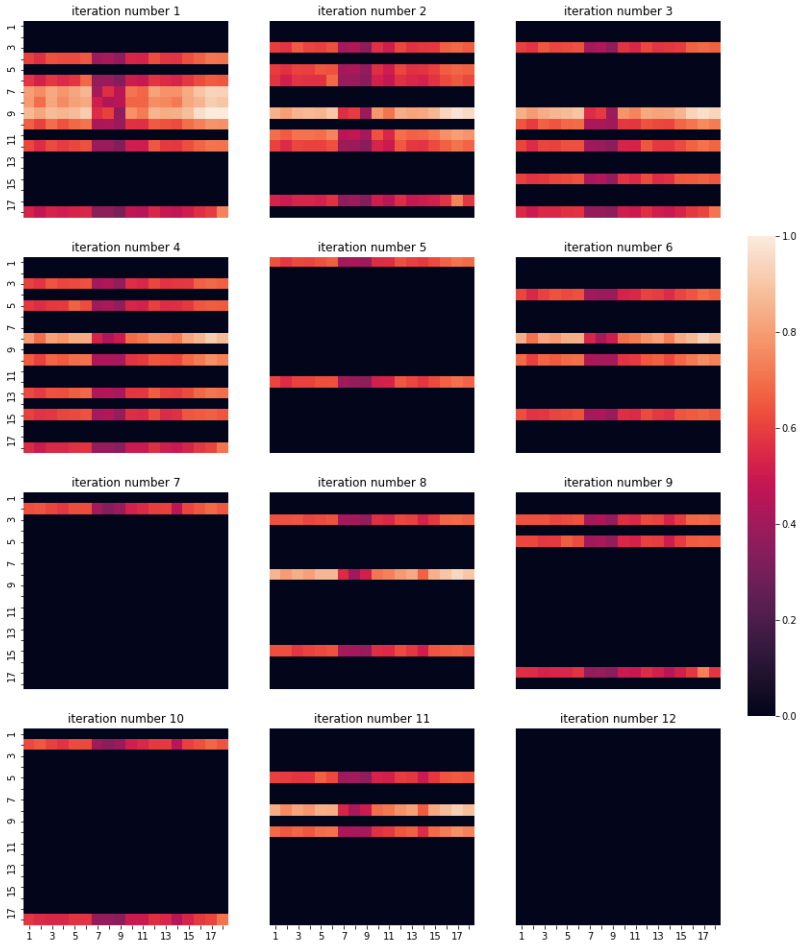


FIG. 2: Flux d'informations pendant la collaboration pour l'ensemble de données Wdbc. Chaque ligne représente un site de données et chaque colonne le poids total attribué à cette source.

5 Conclusion et perspectives

Nous avons présenté Co-LUPI et RCo-LUPI, des algorithmes basés sur l'utilisation d'informations privilégiées et de probabilités pour le clustering collaboratif. Cela permet aux algorithmes d'affiner la collaboration en fonction de leur (in)certitude - et de celle de leurs homologues distants - sur la classification de chaque observation, mesurée par l'entropie. Cette règle de mise à jour à l'avantage d'être simple et la collaboration se fait avec tous les sites distants en même temps. Cela évite le problème classique du choix du site avec lequel collaborer à chaque étape.

Nous avons testé notre approche sur plusieurs ensembles de données dans le cadre de la collaboration horizontale, mais elle est également applicable dans le cadre hybride. Les résultats ont montré une amélioration par rapport à l'état de l'art. Le cadre fournit également un moyen de visualiser le flux d'informations pendant le processus. Il présente des comportements intéressants, car les algorithmes ayant une performance initiale plus faible ont tendance à utiliser davantage les informations entrantes que les autres. De plus, même les algorithmes ayant les meilleurs résultats après l'étape locale ont pu s'améliorer au cours du processus. En effet, la flexibilité apportée par Co-LUPI dans la pondération des informations entrantes permet aux algorithmes de bénéficier d'homologues globalement moins efficaces, car ces derniers peuvent être localement plus efficaces.

Plusieurs améliorations peuvent être apportées à l'algorithme proposé. Un algorithme de transport optimal pourrait être utilisé pour relâcher l'hypothèse selon laquelle chaque algorithme recherche le même nombre de clusters. Nous étudierons également les performances de l'algorithme CoLUPI lorsqu'il est utilisé avec certains des algorithmes les plus représentatifs pour résoudre le problème, par exemple l'algorithme d'optimisation du papillon monarque (MBO) (Wang et al., 2015).

Références

Bezdek, J. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*.

Cabanes, G. et Y. Bennani (2007). A simultaneous two-level clustering algorithm for automatic model selection. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pp. 316–321.

Cornuéjols, A., C. Wemmert, P. Gançarski, et Y. Bennani (2018). Collaborative Clustering : Why, When, What and How. *Information Fusion 39*, 81–95.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7*(1), 1–30.

Ghassany, M., N. Grozavu, et Y. Bennani (2012). Collaborative generative topographic mapping. In T. Huang, Z. Zeng, C. Li, et C. S. Leung (Eds.), *Neural Information Processing*, Berlin, Heidelberg, pp. 591–598. Springer Berlin Heidelberg.

Ghassany, M., N. Grozavu, et Y. Bennani (2013). Collaborative multi-view clustering. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Grozavu, N. et Y. Bennani (2010). Topological collaborative clustering. *Australian Journal of Intelligent Information Processing Systems 12*(3).

Jain, A. K., M. N. Murty, et P. J. Flynn (1999). Data clustering : A review. *ACM Comput. Surv. 31*(3), 264–323.

Pedrycz, W. (2002). Collaborative fuzzy clustering. *Pattern Recognition Letters 23*(14), 1675–1686.

Sublime, J., M. Basarab, G. Cabanes, N. Grozavu, Y. Bennani, et A. Cornuéjols (2017). Entropy Based Probabilistic Collaborative Clustering. *Pattern Recognition 72*, 144–157.

Vapnik, V. et R. Izmailov (2015). Learning using privileged information : Similarity control and knowledge transfer. *J. Mach. Learn. Res. 16*(1), 2023–2049.

Vapnik, V. et A. Vashist (2009). 2009 special issue : A new learning paradigm : Learning using privileged information. *Neural Netw. 22*(5–6), 544–557.

Wang, G.-G., S. Deb, et Z. Cui (2015). Monarch butterfly optimization. *Neural Computing and Applications*.

Wemmert, C., P. Gancarski, et J. Korczak (2000). A collaborative approach to combine multiple learning methods. *International Journal on Artificial Intelligence Tools 9*, 59–78.

Summary

In the collaborative clustering framework, the hope is that by combining several clustering solutions, each one with its own bias and imperfections, one will get a better overall solution. The goal is that each local computation, quite possibly applied to distinct data sets, benefits from the work done by the other collaborators. This article is dedicated to collaborative clustering based on the Learning Using Privileged Information paradigm. Local algorithms weight incoming information at the level of each observation, depending on the confidence level of the classification of that observation. A comparison between our algorithm and state of the art implementations shows improvement of the collaboration process using the proposed approach.

Keywords: Unsupervised learning, Collaborative Clustering

Clustering spectral en utilisant des approximations d'ordre supérieur non homogènes de la distribution de Student

Nistor Grozavu*, Petru Alexandru Vlaicu**
Nicoleta Rogovschi *,*** Basarab Matei*

*LIPN, CNRS UMR 7030, Université Sorbonne Paris Nord
nom.prénom@lipn.univ-paris13.fr,
**Chemistry and Animal Nutrition Physiology Department, INCDBNA-INBA, Baloteşti, România
alexandru.vlaicu@outlook.coml
***LIPADE, Université de Paris, Paris, France
nicoleta.rogovschi@parisdescartes.fr

Résumé. Cet article présente une nouvelle méthode pour visualiser des ensembles de données en grande dimension. L'approche proposée est basée sur des approximations d'ordre supérieur de la distribution de Student. Celle-ci est utilisé pour définir une distance ou similarité entre les données. La stratégie d'approximation proposée permet d'éviter le calcul des exponentielles de matrices ce qui conduit à une complexité faible de l'algorithme proposé. D'autre part l'approximation proposée est inhomogène, dépendant du nombre de degrés de liberté de la distribution de Student correspondante, et elle a l'ordre de précision de 10^{-3}.. Cela permet ainsi d'adapter l'approximation localement aux variétés sur lesquelles se situent les données en très grande dimension. Nous illustrons la puissance de l'approche proposée avec plusieurs experiences sur des ensembles de données réels. Les résultats obtenus surpassent les méthodes SNE et t-SNE classiques.

Mots-clés : Visualisation de Données, Réduction de la dimensionalité, Clustering

1 Introduction

Au cours des dernières décennies, un grand nombre de méthodes de réduction de la dimensionalité (RD) ont été proposées, afin de visualiser de données en grande dimension, telles que Isomap Tenenbaum et al. (2000), Locally linear embedding Roweis et Saul (2000), Laplacian eigenmaps Belkin et Niyogi (2003), Stochastic Neighbor Embedding (SNE) Hinton et Roweis (2003), et les cartes de diffusion Lafon et Lee (2006). L'analyse en composantes principales (ACP), introduite en 1901, est l'une des méthodes phares de réduction de dimensionnalité. L'ACP est largement utilisé dans la communauté Machine Learning, cependant la point faible de cette méthode est constitué par l'utilisation d'un modèle linéaire intrinsequement Zhu et al. (2019). En raison de cette linéarité, l'ACP présente plusieurs inconvénients pour traiter des

données complexes et hétérogènes, cela motivant l'introduction des plusieurs méthodes adaptatives de RD. De par leur adaptabilité, les techniques non-linéaires de RD ont la capacité de traiter des données non linéaires complexes. Relativement récent, une nouvelle méthode de réduction de la dimensionalité, appelée t-Distributed SNE (t-SNE) Van der Maaten et Hinton (2008) a été introduite.

La t-SNE permet de représenter un ensemble de points d'un espace à grande dimension dans un espace de faible dimension en particullier 2 ou 3. Dans l'espace de faible dimension les données sont ainsi représentées par des nuages de points.

L'algorithme t-SNE utilise un critère issu de la théorie de l'information qui peut être ennoncé comme suit : si dans l'espace de grande dimension deux points sont géométriquement proches/éloignes, leurs représentants dans l'espace de faible dimension seront proche/ éloignés.

Pour ce faire, l'algorithme t-SNE propose une interprétation probabiliste des proximités. Plus précisemment, l'algorithme t-SNE définie une distribution de probabilité dans l'espace de grande dimension et une dexuième distribution de probabilité dans l'espace de visualisation de faible dimension.

L'algorithme t-SNE consiste à faire correposndre les deux densités de probabilité, en minimisant la divergence de Kullback-Leibler entre les deux distributions par rapport à l'emplacement géometrique des points dans l'espace de faible dimension Platzer (2013), Zhou et al. (2018),Zhou et Jin (2020), Cieslak et al. (2020). De cette manière si des points sont proches les uns des autres alors la probabilité d'être sélectionnés est très forte et inversement si des points sont éloignés alors la probabilité d'être sélectionnés est faible.

Lorsque l'ensemble de données contient une variété en grande dimension, les résultats de visualisation obtenus par la méthode t-SNE sont meilleures que les méthodes conventionnelles DR. Un exemple d'un tel ensemble de données est l'ensemble de données de chiffres écrits à la main - MNIST. Alors que les méthodes conventionnelles ne parviennent pas à révéler les clusters naturels de données MNIST, la t-SNE révèle ces clusters et cela sans utiliser les informations de classe supervisées.

L'une des malédictions de la dimensionnalité est le problème posé par l'agglomérations des points dans l'espace de faible dimension. Afin de rélier l'espace de modélisation des données de la position géométrique, dans la méthode t-SNE les points distants dans l'espace de grande dimension sont éloignés de force les uns des autres. En minimisant les divergences de Kullback-Leibler entre les deux distributions de deux espaces permet à t-SNE d'atténuer ce problème d'agglomération des points et c'est la principale raison pour laquelle t-SNE est meilleure que les autres méthodes de RD Van der Maaten et Hinton (2008).

Cependant, dans les données contenant une variété de grande dimension, le niveau de l'écart entre les points en grande dimension change en fonction de la dimension de la variété. Pour contourner ce problème, dans l'articleVan der Maaten (2008), l'auteur fait l'hypothèse que le niveau de l'écart est le même pour tous les points d'échantillonnage, cette hypothèse n'est pas raisonnable car la dimension intrinsèque de la variété de données est inhomogène.

Dans Kitazono et al. (2016) les auteurs ont proposé d'optimiser la dépendance de la dimension de la variété en considérant que la distribution de données n'est plus une distribution de Student homogène, mais inhomogène. Cela permet aux auteurs d'optimiser la dépendance de la dimension par la valeur estimée du paramètre de cette distribution de Student inhomo-

gène, c'est à dire le nombre de degrés de liberté. Il est mentionné que dans cette optimisation n'affecte pas significativement la visualisation.

Dans cet article, nous proposons une nouvelle méthode RD, appelée t-SNE inhomogène approchée, dans laquelle la dépendance de la dimension est toujours optimisée d'une manière inhomogène pour chaque point et toujours en estimant le nombre de degrés de liberté point par point non de la distribution de Student inhomogène comme dans Kitazono et al. (2016), mais plutôt une approximation d'ordre supérieur de celle-ci. En utilisant cette approximation inhomogène d'ordre supérieur, le calcul des exponentielles de matrices est évité, en résultant un algorithme de faible complexité . Le reste de l'article est organisé comme suit. Dans les sections 2 et 3, nous décrivons SNE et t-SNE. Nous présentons notre méthode proposée dans la section 4. Les résultats expérimentaux sont présentés dans la section 5. La section 6 conclut l'article.

2 Les méthodes SNE et t-SNE

Soit $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ un ensemble de données de N échantillons dans un espace de grande dimension, espace doté d'une certaine distance $d(\cdot, \cdot)$. En général, dans les méthodes RD, afin de visualiser des données, nous voulons projeter l'échantillon de grande dimension X sur un échantillon situé dans un esapce de faible (deux ou trois) dimensions $Y = \{y_1, \ldots, y_N\}$ tout en préservant certains aspects de la relation topologique entre les points de X.

2.1 SNE

Le point de départ de SNE est de convertir la distance disponible en grande dimension $d(\mathbf{x}_i, \mathbf{x}_j)$ (par exemple la distance euclidienne $\|\mathbf{x}_i - \mathbf{x}_j\|$) dans une certaine probabilité $p_{j|i}$ qui représente la similarité de $\mathbf{x}_j$ à $\mathbf{x}_i$. Pour calculer cette probabilité/similarité, la méthode SNE utilise un noyau gaussien normalisé comme suit :

$$p_{j|i} = \frac{\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}\right)}{\sum\limits_{k=1, k\neq i}^{N} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma_i^2}\right)}, \text{ and } p_{i|i} = 0, \tag{1}$$

où σ_i^2 est la variance du noyau gaussien. Notez que la variance σ_i^2 dépend de la position de l'échantillon $\mathbf{x}_i$.

Afin de déterminer la valeur presque optimale de σ_i la méthode SNE utilise le paramètre de perplexité $2^{-\sum_j p_{j|i} \log_2 p_{j|i}}$ qui est exactement égal à une valeur définie par l'utilisateur u. La définition de la perplexité est issue de la théorie de l'information.

Ce calcul pourrait être effectué même par une recherche binaire Hinton et Roweis (2003) ou une méthode de recherche de racine robuste Vladymyrov et Carreira-Perpinan (2013).

De la même manière, dans l'espace de dimension faible, la méthode SNE calcule la probabilité $q_{j|i}$ à partir de la distance euclidienne $\|y_i - y_j\|$ dans l'espace de dimension faible, en

utilisant noyau gaussien normalisé :

$$q_{j|i} = \frac{\exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2}\right)}{\sum_{k=1,k\neq i}^{N} \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_k\|^2}{2}\right)}, \text{ and } q_{i|i} = 0. \tag{2}$$

Nous pouvons remarquer que dans l'espace de faible dimension nous n'avons pas d'adaptativité du noyau par rapport à la variance. Pour déterminer les coordonnées de $\mathbf{y}_i$, nous utilisons respectivement les distributions P_i et Q_i, correspondant respectivement à la distribution de probabilité dans l'espace de dimension supérieure et inférieure. Donc les coordonnées de by_i sont obtenues par la minimisation de la divergence KL entre les répartition P_i et Q_i :

$$C_{SNE}(Y) = \sum_{i=1}^{N} KL(P_i \| Q_i) = \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \tag{3}$$

Pour la minimisation de la fonctionnelle dans l'équation (3) il est utilisé le gradient suivant :

$$\frac{\partial C_{SNE}(Y)}{\partial \mathbf{y}_i} = 2 \sum_{j\neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(\mathbf{y}_i - \mathbf{y}_j). \tag{4}$$

Remarquons que dans l'espace de faible (resp grande) dimension $q_{j|i} \neq q_{i|j}$ (resp. $p_{j|i} \neq p_{i|j}$), donc ces probabilités ne sont pas symétriques. Un autre point à remarquer est le problème de de l'agglomération, c'est-à-dire même pour de petites valeurs des similarités $p_{j|i}$ et $p_{i|j}$ nous voulons que dans l'espace de faible dimension, les points correspondants $\mathbf{y}_i$ et $\mathbf{y}_j$ dans la visualisation soient bien séparés ce qui est difficile à attendre. Cet inconvénient motive la définition d'une stratégie différente pour la visualisation est ce qu'on appelle t-SNE.

2.2 t-SNE

Nous consacrons ce paragraphe à la présentation de t-SNE. Bien que proches, les méthodes SNE et t-SNE il y ait deux différences principales entre ces deux méthodes. Premièrement, le t-SNE utilise une probabilité symétrisée $p_{ij} = \ fracp_{j|i} + p_{i|j}2N$, par cette symétrisation le gradient résultant du la fonction de coût est plus simple et plus facile à optimiser.

Deuxièmement, dans t-SNE, pour définir la similarité entre les points de données $\mathbf{y}_i$ et $\mathbf{y}_j$ dans la basse dimension, nous utilisons toujours la probabilité q_{ij} qui est définie dans t-SNE non pas par le noyau gaussien, mais par noyau de la distribution t de Student avec un degré de liberté fixé, ce qui donne la formule suivante :

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k,l(k\neq l)} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}, \text{ and } q_{i|i} = 0. \tag{5}$$

Avec cette définition, la fonction de coût dans le t-SNE s'écrit comme suit :

$$C_{tSNE}(Y) = KL(P\|Q) = \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \tag{6}$$

Alors que son gradient est donné par :

$$\frac{\partial C_{tSNE}(Y)}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{j|i} - q_{j|i})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j). \quad (7)$$

Rappelons que la distribution de Student converge en loi vers la la distribution normale, mais que le comportement des deux distributions dans les zones éloignées de leurs valeurs centrales est très différent. Cela est du au fait que le coefficient d'aplatissement *kurtosis* de la distribution de Student est positif et donc plus fort que le coefficient d'aplatissement de la distribution gaussienne. Cela implique que la concentration pour les valeurs extrêmes éloignées de la moyenne, est plus importante en utilisant une distribution de Student. En d'autre mots la queue ou la traîne de la loi de Student est plus lourde que celle de la distribution gaussienne.

Par conséquent, en utilisant la distribution de Student, pour modèliser les petites valeurs de p_{ij}, la distance entre $\mathbf{y}_i$ et $\mathbf{y}_j$ sera forcément plus grande.

3 Notre méthode

En utilisant la distribution de Student, la t-SNE permet de compenser le problème d'agglomération et d'obtenir cela pour de petites valeurs p_{ij} les points de représernation de données $\mathbf{y}_i$ et $\mathbf{y}_j$ bien séparés. Mais en utilisant la distribution de Student avec un nombre fixe de degrés de liberté, une limitation dans la t-SNE d'origine est introduite, c'est-à-dire que la même distance est optimisée en fonction de la lourdeur de la queue de la distribution de Student pour tous les échantillons et pour tout ensemble de données. Cela n'est pas raisonable car les données en grande dimensions contiennent des variétés de dimmensions différentes. Comme nous l'avons dans l'introduction une solution a été proposé dans Kitazono et al. (2016), les auteurs montrent qu'en adaptant la lourdeur de la queue en utilisant la distribution de Student à nombre variable de degrés de liberté pour calculer q_{ij}, ils peuvent améliorer la visualisation par le t-SNE inhomogène. Cependant dans Kitazono et al. (2016) la complexité de la méthode est grande du fait qu'un processus d'optimisation est effectué pour chaque point de données et chaque ensemble de données et que chaque optimisation impliquant le calcul de plusieurs exponentielles matricielles. Cet inconvénient a motivé le présent article, dans lequel nous présentons dans la section suivante notre méthode appelée t-SNE inhomogène approché.

3.1 t-SNE inhomogène approché

Nous proposons une méthode qui adopte une formulation plus générale de la distribution de Student que dans le t-SNE original. Dans la formule générale, de la distribution de Student il y a le paramètre nu désignant le nombre de degrés de liberté, qui contrôlent la lourdeur de la queue.

En optimisant ce paramètre pour chaque point avec une faible compléxité, la méthode proposée permettra d'adapter le calcul de la distance entre $\mathbf{y}_i$ et $\mathbf{y}_j$ pour des petites valeurs de p_{ij}, en fonction de la dimension des données en grande dimension en fonction des variétés sur lesquelles celles-ci se situent en s'adaptant ainsi à l'inhomogénéité intrinsèque dans les jeux de données en grande dimention.

3.2 Degrés de liberté

Généralement, la fonction de densité de probabilité de la distribution de Student est donnée par :

$$f(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{8}$$

où $\Gamma(\cdot)$ est la fonction gamma et ν est le paramètre appelé le nombre de degrés de liberté. Lorque ν diminue, la queue de la distribution devient plus lourde. La distribution $f(x) \propto (1+x^2)^{-1}$ et est utilisée dans le t-SNE d'origine.

3.3 Approximation non homogène de la distribution de Student

Notons $F(x;\nu)$ la fonction de distribution cumulative de Student avec ν degrés de liberté. Due à la convergence en loi de la distribution de Student vers la distribution de Gausse, il est bien connu que la distribution cumulative de Student a une approximation normale ordinaire pour $\nu \geq 30$, c'est-à-dire, $F(x;\nu)$ c'est environ $\Phi(x)$. Li et De Moor Li et de Moor (1999) a proposé une simple approximation de $F(x;\nu)$ comme une alternative aux diverses approximations listées dans Johnson et al. (1999). Li et De Moor Li et de Moor (1999) ont proposé un approximation normale ajustée des distributions cumulatives des familles de Student.

De la formule obtenue dans Li et De Moor Li et de Moor (1999) nous avons déduit les approximations de la distribution de Student comme suit :

$$f(x;\nu) \approx \lambda\phi(\lambda x), \text{ avec } \lambda = \lambda(x;\nu) = \frac{4\nu + x^2 - 1}{4\nu + 2x^2}. \tag{9}$$

où ϕ est la distribution gaussienne. Puisque nous voulons éviter le calcul de l'exponentielle dans ϕ nous utilisons plutôt l'approximation polynomiale de Taylor de ϕ d'ordre s notée $\mathbb{P}_s$ et on utilise

$$f(x;\nu) \approx \lambda\mathbb{P}_s(\lambda x). \tag{10}$$

Dans les cas $nu = 1$ et $nu = 2$, nous avons le calcul exact suivant :

$$f(x;1) = \frac{1}{\pi}\frac{1}{x^2+1} \tag{11}$$

et

$$f(x;2) = \frac{1}{2}\frac{1}{\sqrt{x^2+2}} + \frac{x}{2}\frac{x}{\sqrt{(x^2+2)^3}} = \frac{x^2+1}{\sqrt{(x^2+2)^3}} \tag{12}$$

Dans les sous-sections suivantes, nous définissons d'abord la fonction de coût de la méthode et expliquer comment l'optimiser.

3.4 Fonction de coût et son gradient

Dans la méthode proposée, nous définissons la probabilité dans l'espace de faible dimension comme :

$$q_{j|i} = \frac{\left(1+\frac{\|\mathbf{y}_i-\mathbf{y}_j\|^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}}{\sum\limits_{k=1,k\neq i}^{N}\left(1+\frac{\|\mathbf{y}_i-\mathbf{y}_k\|^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}}, \text{ and } q_{i|i} = 0. \tag{13}$$

La fonction de coût est donnée par la même forme que celle de SNE (Eq. 3) : $C(Y) = \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$. Dans la définition ci-dessus de $q_{j|i}$, le paramètre nu_i est introduit point par point. En optimisant ce paramètre, la méthode proposée adaptera l'épaisseur de la queue de la distribution de Student pour chaque jeu de données et pour chaque point.

Après un peu d'algèbre, on obtient dans l'équation (15) le gradient de la fonction de coût dans l'équation (14) :

$$C_{tSNE}(Y) = 2 \sum_{j \neq i} \left[\frac{\theta_i}{\gamma_{ij}} (p_{j|i} - q_{j|i}) + \frac{\theta_j}{\gamma_{ij}} (p_{i|j} - q_{i|j}) \right] (\mathbf{y}_i - \mathbf{y}_j). \tag{14}$$

$$\frac{\partial C_{tSNE}(Y)}{\partial \mathbf{y}_i} = \sum_{j \neq i} \left[\frac{1}{2} \log \gamma_{ij} - \frac{\theta_i}{2\nu_i \gamma_{ij}} d_{ij}^2 \right] (p_{i|j} - q_{i|j}) \tag{15}$$

où $\theta_i = \frac{\nu_i + 1}{\nu_i}$, $\gamma_{ij} = \left(1 + \frac{d_{ij}^2}{\nu_j}\right)$ et $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$. Quand on fixe $\nu_i = 1$ pour tout i, la méthode proposée coïncide avec SNE. Lorsque $\nu_i = 1$ pour tout i, la méthode proposée devient un version asymétrique du t-SNE.

3.5 Optimisation

Dans cette section, nous présentons la stratégie d'optimisation de la fonction de coût. à cette fin, nous utilisons une variante des méthodes de quasi-Newton appelées Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) à mémoire limitée, comme dans l'articleVladymyrov et Carreira-Perpinan (2014).

Nous utilisons une procédure en trois étapes. Dans la première étape, nous allons faire une " exagération précoce" des similarité entre les points dans l'espace à très garnde dimension,cette technique a été introduite déjà dans la t-SNE d'origine Van der Maaten et Hinton (2008). Plus précisemmeent, nous multiplions toute les probabilités $p_{j|i}$ par une constante (par exemple 4). Toujours dans la première étape, nous allons fixé tous les paramètres ν_i à une valeur relativement plus élevée (par exemple 10). Dans cette étape, seuls $\mathbf{x}_i, (i = 1, \ldots, N)$ sont optimisés. Dans la deuxième étape, nous considérons que les valeurs originales des probabilités $p_{j|i}$ et $\mathbf{y}_i$ sont optimisées. Enfin, dans la troisième étape, $\mathbf{y}_i$ et ν_i sont optimisés. Afin de respecter $\nu > 0$, nous écrivons ν comme $\epsilon + \xi^2$, où $\epsilon = 10^{-3}$ et l'optimisation a été faite par rapport à ξ.

4 Protocole expérimental

Afin d'évaluer notre méthode, nous avons réalisé plusieurs expériences sur quatre jeux de données connus du référentiel UCI :2007 de bases de données d'apprentissage automatique. Asuncion et Newman (2007) et sur le jeu de données Growth weight (GWD) présenté dans Vlaicu et al. (2020). Ces jeux de données sont de taille et de complexité différentes et leurs caractéristiques sont résumées dans le tableau ci-dessous.

Données	nb. obs.	nb. variables	nb. classes
Waveform	5000	40	3
WDBC	569	32	2
SpamBase	4601	57	2
MNIST	1000	784	10
GWD	126	5	3

4.1 Description des données

— Waveform dataset : le jeu de données original comprenait 40 caractéristiques, dont 19 représentent du bruit avec une moyenne de 0 et une variance de 1. Nous disposons de 5000 observations réparties en trois classes, chaque classe étant générée à partir d'une combinaison de 2 des 3 ondes "de base".

— Wisconsin Diagnostic Breast Cancer (WDBC) : Le jeu de données contient 569 observations avec 32 caractéristiques (ID, diagnostic, 30 autres caractéristiques réelles). Chaque observation est étiquetée comme bénigne (357) ou maligne (212). Les caractéristiques sont calculées à partir d'une image numérisée d'une aspiration à l'aiguille fine (FNA) d'une masse mammaire. Elles décrivent les caractéristiques des noyaux cellulaires présents dans l'image.

— SpamBase : Ce jeu de données est composé de 4601 observations décrites par 57 caractéristiques. Chaque caractéristique décrit un e-mail et sa catégorie : spam ou non-spam. La plupart des attributs indiquent si un mot ou un caractère particulier apparaît fréquemment dans le e-mail. Les attributs 55-57 mesurent la longueur de séquences de lettres majuscules consécutives.

— MNIST : Le jeu de données MNIST est constitué d'images en niveaux de gris de chiffres manuscrits. La taille de chaque image est de $28 \times 28 = 784$ pixels. Chaque image correspond à l'une des dix classes (0-9). Nous avons sélectionné 300 images dans chacune des classes.

— Données sur le poids de croissance (GWD) : Cet ensemble de données réelles introduit dans Vlaicu et al. (2020) décrit l'effet des suppléments alimentaires d'orange et d'écorce de pamplemousse sur les performances de croissance, l'état de santé, la qualité de la viande et la microflore intestinale des poulets de chair. L'ensemble de données matricielles contient 126 poulets individuels regroupés sur la ligne suivant trois régimes différents, le régime de contrôle et deux régimes expérimentaux obtenus en supplémentant avec de l'orange et de l'écorce de pamplemousse le régime de contrôle. Dans les variables, nous avons pour chaque poulet la valeur du poids mesurée à 7 jours d'intervalle entre 14 jours et 42 jours de vie des poulets. Cet ensemble de données est ensuite décrit par une matrice de dimension 126×5.

4.2 Clustering spectral basé sur une approximation d'ordre supérieur de la distribution des données.

La méthode de clustering spectral nécessite de construire une matrice d'adjacence et de calculer la décomposition en valeurs propres de la matrice laplacienne correspondante. Ces deux étapes sont coûteuses en termes de calcul. Il n'est donc pas facile d'appliquer le clustering spectral à des ensembles de données à grande échelle. Afin de contourner ce problème, une

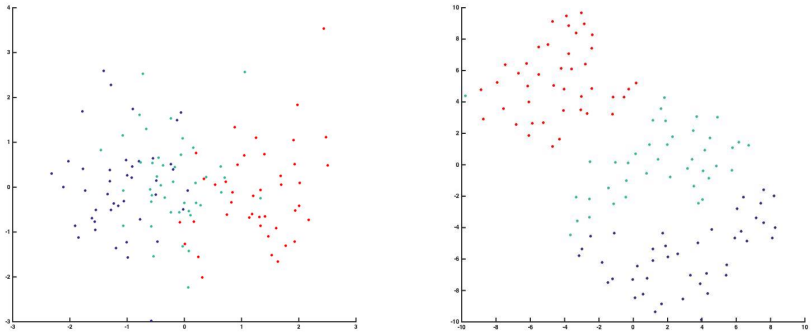


FIG. 1 – *Visualisation de données GWD avec l'ACP et t-SNE*

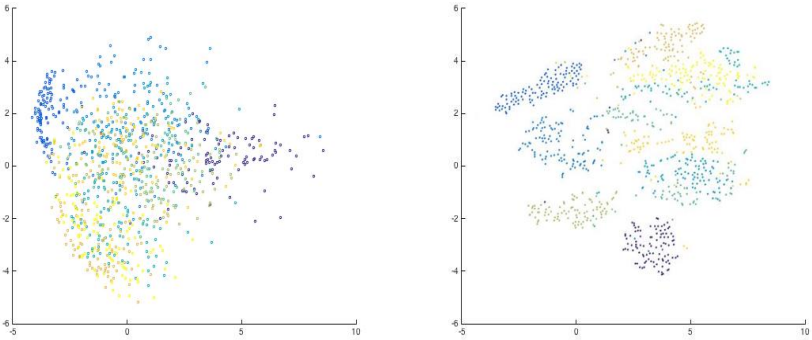


FIG. 2 – *Visualisation de données MNIST avec l'ACP et t-SNE*

première solution a été proposée dans Grozavu et al. (2016). Dans Grozavu et al. (2016), les auteurs proposent d'utiliser la matrice $A \in \mathbb{R}^{K \times K}$ defined by $A_{kk'} = \exp(-\|w_k - w_{k'}\|^2/2\sigma^2)$ if $k \neq k'$, avec $A_{kk} = 0$ pour tous $k, k' \in \{1, \ldots, K\}$, où w_k sont les centroïdes des clusters identifiés. La matrice A étant la représentation spectrale de l'espace des centroïdes. Nous proposons ici de construire la matrice d'affinité A en utilisant une approximation d'ordre supérieur de la fonction exponentielle, comme suit :

$$A_{kk'} = \frac{p_{k|k'} + p_{k'|k}}{2}, k \neq k'. \tag{16}$$

Nous notons que pour l'approximation du second ordre, nous avons $A_{kk'} > (2P)^{-2}$ for all $k, k' \in \{1, \ldots, K\}$ et $k \neq k'$. Comme nous l'avons déjà mentionné, l'approximation de l'exponentielle par une approximation du premier ordre ou d'un ordre supérieur permet d'éviter tous les points dont les probabilités sont proches de zéro, ce qui permet de réduire le temps de calcul par opération. Afin de calculer le Laplacien de la matrice A, nous définissons d'abord la matrice diagonale D comme suit :

$$D_{kk} = \sum_{m=1}^{M} A_{km}, \text{ and } D_{kk'} = 0, \text{ for } k' \neq k. \tag{17}$$

pour tous $k, k' \in \{1, \ldots, K\}$.

Réaliser un clustering sur des ensembles de données à haute dimension peut être difficile. Afin de prendre en compte la structure topologique de nos données, nous calculons la matrice d'affinité H dans l'étape 1 de notre algorithme. Dans l'étape 3, une nouvelle matrice d'affinité est calculée comme le produit entre la matrice H et la matrice symétrique de probabilité de distance définie dans l'équation (16). La nouvelle matrice d'affinité contient donc les deux sources d'information : la source topologique et la source spectrale.

Algorithme 1 : Clustering spectral basé sur une approximation d'ordre supérieur de la distribution des données.

Input : N données $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i \in \mathbb{R}^M$; Nombre de clusters K ;
Output : K clusters ;

1. Construire la matrice d'affinité par l'équation (16)
2. Calculer D défini dans l'équation (17).
3. Trouver les k plus grands vecteurs propres de $D^{-1/2}AD^{-1/2}$. (choisis pour être orthogonaux entre eux dans le cas de valeurs propres répétées), et définissons $U \in \mathbb{R}^{N \times K}$ en empilant ces vecteurs propres en colonnes..
4. Calculez Y la version ortonormalisée de U.
5. Regrouper chaque centroïde de Y en K clusters via l'algorithme des K-moyennes.
6. Affectez les centroïdes à la cluster k si et seulement si la ligne k de la matrice Y a été affectée à la cluster k.

4.3 Les indices de qualité

Lorsque les ensembles de données représentent les mêmes objets dans des espaces différents, les indices sont généralement basés sur la concordance entre les deux partitions, c'est-à-dire que chaque paire d'objets doit être soit dans le même cluster dans les deux partitions, soit dans des clusters différents dans les deux partitions. Pour chaque ensemble de données et chaque expérience, nous avons calculé les indices de validation externes du clustering, car les informations réelles sur les classes sont disponibles pour chacun d'entre eux, et nous avons également comparé le temps de calcul entre le clustering spectral classique et la méthode proposée.

L'évaluation de la performance d'un algorithme de clustering n'est pas aussi triviale que le decomptage du nombre d'erreurs comme pour un algorithme de classification supervisée. Pour évaluer la qualité du clustering, nous nous adoptons l'approche consistant à comparer les résultats à une "vérité de terrain". Nous utilisons l'indice de pureté, l'indice de Rand et l'indice de Jaccard pour mesurer les résultats du clustering. Il s'agit d'une approche commune dans le domaine du clustering. En général, le résultat du clustering est évalué sur la base de connaissances externes sur la façon dont les clusters devraient être structurés. Cette procédure est définie par Jain et Dubes (1988) comme "validation de clustering par une classification extrinsèque", et a été suivie dans de nombreuses autres études (Jain et al. (1999), Khan et Kant (2007), Andreopoulos et al. (2006)). Nous pensons que cette approche est raisonnable si nous ne voulons pas juger les résultats du clustering par un indice de validité, qui n'est rien d'autre qu'un biais vers une propriété préférée du cluster (par exemple, compact, bien séparé ou connecté).

Nous rappelons ci-bas les définitions de indices utilisés. Pour ce faire, nous considérons un ensemble de n objets S. U et V sont deux partitions différentes de S. Supposons que U soit notre critère externe (par exemple, l'ensemble des étiquettes) et que V soit le résultat du clustering, nous définissons ce qui suit :

- a, le nombre de paires d'objets qui sont placés dans la même classe dans U et dans le même cluster dans V ;
- b, le nombre de paires d'objets placés dans la même classe dans U mais pas dans le même cluster dans V ;
- c, le nombre de paires d'objets de la même classe dans V mais différents dans U ;
- d, le nombre de paires d'objets dans des classes différentes et des clusters différents dans les deux partitions.

4.3.1 Indice de pureté

L'indice de pureté produit un résultat dans l'intervalle $[0, 1]$, où une valeur de 1 indique que U et V sont identiques.

La pureté d'un cluster est le pourcentage de données appartenant à la classe majoritaire. En supposant que l'ensemble des clusters soit $V = v_1, v_2, ..., v_{|V|}$ et l'ensemble des étiquettes est $U = u_1, u_2, ..., u_{|L|}$, la formule qui exprime la pureté est la suivante :

$$\text{pureté}(U, V) = \frac{1}{N} \sum_i \max_j |v_i \cap u_j|.$$

4.3.2 Indice de Rand

L'indice de Rand ou la mesure de Rand introduite dans Rand (1971) est une mesure de la similarité entre deux clusters de données. Les deux, $a + b$ peuvent être considérés comme des accords entre U et V, et $c + d$ comme le nombre de désaccords entre ces deux partitions. Par conséquent, l'indice de Rand Rand (1971) est calculé par l'expression suivante :

$$R(U, V) = \frac{a + b}{a + b + c + d} \; ; \tag{18}$$

L'indice de Rand a une valeur comprise entre 0 et 1,, 0 indiquant que les deux clusters de données ne concordent sur aucune paire de points et 1 indiquant que les clusters de données sont exactement les mêmes. Toutefois, cet indice ne tient pas compte du fait que la concordance entre les partitions pourrait être le fruit du hasard. Cela pourrait grandement biaiser les résultats pour des valeurs de concordance plus élevées Hubert et Arabie (1985).

4.3.3 Indice de Jaccard

Dans l'indice de Jaccard, qui a été couramment appliqué pour évaluer la similarité entre différentes partitions d'un même ensemble de données, le niveau de concordance entre un ensemble d'étiquettes de classe U et un résultat de classification V est déterminé par le nombre de paires de points assignés au même cluster dans les deux partitions. L'indice de Jaccard est défini par la formule suivante :

$$J(U, V) = \frac{a}{a + b + c} \tag{19}$$

Il s'agit simplement du nombre d'éléments uniques communs aux deux ensembles divisé par le nombre total d'éléments uniques des deux ensembles. L'indice de Jaccard prend une valeur comprise entre 0 et 1. Un indice de 1 signifie que les deux ensembles de données sont identiques, et un indice de 0 indique que les ensembles de données n'ont aucun élément commun.

Le tableau 1 résume les résultats obtenus pour les quatre jeux de données en appliquant le K-means classique, le clustering spectral et la méthode proposée pour le clustering spectral en utilisant l'approximation du premier et du second ordre de l'exponentielle. Nous pouvons noter que notre méthode est plus performante que la méthode classique des k-moyennes et le clustering spectral, mais nous devons noter ici que le but est également de préserver la structure topologique des données pour la visualisation. Nous notons également que l'ordre de l'approximation joue un rôle clé, car en utilisant une approximation d'ordre supérieur, les résultats sont améliorés. Grâce à la méthode t-SNE, les 3 classes de l'ensemble de données GWD sont bien séparées comme on peut le voir sur la figure 1, par rapport à l'utilisation de l'ACP (Analyse en Composantes Principales). Nous remarquons que la structure topologique des données est préservée grâce à l'utilisation de t-SNE, comme le montre la figure 2 pour le jeu de données MNIST par rapport à l'utilisation de la visualisation ACP.

5 Conclusions

Dans cette étude, nous proposons un nouveau modèle d'apprentissage non supervisé spectral qui permet de regrouper un grand ensemble de données en préservant la structure locale

TAB. 1 – *Résultats expérimentaux pour K-means, le clustering spectral et l'approche proposée.*

Données		Pureté	Rand	Jaccard
Waveform	K-means	51.4600	0.6216	0.5854
	spectral	54.1780	0.6721	0.6107
	1er ordre	54.8230	0.7036	0.6280
	2ième ordre	55.1320	0.7162	0.6430
	5ième ordre	56.3210	0.7432	0.6920
SpamBase	K-means	63.5949	0.5369	0.5086
	spectral	69.6320	0.6152	0.5746
	1er ordre	70.0740	0.6312	0.5791
	2ième ordre	70.7340	0.6552	0.5987
	5ième ordre	71.4016	0.6622	0.6081
Wdbc	K-means	83.4130	0.7104	0.5499
	spectral	85.1327	0.7689	0.6178
	1er ordre	85.1620	0.7710	0.6146
	2ième ordre	86.0230	0.8012	0.6456
	5ième ordre	87.3011	0.8102	0.6568
MNIST	K-means	50.2500	0.4998	0.3329
	spectral	85.0615	0.9307	0.8012
	1er ordre	85.8430	0.9514	0.8217
	2ième ordre	86.2131	0.9643	0.8342
	5ième ordre	87.0134	0.9731	0.8476
GWD	K-means	81.3140	0.7539	0.6566
	spectral	85.0516	0.9170	0.8123
	1er ordre	85.8430	0.9413	0.8137
	2ième ordre	86.0731	0.9534	0.8242
	5ième ordre	87.1316	0.9614	0.8352

des données. La méthode proposée utilise le modèle t-SNE pour réduire la dimensionnalité. Les résultats obtenus montrent que la méthode proposée améliore les résultats du clustering en termes d'indices externes. Dans le cadre de travaux futurs, nous envisageons d'adapter cette méthode aux ensembles de données multi-vues.

Références

Andreopoulos, B., A. An, et X. Wang (2006). Bi-level clustering of mixed categorical and numerical biomedical data. *International Journal of Data Mining and Bioinformatics 1*(1), 19 – 56.

Asuncion, A. et D. Newman (2007). UCI Machine Learning Repository.

Belkin, M. et P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation 15(6)*, 1373–1396.

Cieslak, M. C., A. M. Castelfranco, V. Roncalli, P. H. Lenz, et D. K. Hartline (2020). t-distributed stochastic neighbor embedding (t-sne) : A tool for eco-physiological transcriptomic analysis. *Marine Genomics 51*, 100–123.

Grozavu, N., N. Rogovschi, et L. Lazhar (2016). Spectral clustering trought topological learning for large datasets. In *Neural Information Processing - 23rd International Conference, ICONIP, Proceedings, Part III*, pp. 119–128.

Hinton, G. et S. Roweis (2003). Stochastic neighbor embedding. *Advances in neural information processing systems 15*, 833–840.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of classification*, 2(1) :193–218.

Jain, A. K. et R. Dubes (1988). *Algorithms for clustering data*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

Jain, A. K., M. N. Murty, et P. J. Flynn (1999). Data clustering : a review. *ACM Computing Surveys 31*(3), 264–323.

Johnson, N., S. Kotz, et N. Balakrishnan (1999). *Distributions in Statistics : Continuous Univariate Distributions*. Second edition, New York. Wiley.

Khan, S. et S. Kant (2007). Computation of initial modes for k-modes clustering algorithm using evidence accumulation. In *IJCAI*, pp. 2784–2789.

Kitazono, J., N. Grozavu, N. Rogovschi, T. Omori, et S. Ozawa (2016). t-distributed stochastic neighbor embedding with inhomogeneous degrees of freedom. In A. Hirose et al. (Eds.), *Neural Information Processing*, pp. 119–128. Springer International Publishing.

Lafon, S. et A. Lee (2006). Diffusion maps and coarse-graining : A unified frameworkfor dimensionality reduction, graph partitioning, and data set parameterization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(9)*, 1393–1403.

Li, B. et B. de Moor (1999). A corrected normal approximation for student's t distribution. *Computational Statistics & Data Analysis 29*, 213–216.

Platzer, A. (2013). Visualization of snps with t-sne. *PLOS ONE 8*(2), 1–6.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association.*, 846–850.

Roweis, S. et L. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290(5500)*, 2323–2326.

Tenenbaum, J., V. De Silva, et J. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290(5500)*, 2319–2323.

Van der Maaten, L. (2008). Learning a parametric embedding by preserving local structure. *International Conference on Artificial Intelligence and Statistics, JMLR, W&CP. 5*.

Van der Maaten, L. et G. Hinton (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning 9*, 2579–2605.

Vladymyrov, M. et M. Carreira-Perpinan (2013). Entropic affinities : Properties and efficient numerical computation. *Proceedings of The 30th International Conference on Machine Learning*, 477–485.

Vladymyrov, M. et M. Carreira-Perpinan (2014). Linear-time training of nonlinear lowdimensional embeddings. *Proceedings of AISTATS 2014, International Conference on Artificial*

Intelligence and Statistics, JMLR W&CP. vol. 33.

Vlaicu, P. A., A. Untea, T. Panaite, et R. Turcu (2020). Effect of dietary orange and grapefruit peel on growth performance, health status, meat quality and intestinal microflora of broiler chickens. *Italian Journal of Animal Science 19*(1), 1394–1405.

Zhou, B. et W. Jin (2020). *Visualization of Single Cell RNA-Seq Data Using t-SNE in R*, pp. 159–167. New York, NY : Springer US.

Zhou, H., F. Wang, et P. Tao (2018). t-distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations. *Journal of chemical theory and computation 14*(11), 5499–5510.

Zhu, W., Z. Webb, K. Mao, et J. Romagnoli (2019). A deep learning approach for process data visualization using t-distributed stochastic neighbor embedding. *Industrial & Engineering Chemistry Research 58(22)*, 9564–9575.

Summary

This paper introduces a new approximation strategy to visualize high dimensional datasets. The proposed approach is based on t-SNE (Stochastic Neighbor Embedding) dimensionality reduction method with a different inhomogenous approximation strategy of the t-Distribution. In order to avoid the exponential computation we propose in this paper an inhomogenous approximation of the t-Distribution having the precison order of 10^{-3}. By using this inhomogenous approximation we allow to optimize approximately the t-Distribution with respect to the number of degree of freedom and also to reduce the computational time. We illustrate the power of the proposed approach with several real datasets and the obtained results outperform classical SNE and t-SNE methods.

Keywords: Data-visualization, Dimensional reduction, Clustering

Clustering multi-vues basé sur le transport optimal régularisé

Fatima-Ezzahraa Ben-Bouazzaa[*,**,****] Younès Bennani[*,**]
Abdelfettah Touzani[***], Guénaël Cabanes[*,**]

[*]LIPN UMR 7030 CNRS, Université Sorbonne Paris Nord, France
name.surname@sorbonne-paris-nord.fr
[**]LaMSN, La Maison des Sciences Numériques, France
[***]LAMA-FSDM, Université Sidi Mohamed Ben Abdellah, Morocco
name.surname@usmba.ac.ma
[****] Université Mohammed VI des Sciences de la Santé

Résumé. Dans cet article nous présentons un nouveau cadre de clustering multi-vues formalisé dans la théorie du transport optimal, où l'idée principale est d'apprendre des modèles locaux à partir de chaque vue en se basant sur un clustering basé sur l'algorithme de Sinkhorn, et de chercher un clustering d'ensemble de toutes les vues pour obtenir un modèle de consensus. Pour ce faire, nous proposons deux approches basées sur la distance de Wasserstein régularisée et les barycentres Wasserstein : une approche de projection consensuelle (CPA : Consensus Projection Approach) qui consiste à apprendre un consensus sur l'espace original, et un consensus avec une nouvelle représentation (CNR : Consensus with New Representation) qui s'appuie sur une nouvelle distribution consensuelle apprise à partir de la distribution des différentes vues. Les deux approches sont comparées expérimentalement à d'autres approches de clustering multi-vues, afin de mettent en évidence l'efficacité de nos méthodes.

Mots-clés : Multi-Vues Clustering (MVC), Subspace clustering, Apprentissage multi-tasks, Transport Optimal (TO), Distance de Wasserstein/Earth Mover.

1 Introduction

Les méthodes de clustering non supervisées sont devenues récemment de plus en plus populaires en raison de leur capacité à regrouper en clusters des données non étiquetées en fonction de leurs similarités. Un nombre important de nouveaux algorithmes de clustering ont été développés ces dernières années, et la plupart des approches précédentes ont également été modifiées et améliorées. Cette abondance d'approches peut s'expliquer par la difficulté de proposer des méthodes génériques qui s'adaptent à tous les types de données disponibles (Cabanes et Bennani, 2007; Cabanes et al., 2012). En effet, chaque méthode possède un biais induit par l'objectif choisi pour créer les clusters. Ainsi, deux algorithmes différents peuvent offrir des résultats de clustering très différents à partir des mêmes données. De plus, un même

algorithme peut fournir des résultats différents en fonction de son initialisation ou des valeurs de ses paramètres.

Pour résoudre ce problème, certaines approches proposent d'utiliser plusieurs résultats de clustering différents afin de mieux refléter la complexité potentielle de la structure de données (Ghassany et al., 2013; Tao et al., 2017). Ces approches tirent parti des informations fournies par les différents résultats d'une manière sensiblement différente. Le clustering multi-vues (MVC) (Fu et al., 2020) est une approche populaire et efficace, dont le but est de former un clustering cohérent des données en combinant des algorithmes de clustering appliqués sur différentes vues des données (différentes représentations), au lieu d'une seule représentation (vue) des données. Chaque vue est généralement une description des données dans un espace de caractéristiques spécifique, chaque vue ayant son propre ensemble de caractéristiques. La caractéristique importante de cette méthode réside dans la diversité des caractéristiques utilisées dans les différentes vues. Cette diversité garantit non seulement un clustering cohérent, mais aussi une meilleure interprétation des clusters, grâce à la précision des représentants des clusters dans toutes les vues.

Dans cet article, nous proposons une nouvelle approche de MVC basée sur la théorie du Transport Optimal (TO). Cette nouvelle approche vise à apprendre une nouvelle structure de données à partir de la distribution des données sur chaque vue, afin d'augmenter la qualité et la richesse du résultat du clustering, grâce à la théorie TO qui nous permet non seulement de comparer les distributions, mais aussi de permettre le "transport" d'informations entre les vues pour produire un meilleur consensus. Nous proposons deux algorithmes à cette fin : une approche de projection du consensus (CPA) et un consensus avec nouvelle représentation (CNR), tous deux basés sur la distance de Wasserstein régularisée par l'entropie et les barycentres de Wasserstein entre les distributions de données dans chaque vue.

Le reste de l'article est organisé comme suit. Dans la section 2, nous présentons les travaux les plus importants liés au clustering multi-vues, puis nous introduisons brièvement le contexte théorique TO dans la section 3. La section 4 décrit les approches proposées et les résultats expérimentaux sur des données synthétiques et réelles sont présentés dans la section 5. Enfin, la section 6 conclut l'article et présente quelques directions pour d'éventuelles recherches futures.

2 État de l'art

L'apprentissage multi-vues a été introduit dans par Yarowsky (1995), où il est appliqué pour la désambiguïsation du sens des mots. Principalement, la vue multiple est représentée par deux classificateurs différents : le contexte local d'un mot comme première vue et les sens des autres occurrences de ce mot comme seconde vue. Dans Blum et Mitchell (1998), les auteurs ont introduit une approche de co-apprentissage basée sur deux hypothèses formées sur des vues distinctes. L'algorithme vise à améliorer la qualité globale de l'apprentissage de deux classificateurs avec les instances de confiance les plus élevées à partir de données non étiquetées. Cette approche nécessite que les vues soient indépendantes.

Dans Brefeld et Scheffer (2004), un algorithme co-EM a été présenté comme une version multi-vues de l'algorithme Expectation-Maximization (EM) pour l'apprentissage semi-supervisé. Bickel et Scheffer (2004) ont étudié le problème où l'ensemble d'attributs peut être divisé aléatoirement en deux sous-ensembles. Ces approches cherchent à optimiser l'accord entre les points de vue. Les auteurs ont décrit deux algorithmes différents, un algorithme basé

sur EM qui donne des résultats très significatifs et un algorithme agglomératif multi-vues qui semble être moins efficace que les approches à vue unique.

Il convient de noter qu'il existe deux approches principales en MVC. L'approche générative est basée sur l'apprentissage d'un modèle de mélange. Ce modèle peut être utilisé pour générer de nouvelles données à partir d'un cluster. D'autre part, l'approche discriminante minimise une fonction de coût afin d'estimer une solution optimale de clustering. Il faut noter que, jusqu'à présent, la plupart des algorithmes MVC sont basés sur une approche discriminante, comme le montre l'étude approfondie présentée dans Xu et al. (2013).

Le clustering MVC est également considéré comme une tâche de base pour plusieurs analyses ultérieures en apprentissage automatique, en particulier pour le clustering d'ensemble (Vega-Pons et Ruiz-Shulcloper, 2011), également appelé clustering par consensus ou clustering par agrégation. L'objectif du clustering d'ensemble est de rassembler toutes les informations sur les clusters provenant de différentes sources du même ensemble de données, ou de différentes exécutions du même algorithme de clustering, afin de former un clustering consensuel qui inclut toutes les informations. Cette approche devient un cadre de MVC lorsqu'elle est appliquée à un clustering avec une description multi-vues des données (Xie et Sun, 2013; Tao et al., 2017).

3 Théorie du transport optimal

Le TO a été introduit par Monge (1781) pour résoudre le problème de l'allocation des ressources. Le but initial de cette théorie est de déplacer une particule d'un point à un autre de manière optimale, en minimisant le coût de ce déplacement ou de ce transport. Plus récemment, le problème *Monge* a été relaxé par Kantorovich (2006), où le problème est transformé en appariement de paires de distributions en utilisant la programmation linéaire.

Plus formellement, étant donné deux mesures définies sur deux espaces différents, le problème *Monge-Kantorovich* consiste à trouver un couplage γ défini comme une probabilité conjointe sur le produit des deux espaces. Dans notre cas, nous nous concentrons sur les mesures discrètes en raison de la représentation empirique des distributions. Nous nous référons au livre de Villani (2008) pour plus de détails sur le cas continu et des études mathématiques plus détaillées.

Définition Soit Ω un espace arbitraire avec D une métrique sur cet espace, et $P(\Omega)$ l'ensemble des mesures de probabilité de Borel sur Ω. Pour $p \in [1, \infty)$ et une mesure de probabilité μ et ν dans $P(\Omega)$, la distance p-Wasserstein (Villani, 2008) est donnée par :

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega^2} D(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}} \tag{1}$$

où $\Pi(\mu, \nu)$ est l'ensemble des mesures de probabilité sur Ω^2 avec μ et ν leurs marginales.

Nous ne considérons ici que des distributions discrètes, représentées par des mesures empiriques. Formellement, soit $X_s = \{x_i^s \in \mathbb{R}^n\}_{i=1}^{i=N_s}$ et $X_t = \{x_i^t \in \mathbb{R}^n\}_{i=1}^{i=N_t}$ sont deux familles de points dans Ω, leurs mesures empiriques étant $\mu_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{x_i}$ et $\mu_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \delta_{y_i}$, respectivement définies comme des sommes uniformes de *Dirac*, le problème *Monge-Kantorovich* consiste à trouver un couplage optimal γ comme probabilité conjointe entre μ_s et μ_t sur

$X_s \times X_t$ en minimisant le coût du transport w. r.t une certaine métrique.

Ce problème est basé sur deux éléments principaux : La matrice M des distances par paire entre les instances de X_s et de X_t élevée à la puissance p qui est un paramètre de coût, et le polytope de transport $\Pi(\mu_s, \mu_t) = \left\{\gamma \in \mathbb{R}_+^{N_s \times N_t} \mid \gamma \mathbf{1} = \mu_s, \gamma^T \mathbf{1} = \mu_s\right\}$. Ce problème admet une solution unique γ^* et définit une métrique appelée *distance de Wasserstein* sur l'espace des mesures de probabilité comme suit :

$$W(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} < M, \gamma > \tag{2}$$

où $< ., . >$ est le produit scalaire de Frobenius.

La distance de *Wasserstein* a été très utile récemment, notamment dans les tâches d'apprentissage automatique telles que l'adaptation au domaine (Courty et al., 2014), le clustering (Cuturi et Doucet, 2014), le clustering multiniveau (Ho et al., 2017), le co-clustering (Laclau et al., 2017) le collaborative clustering (Ben Bouazza et al., 2020). La particularité de cette distance est qu'elle prend en compte la géométrie des données à partir des distances entre les instances, ce qui explique son efficacité. D'autre part, en termes de calcul, le succès de cette distance est également dû à Cuturi (2013), qui a introduit une approche basée sur la régularisation par entropie pour approximer les distances TO à l'aide de l'algorithme de Sinkhorn. En effet, bien que la distance *Wasserstein* ait connu un succès très important, en terme de calcul la fonction objectif a toujours souffert d'une convergence très lente, ce qui a poussé Cuturi à proposer une fonction objectif lissée en ajoutant un terme de régularisation entropique qui a été appliqué à TO (Cuturi, 2013) pour augmenter la vitesse de convergence de la fonction objectif originale :

$$\min_{\gamma \in \Pi(\mu_s, \mu_t)} < M, \gamma > -\frac{1}{\lambda} E(\gamma) \tag{3}$$

où $E(\gamma) = -\sum_{i,j}^{N_s, N_t} \gamma_{ij} \log(\gamma_{ij})$ et $\lambda > 0$ un paramètre fixe.

Grâce à cette version régularisée d'TO, les auteurs ont obtenu une solution plus compacte, plus lisse et plus stable que le problème original. De plus, cette formulation permet de résoudre le problème d'TO en utilisant l'algorithme de mise à l'échelle de la matrice de $Sinkhorn - Knopp$.

4 Apprentissage multi-vues basé sur le transport optimal

Dans cette section, nous montrons comment le MVC peut être résolu en utilisant la théorie OT, et comment rassembler toutes les informations issues de toutes les vues pour former un consensus de manière optimale.

4.1 Motivations

Afin de justifier notre approche d'un point de vue théorique, nous expliquons dans cette section les principes fondamentaux du MVC et comment il peut être transformé en un problème

d'OT. Le MVC peut être divisé en deux étapes. L'étape locale, qui consiste essentiellement à trouver un meilleur clustering dans chaque vue, et l'étape globale, qui consiste à agréger ces informations (c'est-à-dire les centroïdes des clusters dans chaque vue) pour former un consensus représentant les informations de toutes les vues en même temps.

Nous considérons $X = \left\{x^1, x^2, .., x^r\right\}$ avec r vues multiples, où $x^v = \{x_1^v, x_2^v, .., x_n^v\}$ avec n points dans Ω dans la $v$ième vue. De manière générale, les approches existantes pour former une vue unifiée ou un consensus consistent à maximiser une fonction objectif qui combine les partitions des clusterings de base $H = \{h_1, h_2, .., h_r\}$ données par un algorithme, afin de trouver une partition de consensus h.

$$\Gamma(h, H) = \sum_{i=1}^{r} w_i U(h, h_i) \tag{4}$$

où $\Gamma : \mathbb{Z}_+^n \times \mathbb{Z}_+^{nr} \mapsto \mathbb{R}$ est la fonction de consensus, et $U : \mathbb{Z}_+^n \times \mathbb{Z}_+^r \mapsto \mathbb{R}$ la fonction d'utilité avec $\sum_i^r w_i = 1$.

La fonction d'utilité U est très importante et doit être choisie avec soin. Ce problème peut être transformé en un problème de minimisation sans changer sa nature en utilisant différentes distances comme la distance de Mirkin (Mirkin, 1987). De plus, Wu et al. (2014) ont prouvé que le problème de consensus est équivalent au problème de K-moyennes sous certaines hypothèses définies dans la définition suivante.

Définition :

Une fonction d'utilité U est une fonction d'utilité de clustering consensuel de type $\mathrm{K}-mean$ si $\forall$ tout $\forall H = \{h_1, \ldots, h_r\}$ et $K \geq 2$, il existe une distance f telle que

$$\max_{h \in H} \sum_{i=1}^{r} w_i U(h, h_i) \Leftrightarrow \min_{h \in H} \sum_{k=1}^{K} \sum_{x_l \in C_k} f(x_l^{(b)}, c_k) \tag{5}$$

est valable pour toute région réalisable H. Où $X^b = \{x_l^b \mid 1 \leq l \leq n\}$ est un ensemble de données binaires dérivé de l'ensemble de r partitionnements de base de H comme suit :

$$x_l^{(b)} =< x_{l,1}^{(b)}, \ldots, x_{l,i}^{(b)}, \ldots, x_{l,r}^{(b)} >, \quad \text{avec} \quad x_{l,i}^{(b)} =< x_{l,i1}^{(b)}, \ldots, x_{l,ij}^{(b)}, \ldots, x_{l,iK}^{(b)} >$$

et $\quad x_{l,ij} = \begin{cases} 1 & \text{si } L_{h_i}(x_l) = j \\ 0 & \text{autrement} \end{cases}$ $\quad$ et $\quad c_k$ sont les centroïdes du cluster.

En partant de l'idée que le clustering consensuel peut être considéré comme un problème de K-moyennes, il peut être converti en un problème de transport optimal basé sur la distance de Wasserstein. Plus précisément, nous pouvons voir chaque vue comme un ensemble de distribution que nous pouvons assembler pour former un consensus optimal, calculé dans l'étape globale.

Nous détaillons l'approche proposée et la manière dont nous améliorons le consensus des clusterings en utilisant la théorie TO dans la section suivante. Nous proposons différents types de consensus qui ont été validés expérimentalement sur plusieurs ensembles de données (Section 5).

TAB. 1 – *Notations*

Notations	
X	l'ensemble des données de toutes les vues, telle que $x_i \in \mathbb{R}^d$
μ	la distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$
X^v	les données de la vue v tel que $x_i \in \mathbb{R}^{d_v}$ avec $d_v < d$
μ^v	la distribution de la vue v $\mu^v = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^v}$
d_v	la dimension de la vue v
c_j^v	le centroïde j dans la vue v entre les données et les centroïdes c_j^v
ν^v	les distributions des centroïdes $\frac{1}{k_v}\sum_{j=1}^{k_v}\delta_{c_j^v}$
$L^v = \{l_{ij}^v\}$	la matrice TO de la vue v
c_k	les centroïdes des clusers du consensus
$L = \{l_{ik}\}$	la matrice TO entre les centroïdes de chaque vue et les centroïdes du consensus
ν	la distribution des centroïdes du consensus $\frac{1}{K}\sum_{k=1}^{K}\delta_{c_k}$
Π	la matrice d'TO entre x_i et c_k

4.2 L'approche proposée

Soit $X = \{X^1, X^2, \ldots, X^r\}, X^v \in \Omega \subset \mathbb{R}^{d\times 1}, 1 \leq v \leq r$ est l'ensemble de données constitué de d attributs numériques. Soit $X^v = \{x_1^v, x_2^v, \ldots, x_n^v\}, x_i^v \in \mathbb{R}^{d_v\times 1}, d_v < d$ le sous-ensemble des attributs traités par la vue v voire tableau 1.

4.2.1 Étape locale :

Nous considérons la mesure empirique des données : $\mu^v = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^v}$, qui représentent les données de chaque vue v. $x_i^v, 1 \leq i \leq n$ dans la vue v est uniformément distribué sur la vue v. Nous cherchons à trouver une mesure de probabilité discrète $\nu^v \in \Sigma_{k_v}$ qui est une approximation de μ^v définie par les centroïdes k_v $C^v = \{c_1^v, c_2^v, \ldots, c_{k_v}^v\}$. A cette fin, nous calculons l'TO de μ^v de Σ_n à Σ_{k_v}, nous définissons donc ν^v comme la solution du problème suivant :

$$\nu^v = \min_{\nu^v \in \Sigma_{k_v}} W_{2,\lambda}^2(\mu^v, \nu^v). \tag{6}$$

Notons que lorsque $d = 1$ et $p = 2$ et sans contrainte sur le poids sur Σ_{k_v}, ce problème est équivalent à l'algorithme de Lloyd (Cuturi et Doucet, 2014). Dans ce qui suit, nous considérons $p = 2$. Afin de résoudre le problème d'optimisation (6), nous procédons de la même manière que pour le clustering par K-means. L'étape locale pour le clustering de la vue v alterne itérativement entre l'affectation de chaque donnée au centroïde le plus proche et l'optimisation des centroïdes $C^v = \{c_1^v, c_2^v, \ldots, c_{k_v}^v\}$.

L'algorithme 1 décrit comment nous regroupons les données localement ; il s'agit d'une alternance entre l'algorithme Sinkhorn pour affecter chaque point de données à son centroïde le plus proche et la mise à jour de la distribution des centroïdes pour qu'elle soit la moyenne pondérée des données qui lui sont affectées.

Il convient de noter que l'algorithme 1 est équivalent à K-means, mais il permet une affectation douce au lieu d'une affectation dure, ce qui signifie que $l_{ij}^v \in [0, \frac{1}{n}]$. De plus, le terme de régularisation $-\frac{1}{\lambda}E(\gamma)$ garantit une solution avec une entropie plus élevée, ce qui signifie que le point sera plus uniformément assigné aux clusters.

Algorithme 1 : Algorithme local

Input : Données de la vue v, $X^v = \{x_i^v\}_{i=1}^n \in \mathbb{R}^{d_v}$ tell que $d_v < d$ et k_v le nombre des clusters
Constante d'entropie λ

Output : Matrice du TO $L^v = \left\{l_{ij}^v\right\}$ et les centroides c_j^v

Initialiser k_v, centroïdes aléatoirement $c^v(0)$ avec la distribution $\nu^v = \frac{1}{k_v}\sum_{j=1}^{k_v}\delta_{c_j}$;

while *ne converge pas (clusters non stables)* **do**

Calculer la matrice TO $L^v = \left\{l_{ij}^v\right\}\, 1 \leq i \leq n, 1 \leq j \leq k_v$;

$$L^v = \min W_{\lambda 2}^2(\mu^v, \nu^v);$$

Mettre à jour les centroïdes de la distribution $c_j^v(t+1)$:

$$c_j^v(t+1) = \sum_i l_{ij} x_i^v \quad 1 \leq j \leq k_v;$$

return $\{L^v\}$ *et* $\left\{c_j^v\right\}_{j=1}^{k_v}$

4.2.2 Etape globale :

L'objectif de l'étape globale est d'assembler toutes les informations que nous obtenons dans chaque vue afin de former un cluster consensuel pour l'ensemble des données. Nous proposons deux approches basées sur la théorie TO : L'approche de projection du consensus et le consensus avec une nouvelle représentation.

Approche de projection par consensus : L'approche de projection par consensus (CPA) consiste à projeter la structure des clusters de chaque vue dans l'espace global. L'idée derrière cette projection est de visualiser la structure de chaque vue dans l'espace global afin d'enrichir l'information des données pour augmenter la qualité du clustering consensuel. Plus précisément, il s'agit d'une sorte de super clustering pour obtenir des prototypes plus précis qui contiennent plus d'informations sur les données. Par conséquent, les matrices de la partition résultant du transport des données seront plus complètes et plus précises et garantiront une meilleure qualité. Dans l'algorithme 2 nous expliquons le mécanisme de cette méthode, la première étape de l'algorithme consiste à projeter les centroïdes de chaque vue dans l'espace global des données et ensuite à calculer un clustering de la projection des centroïdes dans l'espace global en utilisant l'algorithme des K-means de Sinkhon basé sur la distance de Wasserstein, pour obtenir de nouveaux prototypes c_k. La dernière étape de l'algorithme vise à transporter les instances des nouveaux prototypes.

Consensus avec une nouvelle représentation : Le consensus avec une nouvelle représentation (CNR) a pour but d'assembler la structure obtenue dans chaque vue afin de reconstruire une nouvelle représentation des données basée sur la matrice de partition. Cette représentation donne la probabilité postérieure d'appartenance de chaque point à chaque cluster de chaque vue. En d'autres termes, il s'agit d'une sorte de superposition de toutes les vues qui permet de former un clustering consensuel de toutes les informations que nous avons déjà obtenues à partir de chaque vue.

Algorithme 2 : Consensus avec projection (CPA)

Input : $X = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$ représenté par $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ et L_v et $\left\{c_j^v\right\}_{j=1}^{k_v}$ représentée par la distribution ν^v, $1 \leq v \leq r$
Nombre des clusters K
Constante d'entropie λ

Output : Centroïdes du consensus c_k et $\Pi = \{\pi_{ik}\}$ la matrice TO $\forall i$

Initialiser K centroides $c_k(0)$ avec la distribution $\nu = \frac{1}{K}\sum_{k=1}^{K} \delta_{c_k}$;
Calculer les centroïdes de la matrice $C = \{c_k\}$ dans l'espace global :

$$C = L^v X \quad for \quad 1 \leq v \leq r$$

while *ne converge pas (clusters non stables)* **do**
Calculer la matrice TO entre les vues et le consensus $L(t) = \{l_{jk}\}\, 1 \leq j \leq k_v, 1 \leq k \leq K, 1 \leq v \leq r$;

$$L = \min W_{\lambda 2}^2(\nu^v, \nu);$$

Mettre à jour les centroïdes du consensus c_k ;

$$c_k = \sum_{i=1}^{n} l_{ik}.x_i \quad 1 \leq k \leq K;$$

Calculer le TO entre les données et les centroïdes du consensus. c_k ;

$$\pi_{ik} = \min_k W_{\lambda 2}^2(\mu, \nu)$$

return c_k *et* $\Pi = \{\pi_{ik}\}$

L'algorithme 3 explique le processus de cette méthode, la première étape consiste à concaténer toutes les matrices de partitions, puis à calculer un clustering sur cette matrice en utilisant l'algorithme Sinkhorn pour obtenir une meilleure partition des données et des centroïdes qui contiennent les informations émergeant de chaque vue.

5 Résultats expérimentaux

Dans cette section, nous évaluons les approches et les testons sur plusieurs ensembles de données réelles (décrits dans le tableau 2). Nous les comparons également à l'algorithme classique de clustering consensus K-means et au clustering obtenu à partir d'une vue unique. Les tableaux 3 et 4 résument les résultats du clustering des approches proposées (**CNR** et **CPA**) et des autres méthodes en termes d'indices de Davies-Bouldin (DB) et de Rand normalisé (R_n), respectivement. Comme on peut le constater, nos approches sont généralement plus performantes sur tous les ensembles de données selon les deux métriques.

Dans le tableau 3, nous évaluons les deux approches (CNR et CPA), comparées à une approche à vue unique (SVA) utilisant l'indice Davies-Bouldin (DB) (Davies et Bouldin, 1979).

Algorithme 3 : Consensus avec une nouvelle représentation (CNR)

Input : L_v et $\left\{c_j^v\right\}_{j=1}^{k_v}$ représenté par la distribution ν^v, $1 \leq v \leq r$
Nombre de clusters K
Constante d'entropieλ
Output : Les centroïdes du consensus c_k et $\Pi = \{\pi_{ik}\}$ la matrice de TO $\forall i$
Initialiser K centroids $c_k(0)$ avec la distribution $\nu = \frac{1}{K}\sum_{k=1}^{K}\delta_{c_k}$;
Calculer la matrice de la nouvelle représentation des données ;

$$X = \text{concatenation}(L_v) \quad for \quad 1 \leq v \leq r$$

while *note converge(clusters not stable)* **do**
Calculer la matrice TO $\Pi = \{\pi_{ik}\}$;

$$\Pi = \min W_{\lambda 2}^2(\mu, \nu);$$

Mettre à jour les centroïdes du consensus c_k;

$$c_k = \sum_{i=1}^{n} \pi_{ik}.x_i \quad 1 \leq k \leq K;$$

return c_k *et* $\Pi = \{\pi_{ik}\}$

TAB. 2 – *Quelques caractéristiques des ensembles de données expérimentales*

Datasets	#instances	#Attributs	#Classes
Breast	699	9	2
Dermatology	358	33	6
Ecoli	332	7	6
Iris	150	4	3
PenDigits	10992	16	10
Satimage	4435	36	6
Wine	178	13	3

Cet indice évalue la qualité du clustering non supervisé car il est basé sur le rapport de la somme de la dispersion au sein des clusters sur la séparation entre les clusters. Plus la valeur de l'indice DB est faible, meilleure est la qualité du clustering.

Comme nous pouvons le voir dans le tableau 3, la **CNR** a obtenu de meilleures valeurs pour plusieurs ensembles de données. Ceci s'explique par le fait que cette méthode effectue le clustering sur les structures de chaque vue qui garantissent l'amélioration de la qualité du clustering global, alors que **CPA** préforme une assignation forcée des instances aux centroïdes obtenus à partir du clustering de consensus. Nous avons complété l'analyse par un test de Friedman, qui confirme que globalement l'approche **CNR** est plus performante que les autres algorithmes testés. Ces résultats ne sont pas surprenants, car **CNR** utilise une nouvelle représentation des données incluant toutes les structures de toutes les vues.

Dans le tableau 4, nous validons nos approches en comparant le consensus classique basé sur Kmeans par l'indice de Rand normalisé R_n (Wu et al., 2009) qui mesure l'accord entre

TAB. 3 – *Performances de clustering sur sept ensembles de données réelles (indice DB).*

Datasets	CNR	CPA	SVA
Breast	1.735	**1.727**	1.742
Dermatology	1.926	**1.194**	1.310
Ecoli	**1.145**	1.405	1.236
Iris	**0.893**	0.908	0.915
PenDigits	**1.136**	1.334	1.257
Satimage	**1.011**	1.221	1.274
Wine	1.308	**0.556**	**0.556**

deux partitions : une donnée par le processus de clustering et l'autre définie par des critères externes. Les valeurs de R_n sont dans $[0, 1]$. Lorsque la valeur est proche de 1, la qualité du cluster est bien meilleure. Comme nous le voyons, le score le plus élevé est obtenu avec **CPA**. Cela se comprend par le fait que dans la dernière étape de l'algorithme **CPA**, nous forçons l'affectation des instances aux nouveaux centroïdes projetés à partir des vues, ce qui explique l'accord entre les étiquettes prédites à partir du clustering consensuel et les vraies, tandis que sur le **CNR** nous regroupons une nouvelle représentation des données. Il convient de noter que nous choisissons cet indice pour comparer nos approches avec la méthode classique de clustering d'ensemble. Cependant, tant que nous comparons les étiquettes vraies et prédites, cet indice ne met pas en évidence la nature non supervisée du clustering. Pour évaluer davantage les performances, nous calculons un score de mesure (Zhao et Fu, 2015), qui donne une évaluation globale par rapport à tous les ensembles de données. Les résultats montrent que nos approches sont globalement plus performantes que les autres algorithmes.

TAB. 4 – *Performance de clustering sur sept ensembles de données réelles (indice R_n).*

Data sets	CNR	CPA	SVA	KKC
Breast	0.3315	**0.7374**	0.6891	0.0556
Dermatology	**0.4608**	0.1202	0.1472	0.0352
Ecoli	0.2822	0.3447	0.3389	**0.5065**
Iris	0.4423	0.4605	0.4491	**0.7352**
PenDigits	0.5064	**0.6039**	0.5356	0.5347
Satimage	0.3576	**0.4702**	0.4679	0.4501
Wine	**0.2264**	0.2149	0.2149	0.1448
Score	5.2074	**5.5170**	5.3651	4.6340

6 Conclusion

Dans cet article, nous avons proposé une nouvelle façon d'explorer les avantages et l'utilité de la MVC. Nous avons examiné le MVC dans le cadre de la théorie de l'TO et proposé deux

nouveaux algorithmes de clustering qui permettent de combiner les structures découvertes par plusieurs vues, sous la forme d'un consensus. Un algorithme est basé sur des projections et un autre utilise les partitions trouvées par les différentes vues pour créer de nouvelles représentations des données. Des expériences sur sept ensembles de données réelles ont montré l'efficacité des deux algorithmes proposés par rapport à une approche à vue unique et à une autre technique classique de pointe. Les résultats obtenus montrent que l'TO peut fournir pour cette tâche d'apprentissage automatique un cadre formel et une flexibilité algorithmique qui marquent une amélioration des performances par rapport aux approches existantes.

Les travaux futurs comprennent l'étude des capacités d'apprentissage de nouvelles représentations de données grâce aux approches que nous proposons. Nous voulons également voir comment utiliser le paradigme de l'apprentissage profond pour enrichir nos approches multi-vues.

Références

Ben Bouazza, F. E., Y. Bennani, G. Cabanes, et A. Touzani (2020). Collaborative clustering through optimal transport. In *International Conference on Artificial Neural Networks*, pp. 873–885. Springer.

Bickel, S. et T. Scheffer (2004). Multi-view clustering. In *ICDM*, Volume 4, pp. 19–26.

Blum, A. et T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *COLT*, pp. 92–100. ACM.

Brefeld, U. et T. Scheffer (2004). Co-em support vector learning. In *ICML*, pp. 16–24. ACM.

Cabanes, G. et Y. Bennani (2007). A simultaneous two-level clustering algorithm for automatic model selection. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pp. 316–321.

Cabanes, G., Y. Bennani, et D. Fresneau (2012). Enriched topological learning for cluster detection and visualization. *Neural Networks 32*, 186–195. Selected Papers from IJCNN 2011.

Courty, N., R. Flamary, et D. Tuia (2014). Domain adaptation with regularized optimal transport. In *ECML*, pp. 274–289. Springer.

Cuturi, M. (2013). Sinkhorn distances : Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300.

Cuturi, M. et A. Doucet (2014). Fast computation of wasserstein barycenters. In *ICML*, pp. 685–693.

Davies, D. L. et D. W. Bouldin (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* (2), 224–227.

Fu, L., P. Lin, A. V. Vasilakos, et S. Wang (2020). An overview of recent multi-view clustering. *Neurocomputing 402*, 148–161.

Ghassany, M., N. Grozavu, et Y. Bennani (2013). Collaborative multi-view clustering. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Ho, N., X. L. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, et D. Phung (2017). Multilevel clustering via wasserstein means. In *ICML*, pp. 1501–1509.

Kantorovich, L. V. (2006). On the translocation of masses. *Journal of Mathematical Sciences 133*(4), 1381–1382.

Laclau, C., I. Redko, B. Matei, Y. Bennani, et V. Brault (2017). Co-clustering through optimal transport. In D. Precup et Y. W. Teh (Eds.), *Proceedings of the 34th ICML*, Volume 70 of *Proceedings of Machine Learning Research*, pp. 1955–1964. PMLR.

Mirkin, B. (1987). Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification 4*(1), 7–31.

Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale.

Tao, Z., H. Liu, S. Li, Z. Ding, et Y. Fu (2017). From ensemble clustering to multi-view clustering. In *IJCAI*.

Vega-Pons, S. et J. Ruiz-Shulcloper (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence 25*(03), 337–372.

Villani, C. (2008). *Optimal transport : old and new*, Volume 338. Springer Science & Business Media.

Wu, J., H. Liu, H. Xiong, J. Cao, et J. Chen (2014). K-means-based consensus clustering : A unified view. *IEEE transactions on knowledge and data engineering 27*(1), 155–169.

Wu, J., H. Xiong, et J. Chen (2009). Adapting the right measures for k-means clustering. In *SIGKDD*, pp. 877–886. ACM.

Xie, X. et S. Sun (2013). Multi-view clustering ensembles. In *International Conference on Machine Learning and Cybernetics*, Volume 1, pp. 51–56.

Xu, C., D. Tao, et C. Xu (2013). A survey on multi-view learning. *arXiv preprint arXiv :1304.5634*.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*.

Zhao, H. et Y. Fu (2015). Dual-regularized multi-view outlier detection. In *IJCAI*.

Summary

In this work , we present a new framework for multi-view clustering formalized in optimal transport theory, where the main idea is to learn local models from each view using a Sinkhorn-based clustering, and to look for an overall clustering of all views to obtain a consensus model. To do this, we propose two approaches based on the regularized Wasserstein distance and the Wasserstein barycenter: a Consensus Projection Approach (CPA) which learn a consensus on the original space and a Consensus with New Representation (CNR) where the main idea is to rely on a new consensus distribution learned from the distribution of the different views. The two approaches are compared to other multi-view clustering approach, through a set of experiments that demonstrate the efficiency of our methods.

Keywords:Multi-View Clustering (MVC), Co-training, Subspace clustering, Multi-task learning, Transport Optimal(TO), Wasserstein/Earth Mover's distance.

Fouille de Motifs Fermés et Diversifiés Basée sur la Relaxation

Arnold Hien**, Samir Loudni***, Noureddine Aribi * Yahia Lebbah*, Amine Laghzaoui*
Abdelkader Ouali**, Albrecht Zimmermann**

*Université Oran1, Lab. LITIO, 31000 Oran, Algeria
**Normandie Univ., UNICAEN, CNRS – UMR GREYC, France
***TASC (LS2N-CNRS), IMT Atlantique, FR – 44307 Nantes, France

Résumé. Dans cet article, nous proposons une nouvelle approche basée sur la programmation par contraintes pour l'extraction de motifs fréquents fermés et diversifiés (Hien et al., 2020a). La diversité est contrôlée par une contrainte de seuil sur l'indice de Jaccard. Nous montrons que cette mesure n'a pas de propriété de monotonie, ce qui rend le processus d'extraction infaisable. Pour y remédier, nous proposons une nouvelle contrainte globale, CLOSEDDIVERSITY, qui exploite une relaxation anti-monotone de l'indice de Jaccard pour élaguer les motifs non diversifiés. Une seconde relaxation, basée sur une borne supérieure, est exploitée via une nouvelle heuristique de branchement.

Mots-clés : Fouille de motifs, Contrainte globale, Diversité, Jaccard, Relaxation

1 Introduction

Ces dernières années, la fouille de motifs a changé peu à peu de paradigme pour évoluer vers un modèle plus centré utilisateur. Il s'agit de prendre en compte les préférences de l'utilisateur afin de guider la recherche vers des motifs plus intéressants pour lui. Cela est rendu possible par l'introduction de mécanismes de feedback qui permettent à l'utilisateur de spécifier ses préférences sur les motifs extraits (Dzyuba et van Leeuwen, 2013). Un élément important de ce paradigme est la capacité à pouvoir présenter rapidement à l'utilisateur des motifs diversifiés. En effet, lorsque les motifs sont similaires, ou si l'extraction des motifs prend beaucoup de temps, l'utilisateur risque de se lasser et il devient alors difficile pour lui d'exprimer ses préférences. Nous proposons une nouvelle approche déclarative exploitant la programmation par contraintes (PPC) pour extraire efficacement des motifs fréquents, fermés et diversifiés. L'utilisation de la PPC est motivée par son caractère déclaratif, permettant de combiner plusieurs contraintes au même temps, et par la richesse du langage de contraintes qu'elle offre. De plus, la PPC permet une gestion générique des variables et des contraintes ainsi que l'utilisation d'algorithmes efficaces de filtrage, ce qui permet une construction itérative des motifs.

Les travaux précédents sur l'extraction de motifs diversifiés ont proposé l'utilisation d'un post-traitement sur les motifs déjà extraits (voir (Knobbe et Ho, 2006)). Van Leeuwen et Knobbe (2012) ont quant à eux proposé d'utiliser une approche heuristique. Bosc et al. (2018); Belfodil et al. (2019) ont au contraire introduit la diversité dans le processus d'extraction de

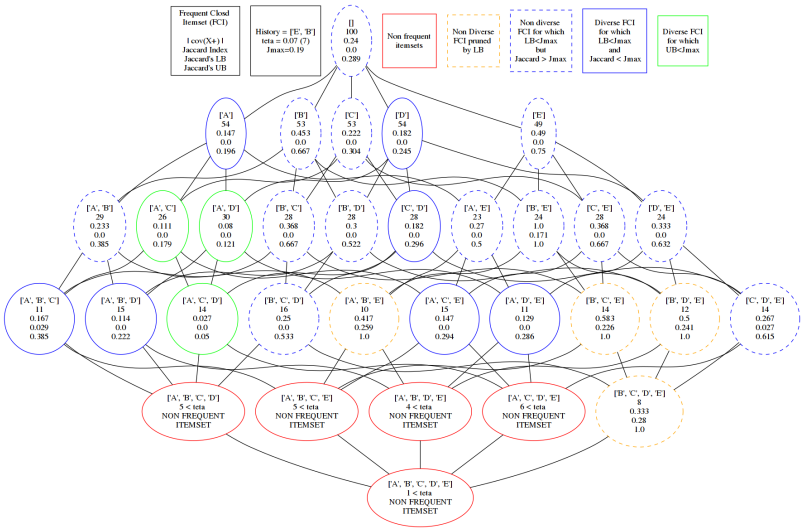


FIG. 1: Le treillis des motifs fréquents fermés associé à la base transactionnelle $\mathcal{D}$ de l'exemple 1.

motifs. Cette dernière approche nécessite d'ajouter des contraintes supplémentaires pour assurer la diversité en élaguant les motifs non diversifiés.

Dans cet article, nous utilisons la programmation par contraintes pour extraire efficacement les motifs fréquents fermés et diversifiés. La diversité est contrôlée par une contrainte de seuil sur l'indice de Jaccard. Nous montrons que cette mesure n'a pas de propriété de monotonie, ce qui rend le processus d'extraction infaisable. Pour y remédier, nous proposons une nouvelle contrainte globale, CLOSEDDIVERSITY, qui exploite une relaxation anti-monotone de l'indice de Jaccard pour élaguer les motifs non diversifiés. Une seconde relaxation, basée sur une borne supérieure, est exploitée via une nouvelle heuristique de branchement.

2 Préliminaires

2.1 Fouille d'itemset

Soit $\mathcal{I}$ un ensemble de n *items*, un *motif* P est un sous-ensemble non vide de $\mathcal{I}$. Une base transactionnelle $\mathcal{D}$ est un multi-ensemble de transactions sur $\mathcal{I}$, où chaque *transaction* t est un sous-ensemble de $\mathcal{I}$, i.e., $t \subseteq \mathcal{I}$. Un motif P *apparaît* dans une transaction t, ssi $P \subseteq t$. La *couverture* de P dans $\mathcal{D}$ est l'ensemble des transactions dans lesquelles il apparaît : $\boldsymbol{t}(P) = \{t \in \mathcal{D} \mid P \subseteq t\}$. Le *support* de P dans $\mathcal{D}$ est le cardinal de sa couverture : $sup(P)$ = $|\boldsymbol{t}(P)|$. Un motif P est dit *fréquent* si son support dépasse un seuil de fréquence minimal θ, $sup(P) \geq \theta$. La *clôture* d'un motif P, notée $Clos(P)$, est l'ensemble des items communs à toutes les transactions dans $\boldsymbol{t}(P)$: $Clos(P) = \{i \in \mathcal{I} \mid \forall t \in \boldsymbol{t}(P), i \in t\}$. Un motif P est dit *fermé* ssi $Clos(P) = P$.

Example 1 *La figure 1 montre le treillis de motifs fréquents fermés dérivés d'une base transactionnelle ayant* 5 *items et* 100 *transactions, avec* $\theta = 7$.

2.2 Mesure de Diversité

L'indice de Jaccard est une mesure de similarité classique sur les ensembles. Nous l'utilisons pour quantifier le chevauchement des couvertures entre deux motifs.

Definition 1 (Indice de Jaccard) *Soient deux motifs P et Q, l'indice de Jaccard mesure la proportion de chevauchement entre les couvertures des deux motifs :* $Jac(P,Q) = \frac{|\boldsymbol{t}(P) \cap \boldsymbol{t}(Q)|}{|\boldsymbol{t}(P) \cup \boldsymbol{t}(Q)|}$.

Un indice de Jaccard plus petit est synonyme d'une faible similarité en termes de couverture entre motifs et peut donc être utilisé comme mesure de diversité entre paires de motifs.

Definition 2 (Contrainte de Diversité/Jaccard) *Soient P et Q deux motifs. Étant donné la mesure Jac et un seuil de diversité J_{max}, on dit que P et Q sont diversifiés entre eux ssi $Jac(P,Q) \leq J_{max}$. Nous noterons cette contrainte c_{Jac}.*

Notre objectif est d'exploiter la contrainte de Jaccard durant la recherche pour élaguer les motifs non-diversifiés. Pour cela, nous maintenons un *historique* $\mathcal{H}$ de motifs extraits pendant la recherche et qui sont diversifiés entre eux. Les prochains motifs P extraits devront alors respecter la contrainte c_{Jac} par rapport à chaque motif $H \in \mathcal{H}$.

Definition 3 (k motifs fréquents et diversifiés) *Étant donné un historique $\mathcal{H} = \{H_1, \ldots, H_k\}$ de k motifs fréquents, fermés et diversifiés, la mesure Jac et un seuil de diversité J_{max}, le problème consiste à trouver de nouveaux motifs P tel que $\forall\, H \in \mathcal{H}$, $Jac(P,H) \leq J_{max}$.*

Example 2 *Le treillis de la figure 1 montre un ensemble de motifs fréquents fermés et diversifiés (représentés par des cercles bleus et verts) obtenus avec $J_{max} = 0.19$ et $\mathcal{H} = \{BE\}$. ACE est un motif fréquent, fermé et diversifié (i.e., $Jac(ACE, BE) = 0.147 < 0.19$).*

Proposition 1 *Soient P, Q et P' trois motifs avec $P \subset P'$. $Jac(P,Q)$ peut être plus petit, égal ou supérieur à $Jac(P',Q)$.*

Comme l'indique la proposition 1, la contrainte de Jaccard n'est ni monotone ni anti-monotone, ce qui implique un élagage limité lors de la recherche. Pour faire face à ce problème, nous proposons deux relaxations anti-monotones : (i) Une relaxation par la borne inférieure, permettant d'élaguer les motifs non-diversifiés lors de la recherche, (ii) une relaxation par la borne supérieure pour trouver les items menant vers des motifs diversifiés.

2.3 Programmation par contrainte

La programmation par contraintes (PPC) offre une approche générique pour modéliser les problèmes combinatoires. Un modèle PPC consiste en un ensemble de variables $X = \{x_1, \ldots, x_n\}$, un ensemble de domaines finis D pour chaque variable $x_i \in X$, et un ensemble de contraintes $\mathcal{C}$ sur X. Une contrainte $c \in \mathcal{C}$ est une relation entre différentes variables $X(c)$, qui précise les combinaisons possibles de valeurs pour ces variables. Une instanciation d'un sous-ensemble de variables $Y \subseteq X$ est une affectation de valeurs $v \in dom(x_i)$ à chaque variable x_i. Une solution est alors une instanciation de X satisfaisant toutes les contraintes $\mathcal{C}$. Pour la résolution, les solveurs utilisent des méthodes de recherche par retour-arrière pour explorer l'espace de recherche et instancier progressivement les variables. L'algorithme 1 donne le schéma général de résolution. À chaque nœud, *Recherche* sélectionne une

Algorithme 1 : Recherche(D)

```
1  In : X : variables de décision ; C : contraintes ;
2  InOut : D : domaines des variables ;
3  begin
4      D ← Filtrage(D, C)
5      if il existe x_i ∈ X t.q. dom(x_i) est vide then
6          return Echec
7      if il existe x_i ∈ X t.q. |dom(x_i)| > 1 then
8          Selectionner x_i ∈ X t.q. |dom(x_i)| > 1
9          forall v ∈ dom(x_i) do
10             Recherche(Dom ∪ {x_i ← {v}})
11     else
12         retourner la solution D
```

variable non instanciée (ligne 8) selon l'heuristique définie par l'utilisateur et l'instancie avec une valeur (ligne 9). Lorsqu'une instanciation ne respecte pas toutes les contraintes (lorsqu'un des domaines devient vide), un retour-arrière a lieu (ligne 5). On obtient une solution (ligne 12) lorsque tous les domaines $dom(x_i)$ ne contiennent que des singletons et que toutes les contraintes sont respectées. Afin d'accélérer la recherche, des *algorithmes de filtrages* sont utilisés. En effet, à chaque instanciation d'une variable à une valeur de son domaine, l'algorithme de filtrage réduit l'espace de recherche tout en garantissant une certaine propriété de consistance comme la *consistance de domaine*. La consistance de domaine garantit que pour chaque variable x_i d'une contrainte c ($x_i \in X(c)$) et pour chaque $v \in dom(x_i)$, il existe une instanciation $(x_i = v)$ qui satisfait c.

2.4 Modèle PPC pour la fouille de motifs fermés

Le premier modèle PPC utilisé pour la fouille de motifs fréquents et fermés a été proposé par De Raedt et al. (2008). Ce modèle est basé sur des contraintes réifiées (Apt, 2003) faisant intervenir les items et les transactions d'un jeu de données. Par la suite, Lazaar et al. (2016) ont proposé la première contrainte globale pour produire des motifs fréquents et fermés. Ils utilisent un vecteur x de variables booléennes $(x_1, \ldots, x_{|\mathcal{I}|})$ pour représenter les motifs. Chaque variable x_i représente la présence de l'item $i \in \mathcal{I}$ dans le motif. Nous utiliserons les notations suivantes : $x^+ = \{i \in \mathcal{I} \,|\, dom(x_i) = \{1\}\}$ l'ensemble des items présents, $x^- = \{i \in \mathcal{I} \,|\, dom(x_i) = \{0\}\}$ l'ensemble des items absents et $x^* = \{i \in \mathcal{I} \,|\, i \notin x^+ \cup x^-\}$.

Definition 4 (CLOSEDPATTERNS) *Soit x un vecteur de variables booléennes, θ un seuil de support minimum et $\mathcal{D}$ un jeu de données. La contrainte globale* CLOSEDPATTERNS$_{\mathcal{D},\theta}(x)$ *est vérifiée si et seulement si x^+ est à la fois fermé et fréquent.*

Definition 5 (Extension propre (Wang et al. (2003))) *Un motif non nul P est une extension propre de Q ssi $\boldsymbol{t}(P \cup Q) = \boldsymbol{t}(Q)$.*

Règles de filtrage. Lazaar et al. (2016) ont proposé trois règles de filtrage pour CLOSEDPATTERNS. La première règle permet d'étendre un motif x^+ avec un item i lorsque $x^+ \cup \{i\}$ est une extension propre de x^+ (voir Définition 5). Dans ce cas, on supprime la valeur 0 de $dom(x_i)$. La seconde règle permet de vérifier la fréquence du motif $x^+ \cup \{i\}$ et de supprimer la valeur 1

de $dom(x_i)$ si son support est inférieur au seuil θ. La troisième règle supprime la valeur 1 de $dom(x_i)$ lorsque $\boldsymbol{t}(x^+ \cup \{i\}) \subset \boldsymbol{t}(x^+ \cup \{j\})$, avec j un item absent ($j \in x^-$).

3 Extraction de Motifs Fréquents Fermés et Diversifiés

Cette section présente deux relaxations anti-monotones de l'indice de Jaccard : (i) Une relaxation par la borne inférieure pour élaguer les motifs non-diversifiés lors de la recherche, (ii) une relaxation par la borne supérieure pour trouver les items menant vers des motifs diversifiés. Les preuves des différentes propositions sont disponibles dans (Hien et al., 2020b).

3.1 Reformulation du problème

La proposition 1 établie que la contrainte de Jaccard n'est ni monotone ni anti-monotone. Nous proposons alors d'approximer la contrainte c_{Jac} par la collection de motifs solutions de sa relaxation : c^r_{Jac} : $Th(c_{Jac}) \subseteq Th(c^r_{Jac})$. Notre approche consiste à formuler une contrainte relâchée, ayant une propriété de monotonie, qui sera utilisée pour élaguer l'espace de recherche. Plus précisément, nous proposons d'exploiter des bornes inférieure et supérieure de l'indice de Jaccard dans le but de dériver une relaxation anti-monotone de c_{Jac}.

Definition 6 (Relaxation de l'indice de Jaccard) *Soit un historique $\mathcal{H} = \{H_1, \ldots, H_k\}$ de k motifs fréquents, fermés et diversifiés, un seuil de diversité J_{max}, une borne inférieure LB_J et une borne supérieure UB_J de l'indice de Jaccard, la relaxation du problème d'extraction de motifs diversifiés consiste à trouver les motifs candidats P tels que $\forall\, H \in \mathcal{H}$, $LB_J(P, H) \leq J_{max}$. La contrainte de Jaccard est satisfaite lorsque $\forall H \in \mathcal{H}, UB_J(P, H) \leq J_{max}$.*

3.2 Borne inférieure de l'indice de Jaccard

À partir de la définition 1, nous formulons une borne inférieure de l'indice de Jaccard qui minimise le chevauchement entre les couvertures des deux motifs et maximise la couverture propre de chaque motif.

Definition 7 (Couverture propre) *Soient P et Q deux motifs. La couverture propre de P par rapport à Q est définie par : $\boldsymbol{t}^{pr}_Q(P) = \boldsymbol{t}(P) \backslash \{\boldsymbol{t}(P) \cap \boldsymbol{t}(Q)\}$.*

Une borne inférieure LB de Jaccard minimise le numérateur et maximise le dénominateur du quotient donné dans la définition 1.

Proposition 2 (Borne inférieure LB) *Soit un motif $H \in \mathcal{H}$, P un motif partiel en cours de construction tel que $sup(P) \geq \theta$, et $\boldsymbol{t}^{pr}_H(P)$ la couverture propre de P par rapport à H. $LB_J(P, H) = \frac{\theta - |\boldsymbol{t}^{pr}_H(P)|}{|\boldsymbol{t}(P)| + |\boldsymbol{t}(H)| + |\boldsymbol{t}^{pr}_H(P)| - \theta}$ est une borne inférieure de $Jac(P, H)$.*

Cette borne inférieure de Jaccard nous permet de filtrer des motifs non diversifiés, c'est à dire ceux qui ont un LB_J supérieur à J_{max}. Ces motifs sont appelés des *témoins négatifs*.

Example 3 *Dans la figure 1, les motifs non diversifiés avec un LB_J supérieur à $J_{max} = 0.19$ sont représentés avec la couleur orange.*

Proposition 3 (Monotonie de LB_J) *Soit $H \in \mathcal{H}$ un motif. Pour tout motif P et Q tel que $P \subseteq Q$, alors nous avons $LB_J(P,H) \leq LB_J(Q,H)$.*

La propriété 3 est très importante car elle établit une condition nécessaire pour pouvoir filtrer les motifs non diversifiés (voir Section 3.4). En effet, lorsque $LB_J(P,H) > J_{max}$, alors aucun motif $Q \supseteq P$ ne pourra satisfaire la contrainte de Jaccard, ce qui rend la contrainte anti-monotone. On pourra donc filtrer le motif Q.

3.3 Borne supérieure de l'indice de Jaccard

En relâchant la contrainte de Jaccard et en approximant sa théorie $Th(c_{Jac})$ par $Th(c^r_{Jac})$ ($Th(c_{Jac}) \subseteq Th(c^r_{Jac})$), il est possible d'extraire des motifs P tel que $LB_J(P,H) < J_{max}$ alors que $Jac(P,H) > J_{max}$ (voir Figure 1). Pour remédier à cette situation (cas de faux positifs), nous définissons une borne supérieure UB de Jaccard qui évalue la satisfaction de la contrainte. Ainsi, les motifs P tels que $UB_J(P,H) \leq J_{max}$, $\forall H \in \mathcal{H}$ vont satisfaire la contrainte de Jaccard et seront appelés *témoins positifs*.

La borne UB a été construite en prenant la démarche inverse de celle de la borne inférieure : nous maintenons le numérateur $\boldsymbol{t}(H) \cap \boldsymbol{t}(P)$ inchangé et nous réduisons l'ensemble $\boldsymbol{t}^{pr}_H(P)$ afin de maximiser le dénominateur $\boldsymbol{t}(H) \cup \boldsymbol{t}(P)$. Ainsi, si l'intersection est supérieure ou égale à θ, les futurs motifs P' couvriront uniquement des transactions de l'intersection. Dans le cas contraire, le dénominateur devra contenir quelques transactions de $\boldsymbol{t}^{pr}_H(P)$ (exactement $\theta - |\boldsymbol{t}(H) \cap \boldsymbol{t}(P)|$ transactions).

Proposition 4 (Borne supérieure UB) *Étant donné un motif $H \in \mathcal{H}$, et un motif P tel que $sup(P) \geq \theta$. $UB_J(P,H) = \frac{|\boldsymbol{t}(H) \cap \boldsymbol{t}(P)|}{|\boldsymbol{t}^{pr}_P(H)| + \max\{\theta, |\boldsymbol{t}(H)| \cap |\boldsymbol{t}(P)|\}}$ est une borne supérieure de $Jac(P,H)$.*

Example 4 *Dans la figure 1, les motifs diversifiés avec un UB_J inférieur à $J_{max} = 0.19$ sont représentés avec la couleur verte.*

Notre borne supérieure UB peut être utilisée pour évaluer la contrainte de Jaccard pendant l'extraction des motifs. En effet, pendant l'étape d'énumération des motifs, lorsqu'un motif candidat P a une borne supérieure de Jaccard inférieure à J_{max} alors la contrainte c_{Jac} est satisfaite. Par ailleurs, comme nous le montrons dans la proposition 5, la borne supérieure UB est anti-monotone. De ce fait, tous les motifs $Q \supseteq P$ serons aussi diversifiés.

Proposition 5 (Anti-monotonie de UB_J) *Soit H un motif de l'historique $\mathcal{H}$. Pour tous les motifs P et Q, tels que $P \subseteq Q$, nous avons $UB_J(P,H) \geq UB_J(Q,H)$.*

3.4 Contrainte globale CLOSEDDIVERSITY

La contrainte globale CLOSEDDIVERSITY exploite la relaxation LB de l'indice de Jaccard pour extraire des motifs fréquents, fermés et diversifiés.

Definition 8 (CLOSEDDIVERSITY) *Soit x un vecteur de variables booléennes, $\mathcal{H}$ un historique de motifs fréquents, fermés et diversifiés (initialement vide), θ un seuil de support, J_{max} un seuil de diversité et $\mathcal{D}$ un jeu de données. La contrainte* CLOSEDDIVERSITY$_{\mathcal{D},\theta}(x, \mathcal{H}, J_{max})$ *est vérifiée si et seulement si : (1) x^+ est fermé ; (2) x^+ est fréquent, $sup(x^+) \geq \theta$; (3) x^+ est diversifié, $\forall H \in \mathcal{H}, LB_J(x^+, H) \leq J_{max}$.*

Algorithme 2 : Filtrage pour CLOSEDDIVERSITY

```
1  In : θ, J_max : seuils de fréquence et de diversité ; H : historique des solutions trouvées ;
2  InOut : x = {x_1 ... x_n} : variables booléennes ;
3  begin
4      if ( |t(x+)| < θ ∨ !PGrowth_LB(x+, H, J_max)) then return false;
5      foreach i ∈ x* do
6          if (|t(x+ ∪ {i})| < θ) then
7              dom(x_i) ← dom(x_i) − {1}; x⁻_Freq ← x⁻_Freq ∪ {i}; x* ← x* \ {i}; continue;
8          if (|t(x+ ∪ {i})| = |t(x+)|) then
9              dom(x_i) ← dom(x_i) − {0}; x+ ← x+ ∪ {i};    x* ← x* \ {i};
10         if (! PGrowth_LB(x+ ∪ {i}, H, J_max)) then
11             dom(x_i) ← dom(x_i) − {1}; x⁻_Div ← x⁻_Div ∪ {i}; x* ← x* \ {i}; continue;
12         foreach k ∈ (x⁻_Freq ∪ x⁻_Div) do
13             if (t(x+ ∪ {i}) ⊆ t(x+ ∪ {k})) then
14                 dom(x_i) ← dom(x_i) − {1}
15                 if k ∈ x⁻_Freq then x⁻_Freq ← x⁻_Freq ∪ {i};
16                 else x⁻_Div ← x⁻_Div ∪ {i};
17                 x* ← x* \ {i}; break;
18     return true;
19 Function PGrowth_LB(x, H, J_max) : Booléen
20     foreach H ∈ H do
21         if (LB_J(x, H) > J_max) then return false
22     return true
```

L'historique $\mathcal{H}$ est mis à jour de façon itérative en y ajoutant les motifs extraits avec CLOSEDDIVERSITY. La condition (3) est une condition nécessaire pour assurer la diversité des motifs. Nous montrerons en section 3.5 comment exploiter la borne supérieure UB pour garantir la satisfaction de la contrainte globale. CLOSEDDIVERSITY exploite les règles de filtrage de CLOSEDPATTERNS (voir Sect. 2.4) auxquels nous avons ajouté nos règles détaillées ci-dessous. On notera par x^-_{Freq} l'ensemble des variables non fréquentes, et par x^-_{Div} l'ensemble des variables filtrées par la règle LB.

Proposition 6 (Règles de filtrage) *Soit $\mathcal{H} = \{H_1, \ldots, H_k\}$ un historique de k motifs fréquents, fermés et diversifiés, x une instanciation partielle des variables et une variable non instanciée $i \in x^*$, la variable i sera filtrée si l'un des deux cas suivants est vérifié :*
1) si $\exists\, H \in \mathcal{H}$ s.t. $LB_J(x^+ \cup \{i\}, H) > J_{max}$, alors on filtre 1 du domaine $dom(x_i)$ de i.
2) si $\exists\, k \in x^-_{Div}$ s.t. $\boldsymbol{t}(x^+ \cup \{i\}) \subseteq \boldsymbol{t}(x^+ \cup \{k\})$, alors $LB_J(x^+ \cup \{i\}, H) > LB_J(x^+ \cup \{k\}, H) > J_{max}$ et on filtre 1 du domaine $dom(x_i)$ de i. Ces deux cas indiquent que le motif $x^+ \cup \{i\}$ n'est pas diversifié et sera donc filtré.

Algorithme. Le propagateur de CLOSEDDIVERSITY prend en paramètre les variables x, le support minimum θ, le seuil de diversité J_{max} et l'historique $\mathcal{H}$ initialement vide. Il commence par vérifier si le motif partiel x^+ est fréquent en comparant sa couverture à θ. Il teste également sa diversité avec la fonction $\mathcal{P}Growth_{LB}$. Si le motif n'est pas fréquent ou s'il n'est pas diversifié, alors la contrainte globale n'est pas respectée et la branche explorée est abandonnée (ligne 4). Par la suite, l'algorithme 2 applique les règles de filtrage de CLOSEDPATTERNS (voir Section 2.4) auxquelles on ajoute une règle de filtrage avec LB_J. Ainsi, $\forall H \in \mathcal{H}$, la

valeur de $LB_J(x^+ \cup \{i\}, H)$ est évaluée avec la fonction $\mathcal{P}Growth_{LB}(x^+ \cup \{i\}, \mathcal{H}, J_{max})$. S'il existe un motif H tel que $LB_J(x^+ \cup \{i\}, H) > J_{max}$ (ligne 21), alors la variable x_i est filtrée (la valeur 1 est supprimée de son domaine) car le motif $x^+ \cup \{i\}$ ne conduira pas vers un motif diversifié (ligne 11). On met à jour x^-_{Div} et x^*, et on répète l'opération sur les autres variables non instanciées. De même, lorsque la couverture d'un motif $x^+ \cup \{i\}$ est incluse dans la couverture du motif $x^+ \cup \{k\}$, tel que $k \in (x^-_{Freq} \cup x^-_{Div})$ (lignes 12-17), alors la variable i est filtrée.

Proposition 7 (Consistance de domaine et complexité) *L'algorithme 2 supprime toutes les valeurs inconsistantes avec une complexité temporelle en* $\mathcal{O}(n^2 \times m)$.

3.5 Motifs témoins et Fréquences estimées

Fréquences estimées. La fréquence d'un motif peut être calculée en faisant l'intersection des couvertures des items qui le constituent puis calculer leur cardinalité : $sup(x^+) = |\cap_{i \in x^+} \boldsymbol{t}(i)|$. Pour limiter les nombreuses et coûteuses opérations d'intersection qui correspondent à des *OU logiques*, nous proposons d'estimer la fréquence de chaque item $i \in \mathcal{I}$ en fonction des items du motif x^+. Cette estimation, notée $eSup_{\mathcal{D}}(i, x^+)$, constitue une *borne inférieure* de $|\boldsymbol{t}(x^+ \cup \{i\})|$. De ce fait, lorsque $eSup_{\mathcal{D}}(i, x^+) \geq \theta$ alors $|\boldsymbol{t}(x^+ \cup \{i\})| \geq \theta$. Ainsi, la fréquence du motif n'est calculée que lorsque $eSup_{\mathcal{D}}(i, x^+) < \theta$, ce qui nous permet des gains de performance non négligeables. Par ailleurs, avec les fréquences estimées, nous proposons une nouvelle heuristique de choix de variables notée MINCOV. Elle consiste, pour une itération donnée, à étendre le motif partiel courant avec la variable qui, à l'itération précédente, avait la plus petite fréquence estimée. En effet, ces variables sont les plus susceptibles d'activer rapidement une règle de filtrage (voir Algorithme 2) et donc de réduire l'espace de recherche.
Témoins positifs. Durant la recherche, nous calculons de façon incrémentale $UB(x^+ \cup \{i\}, H)$ de chaque extension du motif partiel x^+. Ainsi, avec la propriété d'anti-monotonie de UB_J (voir Proposition 5), si $\forall H \in \mathcal{H}, UB(x^+ \cup \{i\}, H) < J_{max}$ alors tous les sur-motifs de $x^+ \cup \{i\}$ satisferont la contrainte de Jaccard. Cette propriété nous permet de déduire une nouvelle heuristique de choix de variable que nous notons FIRSTWITCOV et qui consiste à étendre le motif partiel courant avec la variable i qui a un UB_J inférieur au seuil J_{max}.

4 Résultats expérimentaux

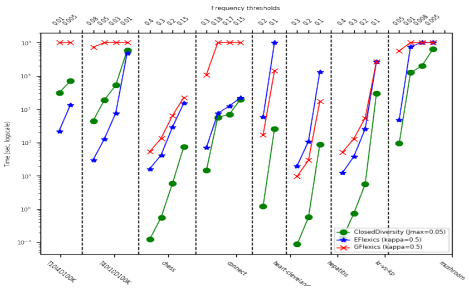
Nous avons évalué notre méthode en nous intéressant aux trois points suivants : (1) le temps d'exécution et le nombre de motifs générés : pour cela, nous comparons CLOSEDDIV avec CLOSEDP et FLEXICS de (Dzyuba et al., 2017) ; (2) la qualité des motifs de CLOSEDDIV par rapport à ceux générés avec CLOSEDP et FLEXICS ; (3) la qualité de nos bornes LB/UB : nous avons mesuré la distance qu'il y a entre ces deux bornes et l'indice de Jaccard. Nous avons utilisé les jeux de données UCI (fimi.ua.ac.be/data) et avons choisi des jeux de données de différentes tailles et densités. Certains jeux de données, comme HEPATITIS et CHESS sont très denses (resp. 50% et 49%). D'autres au contraire sont très peu denses, comme T10I4D100K et RETAIL (resp. 1% and 0.06%). Les expérimentations ont été menées sur une machine avec un processeur AMD Opteron 6174 de 2.2 GHz ayant 256 Go de RAM et avec une limite de temps d'exécution de 24 heures. Nous avons sélectionné pour chaque dataset des seuils de fréquence pour avoir différents nombres de motifs fermés et fréquents ($|Th(c)| \leq 15000$,

Dataset $\lvert\mathcal{I}\rvert \times \lvert\mathcal{T}\rvert$ $\rho(\%)$	$\theta(\%)$	#Motifs		Temps (s)		#Nœuds	
		(1)	(2)	(1)	(2)	(2)	(2)
CHESS	20	22,808,625	**96**	2838.30	**5.87**	45,617,249	**436**
75 × 3196	15	50,723,131	**393**	5666.03	**75.40**	101,446,261	**1,855**
49.33%	10	OOM	**4,204**	OOM	**3825.29**	OOM	**18,270**
HEPATITIS	30	83,048	**12**	9.64	**0.09**	166,095	**29**
68 × 137	20	410,318	**57**	42.00	**0.57**	820,635	**162**
50.00%	10	1,827,264	**2,270**	**169.59**	**76.91**	3,654,527	**5,256**
KR-VS-KP	30	5,219,727	**17**	682.94	**0.74**	10,439,453	**82**
73 × 3196	20	21,676,719	**96**	2100.79	**5.64**	43,353,437	**448**
49.32%	10	OOM	**4,120**	OOM	**3035.49**	OOM	**17,861**
CONNECT	30	460,357	**18**	1666.14	**14.81**	920,713	**77**
129 × 67557	18	2,005,476	**197**	5975.44	**573.66**	4,010,951	**900**
33.33%	15	3,254,780	**509**	9534.07	**1989.35**	6,509,559	**2,188**
HEART-CLEVELAND	10	12,774,456	**3,496**	1308.63	**257.39**	25,548,911	**7,977**
95 × 296	8	23,278,687	**12,842**	**2278.97**	2527.38	46,557,373	**28,221**
47.37%	6	43,588,346	58,240	**4126.84**	46163.06	87,176,691	124,705
SPLICE1	10	1,606	**422**	**6.55**	25.25	3,211	**843**
287 × 3190	5	31,441	**8,781**	**117.15**	5616.47	62,881	**17,594**
20.91%	2	589,588	-	**1179.55**	-	1,179,175	-
MUSHROOM	5	8,977	**727**	**10.02**	60.70	17,953	**1,704**
112 × 8124	1	40,368	**12,139**	**34.76**	12532.95	80,735	**25,154**
18.75%	0.5	62,334	**27,768**	**50.05**	64829.06	124,667	**56,873**
T40I10D100K	8	138	127	**75.91**	447.20	275	253
942 × 100000	5	317	288	**331.47**	1561.34	633	575
4.20%	1	65,237	7,402	**5574.31**	58613.88	130,473	14,887
PUMSB	40	-	**4**	-	**57.33**	-	**16**
2113 × 49046	30	-	**15**	-	**267.72**	-	**64**
3.50%	20	-	**52**	-	**852.39**	-	**250**
T10I4D100K	5	11	11	**1.73**	6.31	21	21
870 × 100000	1	386	**361**	**434.25**	3125.06	771	722
1.16%	0.5	1,074	**617**	**881.31**	7078.90	2,147	**1,257**
BMS1	0.15	1,426	**609**	**11362.71**	68312.38	2,851	**1,220**
497 × 59602	0.14	1,683	**668**	**11464.93**	68049.00	3,365	**1,339**
0.51%	0.12	2,374	**823**	**13255.79**	79704.88	4,747	**1,651**
RETAIL	5	17	**13**	**10.74**	33.44	33	25
16470 × 88162	1	160	**111**	**297.21**	1625.73	319	227
0.06%	0.4	832	**528**	**6073.53**	31353.23	1,663	**1,093**

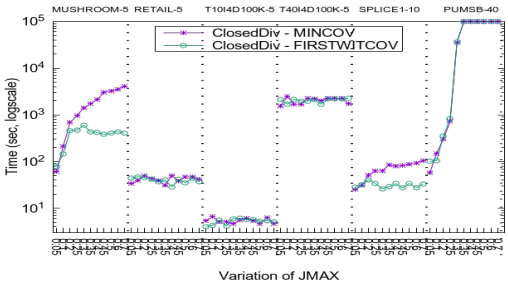
TAB. 1: CLOSEDDIV ($J_{max} = 0.05$) vs CLOSEDP. Pour les colonnes #Motifs and #Noeuds, les valeurs en gras indiquent une réduction dépassant 20% du nombre total de motifs et nœuds." − " s'affiche lorsque la limite de temps est dépassée. OOM : Mémoire insuffisante. (1) : CLOSEDP (2) : CLOSEDDIV

$30000 \leq |Th(c)| \leq 10^6$, and $|Th(c)| > 10^6$). La seule exception concerne les datasets très volumineux et peu denses RETAIL et PUMSB, où le nombre de solutions est petit. Nous avons utilisé CLOSEDPATTERNS comme base de référence pour déterminer les seuils appropriés utilisés par CLOSEDDIV. Pour évaluer la qualité de nos motifs, nous proposons de calculer l'ECR (Exclusive Coverage Ratio) qui mesure le taux moyen de couverture propre des motifs extraits : $ECR(P_1, ..., P_k) = avg_{1 \leq i \leq k}(\frac{sup(P_i) - |\boldsymbol{t}(P_i) \cap \bigcup_{j \neq i} \boldsymbol{t}(P_j)|}{sup(P_i)})$.

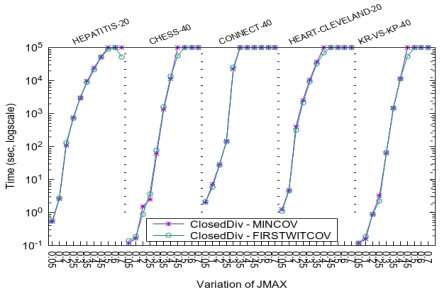
(a) Comparaisons entre CLOSEDDIV, CLOSEDP et FLEXICS. La table 1 compare CLOSEDDIV et CLOSEDP pour différentes valeurs de θ. Pour CLOSEDDIV, nous avons pris un seuil de diversité J_{max} égal à 0.05 et avons choisi MINCOV comme heuristique de choix de variables. Comme nous pouvons le constater, CLOSEDDIV permet de réduire considérablement le nombre de motifs, surtout avec les jeux de données denses et pour des valeurs faibles de θ. Ainsi, pour CHESS, CLOSEDDIV permet une réduction de près de 99% du nombre de motifs par rapport à CLOSEDP (de $\sim 50 \cdot 10^6$ à 393 motifs) pour $\theta = 15\%$. Ce résultat peut s'expliquer par la densité des jeux de données qui induisent de nombreuses redondances dans la



(a) Temps d'exécution de MINCOV et FLEXICS



(b) Jeux de données non denses.



(c) Jeux de données denses.

Jeux de données	$\theta(\%)$	Exclusive Coverage Ratio (ECR)		
		CLOSEDP	CLOSEDDIV–MINCOV	EFLEXICS
T40I10D100K	8	1.6E-01	**4.4E-01**	1.29E-01
	1	1.9E-01	**9.2E-01**	2.90E-01
SPLICE1	10	4.0E-02	**1.6E-01**	2.98E-02
	5	4.0E-02	**2.5E-01**	1.79E-01
CONNECT SPLICE1	18	0.00	**1.0E-01**	4.31E-02
	10	4E-02	**1.6E-01**	2.98E-02
	5	4E-2	**2.5E-01**	1.79E-01
CONNECT	18	1.25E-6	**1.0E-01**	4.31E-02
	15	0.00	**1.7E-01**	4.16E-02
MUSHROOM	5	3.3E-4	**5.6E-01**	2.63E-01
	1	3.73E-4	**4.3E-01**	4.2E-01
HEPATITIS	10	0.00	**2.5E-01**	1.0E-01
HEPATITIS	10	1.50E-3	**2.5E-01**	1.0E-01
T10I4D100K	1	5.9E-01	**9.3E-01**	5.72E-01
	0.5	4.8E-01	**9.6E-01**	4.63E-01

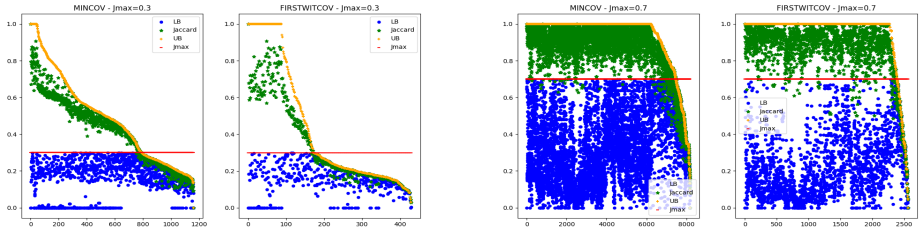
(d) ECR.

FIG. 2: Analyse des temps d'exécutions (MINCOV vs FIRSTWITCOV et CLOSEDDIV vs FLEXICS) et ECR

couverture des motifs fermés. Avec ces jeux de données, le temps d'exécution de CLOSEDDIV est également plus court car le filtrage des motifs redondants permet de réduire considérablement l'espace de recherche. Avec les jeux de données moins denses, la réduction du nombre de motifs par CLOSEDDIV est plus faible. En effet, avec ces jeux de données, il y a peu de motifs fermés et donc moins de redondances. Par conséquent, les calculs de LB_J pénalisent grandement CLOSEDDIV qui devient plus lent comparé à CLOSEDP.

La figure 2a montre une comparaison des temps d'exécution entre CLOSEDDIV et FLEXICS. FLEXICS est un outil permettant de faire de l'échantillonnage de motifs. Pour cela, il partitionne l'espace de recherche en cellules avec des contraintes XOR puis utilise un oracle pour faire un tirage pondéré des motifs dans différentes cellules. Il se décline en deux versions, EFLEXICS qui utilise ECLAT comme oracle et GFLEXICS qui utilise plutôt CP4IM (De Raedt et al., 2008). La partition de l'espace en cellules permet d'introduire une diversité dans les données, d'où notre intérêt pour cette approche. D'une part, CLOSEDDIV est plus rapide que GFLEXICS avec plus d'un ordre de grandeur. D'autre part, EFLEXICS est plus rapide que GFLEXICS, et notre approche est presque toujours classée en premier, illustrant son utilité pour extraire des motifs diversifiés de manière *anytime*.

(b) Impact de la variation de J_{max}. Nous avons étudié les deux heuristiques de choix de variables MINCOV et FIRSTWITCOV pour différentes valeurs du seuil J_{max}. Dans les figures 2b et 2c, nous pouvons constater que, de façon générale, le temps d'exécution augmente avec le



(a) SPLICE1 ($\theta = 10\%$, $J_{max} = 0.3$) (b) MUSHROOM ($\theta = 5\%$, $J_{max} = 0.7$)

FIG. 3: Analyse qualitative de LB et UB

seuil J_{max}. En effet, avec un seuil de diversité plus grand, CLOSEDDIV génère plus de motifs, ce qui induit un sur-coût dans le calcul de LB et UB. Nous pouvons constater néanmoins que lorsque J_{max} devient assez grand, CLOSEDDIV filtre beaucoup moins (voir figure 3). Par ailleurs, nous constatons que les deux heuristiques MINCOV et FIRSTWITCOV ont presque les même performances, FIRSTWITCOV étant meilleur sur certaines instances. De plus, on peut constater (voir l'annexe complémentaire (Hien et al., 2020b)) que FIRSTWITCOV permet de générer des motifs de meilleure qualité grâce à sa capacité de guider la recherche vers des motifs témoins positifs.

(c) Analyse qualitative des bornes LB et UB. La figure 3a montre l'évolution des valeurs de LB_J, Jac et UB_J des motifs générés par MINCOV et FIRSTWITCOV pour les jeux de données SPLICE1 et MUSHROOM. Pour les différentes courbes, les motifs ont été triés par ordre décroissant de leur UB_J. Nous constatons ainsi que la valeur de LB est toujours en-dessous de J_{max}. Concernant la valeur de UB_J, on peut constater qu'elle est toujours très proche de celle du Jaccard, ce qui dénote une bonne relaxation pour notre borne supérieure. Par ailleurs, nous pouvons noter que l'heuristique FIRSTWITCOV permet de générer rapidement des motifs de bonne qualité (avec un UB inférieur à J_{max}). Ces résultats démontrent l'intérêt de notre borne supérieure UB_J et de l'heuristique FIRSTWITCOV à générer des motifs plus diversifiés.

(d) Analyse qualitative des motifs. La figure 2d compare les ECR de CLOSEDDIV, CLOSEDP et FLEXICS, un ECR important traduit une plus grande diversité dans la couverture des motifs. Pour pallier au grand nombre de motifs générés par CLOSEDP qui complique l'évaluation de l'ECR, nous avons décidé d'effectuer le calcul sur des échantillons de $k = 10$ motifs, et répéter l'évaluation 100 fois afin d'évaluer un grand nombre de motifs. Nous reportons une moyenne des ECR calculés. CLOSEDDIV conduit clairement à des solutions avec une plus grande diversité entre motifs. Ceci est indicatif de motifs dont la couverture est (approximativement) mutuellement exclusive.

5 Conclusions

Dans ce papier, nous avons proposé une contrainte globale qui exploite deux relaxations LB/UB (anti-)monotones de l'indice de Jaccard pour l'extraction de motifs fréquents, fermés et diversifiés. La diversité est contrôlée par une contrainte de seuil mesurant la similarité des

occurrences des motifs. Nos expérimentations ont montré que notre approche permet de réduire de façon significative le nombre de motifs par rapport à ceux générés par la contrainte globale CLOSEDPATTERNS. Par ailleurs, les ensembles de motifs générés sont plus diversifiés.

Références

Apt, K. (2003). *Principles of Constraint Programming*. USA : Cambridge University Press.

Belfodil, A., A. Belfodil, A. Bendimerad, P. Lamarre, C. Robardet, M. Kaytoue, et M. Plantevit (2019). Fssd-a fast and efficient algorithm for subgroup set discovery. In *Proceedings of DSAA*, pp. 91–99.

Bosc, G., J.-F. Boulicaut, C. Raïssi, et M. Kaytoue (2018). Anytime discovery of a diverse set of patterns with monte carlo tree search. *Data mining and knowledge discovery 32*(3), 604–650.

De Raedt, L., T. Guns, et S. Nijssen (2008). Constraint programming for itemset mining. In *14th ACM SIGKDD*, pp. 204–212.

Dzyuba, V. et M. van Leeuwen (2013). Interactive discovery of interesting subgroup sets. In *International Symposium on Intelligent Data Analysis*, pp. 150–161. Springer.

Dzyuba, V., M. van Leeuwen, et L. De Raedt (2017). Flexible constrained sampling with guarantees for pattern mining. *Data Mining and Knowledge Discovery 31*(5), 1266–1293.

Hien, A., S. Loudni, N. Aribi, Y. Lebbah, M. Laghzaoui, A. Ouali, et A. Zimmermann (2020a). A relaxation-based approach for mining diverse closed patterns. In *Proceedings of ECML PKDD 2020*, Volume 12457, pp. 36–54. Springer.

Hien, A., S. Loudni, N. Aribi, Y. Lebbah, M. Laghzaoui, A. Ouali, et A. Zimmermann (June 2020b). Supplementary Material : https://github.com/lobnury/ClosedDiversity.

Knobbe, A. J. et E. K. Ho (2006). Pattern teams. In *Proceedings of ECML-PKDD*, pp. 577–584. Springer.

Lazaar, N., Y. Lebbah, S. Loudni, M. Maamar, V. Lemière, C. Bessiere, et P. Boizumault (2016). A global constraint for closed frequent pattern mining. In *Proceedings of the 22nd CP*, pp. 333–349.

Van Leeuwen, M. et A. Knobbe (2012). Diverse subgroup set discovery. *Data Mining and Knowledge Discovery 25*(2), 208–242.

Wang, J., J. Han, et J. Pei (2003). CLOSET+ : searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the Ninth KDD*, pp. 236–245. ACM.

Summary

In this paper, we use constraint programming (well suited to user-centric mining due to its rich constraint language) to efficiently mine a diverse set of closed patterns. Diversity is controlled through a threshold on the Jaccard similarity measure. We show that the Jaccard measure has no monotonicity property, which prevents usual pruning techniques and makes classical pattern mining unworkable. This is why we propose antimonotonic lower and upper bound relaxations, which allow effective pruning, with an efficient branching rule, boosting the whole search process. We show experimentally that our approach significantly reduces the number of patterns and is very efficient in terms of running times, particularly on dense datasets.

Keywords: Pattern mining, Global Constraint, Diversity, Jaccard, Relaxations

Vers l'extraction efficace des représentations condensées de motifs; Application aux motifs Pareto Dominants

Charles Vernerey**, Samir Loudni**, Noureddine Aribi * Yahia Lebbah*

*University of Oran1, Lab. LITIO, 31000 Oran, Algeria
**TASC (LS2N-CNRS), IMT Atlantique, FR – 44307 Nantes, France

Résumé. Nous proposons dans ce papier un cadre générique basé sur la programmation par contraintes pour découvrir des représentations condensées de motifs par rapport à un ensemble de mesures. Nous montrons comment notre cadre peut être combiné avec des contraintes de Pareto dominance afin de découvrir des motifs non dominés. Les expérimentations menées sur différents jeux de données démontrent l'efficacité de notre approche et les avantages significatifs qu'elle présente comparé aux approches existantes.

1 Introduction

La fouille de motifs vise à découvrir des régularités intéressantes dans des bases de données (Novak et al., 2009; Wrobel, 1997). La majorité des approches existantes réalise une énumération complète des motifs qui respectent un ensemble de contraintes. Cependant, le nombre important de motifs rend l'analyse de ces derniers très compliquée pour l'utilisateur. Une solution à ce problème repose sur le principe de *représentation condensée*. Cette approche a été utilisée principalement avec la mesure de fréquence (Calders et al., 2004) et il y a peu d'études sur les autres mesures (Giacometti et al., 2002; Soulet et al., 2004). Soulet et Crémilleux (2008) ont étendu le principe de représentation condensée de motifs à un ensemble de mesures. Ils ont proposé l'algorithme MICMAC pour la fouille de représentations condensées adéquates grâce à un nouvel opérateur de fermeture basé sur la notion de *fonction condensable*. Cependant, le problème principal de cette approche est son passage à l'échelle.

Les autres approches pour réduire le nombre de motifs sont basées sur les préférences de l'utilisateur. L'approche la plus populaire est la procédure top-k, qui retourne les k meilleurs motifs par rapport à une mesure choisie par l'utilisateur (Ke et al., 2009; Wang et al., 2005). Récemment, de nouvelles méthodes pour intégrer les idées qui viennent de l'analyse multi-critères tel que la *Pareto dominance*, ou *skylines*, ont été proposées. Soulet et al. (2011) ont utilisé la notion de requêtes skylines (Börzsönyi et al., 2001) afin de découvrir les motifs Pareto (skypatterns). L'approche proposée, intitulée AETHERIS, exploite une représentation condensée adéquate de motifs et la notion de *skylineabilité* afin de réduire le temps d'exécution. Dans (Ugarte et al., 2014), une méthode (intitulée CP+SKY) qui utilise des contraintes dynamiques a été proposée. Elle exploite un modèle réifié pour encoder les motifs (Raedt et al., 2008). Cependant, l'utilisation de contraintes réifiées dans le modèle constitue un frein majeur pour son passage à l'échelle.

Récemment, la programmation par contraintes (PPC) a été utilisée avec succès pour modéliser différents problèmes de fouilles de données (Guns et al., 2011; Hien et al., 2020; Lazaar et al., 2016). L'avantage principal d'utiliser la PPC pour la fouille de données est sa déclarativité et sa flexibilité, ce qui permet d'ajouter de nouvelles contraintes spécifiées par l'utilisateur sans avoir à modifier le système sous-jacent. Dans cet article, nous proposons une nouvelle contrainte globale pour extraire efficacement des représentations condensées adéquates de motifs par rapport à un ensemble de mesures. Cela est possible grâce à un opérateur de fermeture qui exploite le concept de mesure préservante. Nous démontrons ensuite l'utilité de notre contrainte globale pour la découverte de skypatterns. Enfin, nous présentons une étude expérimentale qui compare notre approche à celles existantes pour la fouille de représentations condensées adéquates de motifs et la fouille de skypatterns afin de démontrer son efficacité et son passage à l'échelle.

2 Préliminaires

2.1 Fouille de motifs

Soit $\mathcal{I} = \{1, ..., n\}$ un ensemble de n *items*, un *motif* P est un sous-ensemble non vide de $\mathcal{I}$. Le langage des motifs correspond à $\mathcal{L}_\mathcal{I} = 2^\mathcal{I} \backslash \emptyset$. Un jeu de données transactionnel $\mathcal{D}$ est un ensemble de transactions, où chaque *transaction* t est un sous ensemble de $\mathcal{I}$, i.e., $t \subseteq \mathcal{I}$; $\mathcal{T} = \{1, ..., m\}$ est un ensemble de m indices de *transactions*. Un motif P *apparaît* dans une transaction t, ssi $P \subseteq t$. La *couverture* de P dans $\mathcal{D}$ est l'ensemble des transactions dans lesquelles il apparaît : $\boldsymbol{t}_\mathcal{D}(P) = \{t \in \mathcal{D} \mid P \subseteq t\}$. Le *support* de P dans $\mathcal{D}$ est la taille de sa couverture : $sup_\mathcal{D}(P) = |\boldsymbol{t}_\mathcal{D}(P)|$. Un motif P est dit *fréquent* dans $\mathcal{D}$ si $sup_\mathcal{D}(P) \geq \theta$, où θ est un seuil minimal fixé par l'utilisateur. Étant donné $T \subseteq \mathcal{D}$, $\boldsymbol{i}(T)$ est l'ensemble des items qui sont communs à toutes les transactions de T : $\boldsymbol{i}(T) = \{i \in \mathcal{I} \mid \forall t \in T, i \in t\}$. On définit par $\boldsymbol{c}$ un *opérateur de fermeture*, tel que $\boldsymbol{c}(P) = \boldsymbol{i} \circ \boldsymbol{t}(P) = \boldsymbol{i}(\boldsymbol{t}(P))$. La *fermeture* d'un motif P est l'ensemble des items qui sont contenus dans toutes les transactions de $\boldsymbol{t}(P)$: $\boldsymbol{c}(P) = \{i \in \mathcal{I} \mid \forall t \in \boldsymbol{t}(P), i \in t\}$. Un motif P est dit *clos* (Pasquier et al., 1999) ssi $\boldsymbol{c}(P) = P$. L'opérateur de fermeture permet de définir les classes d'équivalence, et donc la représentation *condensée* des motifs.

Plusieurs mesures basées sur la fréquence sont utilisées afin d'évaluer l'intérêt d'un motif. Soit $\mathcal{D}$ un jeu de données partitionné en deux sous-ensembles $\mathcal{D}_1$ et $\mathcal{D}_2$. Le taux de croissance (gr_1) est une mesure qui permet de mettre en valeur les motifs dont la fréquence augmente significativement d'un sous-jeu de données à l'autre (Novak et al., 2009). Le support disjonctif d'un motif X est $sup_\vee(X) = |\{t \in \mathcal{D} \mid \exists i \in X : i \in t\}|$ et $size$ sa cardinalité. Des informations additionnelles (tel que des valeurs numériques associées aux items) peuvent également être utilisées. Étant donné une fonction $val : \mathcal{I} \to \mathbb{R}_+$, nous l'étendons à un motif X et nous notons $X.val$ l'ensemble $\{val(i) \mid i \in X\}$. Ce type de fonction peut être utilisé avec les primitives usuelles telles que sum, min et max. Par exemple, $sum(X.val)$ est la somme des val pour chaque item de X.

La condensation des motifs vient du fait qu'il y a des dépendances entre eux. Le concept de *mesure préservante* (Soulet et Crémilleux, 2008), qui révèle cette dépendance entre un motif et ses spécialisations, est à la base des représentations condensées basées sur la fermeture.

Trans.	Items						
t_1		B			E	F	$\mathcal{D}_1$
t_2		B	C	D			
t_3	A				E	F	
t_4	A	B	C	D	E		
t_5		B	C	D	E		$\mathcal{D}_2$
t_6		B	C	D	E	F	
t_7	A	B	C	D	E	F	

(a)

Item	val
A	30
B	40
C	10
D	40
E	70
F	50

Name	Definition
$area$	$X \mapsto sup(X) \times size(X)$
$mean$	$X \mapsto \frac{\min(X.val)+\max(X.val)}{2}$
min	$X \mapsto \min(X.val)$
$size$	$X \mapsto \lvert X \rvert$
$bond$	$X \mapsto \frac{sup(X)}{sup_\vee(X)}$
gr_1	$X \mapsto \frac{\lvert\mathcal{D}_2\rvert}{\lvert\mathcal{D}_1\rvert} \times \frac{sup_{\mathcal{D}_1}(X)}{sup_{\mathcal{D}_2}(X)}$

(b)

Itemset	$\langle sup, \min \rangle$	Itemset	$\langle sup, \min \rangle$
B	$\langle 6, 40 \rangle$	E	$\langle 6, 70 \rangle$
AE	$\langle 3, 30 \rangle$	BD	$\langle 5, 40 \rangle$
BE	$\langle 5, 40 \rangle$	EF	$\langle 4, 50 \rangle$
AEF	$\langle 2, 30 \rangle$	BCD	$\langle 5, 10 \rangle$
BDE	$\langle 4, 40 \rangle$	BEF	$\langle 3, 40 \rangle$
$ABDE$	$\langle 2, 30 \rangle$	$BCDE$	$\langle 4, 10 \rangle$
$BDEF$	$\langle 2, 40 \rangle$	$ABCDE$	$\langle 2, 10 \rangle$
$ABDEF$	$\langle 1, 30 \rangle$	$BCDEF$	$\langle 2, 10 \rangle$
$ABCDEF$	$\langle 1, 10 \rangle$		

(c)

TAB. 1: Un jeu de données transactionnel (a). Une valeur est associée à chaque item. Examples de mesures (b). Les motifs fréquents et clos par rapport à $M = \{sup, \min\}$ et leurs valeurs pour $\theta = 1$ (c).

Définition 1 (Mesure préservante) *Une mesure m est dite préservante ssi $\forall i \in \mathcal{I}$ et pour chaque motif $P \subseteq Q$ si $m(P \cup \{i\}) = m(P)$ alors $m(Q \cup \{i\}) = m(Q)$.*

Proposition 1 (Quelques mesures préservantes) $\min$*,* sup*,* $sup_\vee$*,* $\max$*,* $mean$ *et* sum *sont des mesures préservantes.*

Un *opérateur de fermeture adéquat* à des mesures autre que la fréquence a été proposé dans (Soulet et Crémilleux, 2008). Cette opérateur exploite la notion de mesure préservante.

Définition 2 (Opérateur de fermeture adéquat) *Soit M un ensemble de mesures préservantes. La fermeture d'un motif P par rapport à M, noté $\texttt{clos}_M(P)$, est l'ensemble d'items tel que $\texttt{clos}_M(P) = \{i \in \mathcal{I} \mid \forall m \in M, m(P \cup \{i\}) = m(P)\}$.*

Proposition 2 (Motifs clos) *Soit M un ensemble de mesures préservantes, $\texttt{clos}_M$ est un opérateur de fermeture. De plus, P est clos par rapport à M ssi $\texttt{clos}_M(P) = P$.*

Exemple 1 *Soit le jeu de données de la table 1 et $M = \{sup, \min\}$. B est un motif clos par rapport à M car $\texttt{clos}_M(B) = B$, i.e. $\nexists i \in \mathcal{I}$ tel que $sup(B) = sup(B \cup \{i\}) \wedge min(B.val) = min(B \cup \{i\}.val)$. Cependant, A n'est pas un motif clos par rapport à M car $\texttt{clos}_M(A) = AE$, i.e. il existe AE tel que $sup(A) = sup(AE) = 3$ et $min(A.val) = min(AE.val) = 30$. La table 1c montre les 17 motifs clos par rapport à $M = \{sup, \min\}$ avec $\theta = 1$.*

2.2 Fouille de skypatterns

Définition 3 (Dominance Pareto) *Soit $M = \{m_1, \ldots, m_n\}$ un ensemble de n mesures et $N = \{1, .., n\}$ un ensemble d'indices. Un motif P est caractérisé par un vecteur d'utilité $u(P) = (m_1(P), ..., m_n(P)) \in \mathbb{R}^n$. On compare généralement les vecteurs d'utilité à l'aide d'une relation de dominance Pareto ($\mathcal{P}$-dominance). La weak-$\mathcal{P}$-dominance $\succsim_\mathcal{P}$ entre deux motifs P, P' est définie par : $P \succsim_\mathcal{P} P' \Leftrightarrow [\forall i \in N, m_i(P) \geq m_i(P')]$, tandis que la strict $\mathcal{P}$-dominance $\succ_\mathcal{P}$ entre P et P' est définie par : $P \succ_\mathcal{P} P' \Leftrightarrow [P \succsim_\mathcal{P} P' \wedge not(P' \succsim_\mathcal{P} P)]$.*

Une solution P^* est Pareto-optimale (a.k.a *Skypattern*) ssi il n'existe pas de motif Q qui domine P^*. La $\mathcal{P}$-dominance peut être exprimée ainsi : $max\ \{(m_1(P), ..., m_n(P)) : P \in \mathcal{S}\}$, où $\mathcal{S} \subseteq \mathcal{L}_\mathcal{I}$ est l'ensemble des solutions possibles.

Exemple 2 *Considérons l'exemple dans la table 1a avec M=$\{sup, min\}$. Le motif E domine le motif B par rapport à M car $sup(B)$=$sup(E)$=6 et $min(E.val)$>$min(B.val)$.*

Définition 4 (Opérateur Sky) *Étant donné un ensemble de motifs $\mathcal{S} \subseteq \mathcal{L}_\mathcal{I}$ et un ensemble de mesures M, un skypattern de $\mathcal{S}$ par rapport à M est un motif de $\mathcal{S}$ qui n'est pas dominé par rapport à M. L'opérateur de motifs Pareto Sky($\mathcal{S}$,M) retourne tous les skypatterns de $\mathcal{S}$ par rapport à M : Sky($\mathcal{S}$,M) $= \{P \in \mathcal{S} \mid \nexists Q \in \mathcal{S}, Q \succsim_\mathcal{P} P\}$.*

Le problème de fouille de skypatterns peut être formulé ainsi : *Étant donné un ensemble de mesures M, le problème consiste à évaluer la requête $Sky(\mathcal{L}_\mathcal{I}, M)$*. Le problème de fouille de skypatterns est difficile en raison du nombre exponentiel de candidats potentiels (i.e. $|\mathcal{L}_\mathcal{I}|$) Yang (2004). Pour réduire le coût d'évaluation de la requête $Sky(\mathcal{L}_\mathcal{I}, M)$, nous proposons d'appliquer l'opérateur Sky sur un ensemble réduit mais pertinent de motifs $S \subseteq \mathcal{L}_\mathcal{I}$ qui contient tous les motifs Pareto, *i.e.* $S \subseteq Sky(\mathcal{L}_\mathcal{I}, M)$. Les représentations condensées peuvent être utilisées pour réduire le temps de calcul sans perte de précision (Ugarte et al., 2017).

Skylinéabilité. Bien que les représentations condensées réduisent le temps de calcul, pour certaines mesures, telles que $area$ ou $size$, la représentation condensée est égale à $\mathcal{L}_\mathcal{I}$. Par conséquent, calculer une représentation condensée pour chaque mesure $m \in M$ rendrait le processus d'extraction non efficace. Pour résoudre ce problème, Soulet et al. (2011) ont proposé la notion de *skylinéabilité*. L'idée majeure est de trouver un ensemble plus petit de mesures $M' \subseteq M$ tel que les motifs Pareto à M peuvent être récupérés à partir de la représentation condensée par rapport à M'. Un opérateur, noté $\bar{c}$, permettant d'obtenir M' à partir de M est introduit dans Soulet et al. (2011). Il retourne un ensemble de mesures M' qui garantit que pour tout motif $P \subset Q$, si $P =_{M'} Q$, alors $Q \succeq_M P$ (voir (Ugarte et al., 2017) pour plus de détails).

2.3 Programmation par contraintes

Un modèle PPC consiste en un ensemble de variables $X = \{x_1, \ldots, x_n\}$, un ensemble de domaines finis D pour chaque variable $x_i \in X$, et un ensemble de contraintes $\mathcal{C}$ sur X. Une contrainte $c \in \mathcal{C}$ est une relation qui spécifie les combinaisons autorisées de valeurs pour les variables $X(c)$. Une instanciation d'un sous-ensemble de variables $Y \subseteq X$ est une affectation de valeurs $v \in dom(x_i)$ à chaque variable x_i. Une solution est une instanciation de X satisfaisant toutes les contraintes $\mathcal{C}$. Les solveurs de contraintes utilisent des méthodes de recherche par retour-arrière pour explorer l'espace de recherche. Le concept principal utilisé pour accélérer la recherche est la propagation de contraintes à l'aide d'*algorithmes de filtrage*. En effet, à chaque instanciation d'une variable, l'algorithme de filtrage réduit l'espace de recherche tout en garantissant certaines propriétés de consistance comme la *consistance de domaine*. La consistance de domaine garantit que pour chaque variable x_i d'une contrainte c ($x_i \in X(c)$) et pour chaque $v \in dom(x_i)$, il existe une instanciation $(x_i = v)$ qui satisfait c.

Un modèle PPC pour la fouille de motifs clos et fréquents. Le premier modèle PPC pour la fouille de motifs clos et fréquents a été introduit dans (Guns et al., 2011). Il est basé sur des contraintes réifiées qui connectent les variables d'items aux variables de transactions. La première contrainte globale CLOSEDPATTERNS pour la fouille de motifs clos et fréquents a été proposée dans (Lazaar et al., 2016). La contrainte globale COVERSIZE pour le calcul exact de la couverture d'un motif a été proposée dans (Schaus et al., 2017).

Contrainte globale CLOSEDPATTERNS. La majorité des méthodes déclaratives utilisent un vecteur x de variables booléennes $(x_1, \ldots, x_{|\mathcal{I}|})$ pour représenter les motifs, où x_i représente la présence de l'item $i \in \mathcal{I}$ dans le motif. Nous utiliserons la notation suivante : $x^+ = \{i \in \mathcal{I} \mid dom(x_i) = \{1\}\}$, $x^- = \{i \in \mathcal{I} \mid dom(x_i) = \{0\}\}$ and $x^* = \{i \in \mathcal{I} \mid i \notin x^+ \cup x^-\}$.

Définition 5 (CLOSEDPATTERNS) *Soit x un vecteur de variables booléennes, θ un support minimal et $\mathcal{D}$ un jeu de données. La contrainte globale* CLOSEDPATTERNS$_{\mathcal{D},\theta}(x)$ *est respectée ssi x^+ est clos par rapport à $\{sup\}$ et fréquent par rapport à θ.*

Filtrage de CLOSEDPATTERNS. Lazaar et al. (2016) ont aussi introduit un algorithme de filtrage complet pour CLOSEDPATTERNS basé sur trois règles. La première règle filtre 0 de $dom(x_i)$ si $\{i\}$ est une extension de fermeture [1] de x^+. La seconde règle filtre 1 de $dom(x_i)$ si le motif $x^+ \cup \{i\}$ est infréquent par rapport à θ. Enfin, la troisième règle filtre 1 de $dom(x_i)$ si $\boldsymbol{t}(x^+ \cup \{i\})$ est un sous ensemble de $\boldsymbol{t}(x^+ \cup \{j\})$ où j est un item absent, i.e. $j \in x^-$.

3 La contrainte globale ADEQUATECLOSURE

Cette section présente une nouvelle contrainte globale ADEQUATECLOSURE pour la fouille de motifs fréquents et clos par rapport à un ensemble de mesures préservantes M. Les preuves des différentes propositions sont disponibles dans (Vernerey et al., 2021).

Définition 6 (ADEQUATECLOSURE) *Soit x un vecteur de variables booléennes, f et f_1 deux variables entières, θ un support minimum, $\mathcal{D}$ un jeu de données transactionnel, et M un ensemble de mesures préservantes. La contrainte globale* ADEQUATECLOSURE$_{\mathcal{D},M,\theta}(x, f, f_1)$ *est respectée ssi* $\texttt{clos}_M(x^+) = x^+$ *et x^+ est frequent par rapport à θ (i.e. $f \geq \theta$).*

Les variables f et f_1 permettent de stocker les valeurs de $sup(x^+)$ et $sup_{\mathcal{D}_1}(x^+)$. Elles sont utilisées pour imposer des contraintes sur le support et le taux de croissance du motif. Nous introduisons l'opérateur d'inclusion de fermeture $\texttt{cl}_{inc}$ qui est utilisé par notre contrainte globale pour la fouille de représentations condensées adéquates par rapport à M.

Définition 7 (Closure inclusion) *Soit x une instanciation partielle des variables $x_1, \ldots, x_{|\mathcal{I}|}$, M un ensemble de mesures préservantes et i un item libre (i.e. $i \in x^*$).* $\texttt{cl}_{inc}(x^+, i, M)$ *retourne* **vrai** *ssi $\forall m \in M, m(x^+ \cup \{i\}) = m(x^+)$, i.e.* $\texttt{cl}_{inc}(x^+, i, M) \Leftrightarrow i \in \texttt{clos}_M(x^+)$.

Le lemme 1 caractérise une instanciation partielle cohérente par rapport à la contrainte ADEQUATECLOSURE, c'est à dire une instanciation partielle qui peut être étendue à une instanciation complète qui satisfait la contrainte.

Lemme 1 (Instanciation partielle cohérente) *Soit x une instanciation partielle des variables $x_1, \ldots, x_{|\mathcal{I}|}$ et M un ensemble de mesures préservantes. x est une instanciation partielle cohérente ssi x^+ est fréquent par rapport à θ et $\nexists j \in x^-$ tel que* $\texttt{cl}_{inc}(x^+, j, M)$ *est vérifiée.*

Proposition 3 (Règles de filtrage de ADEQUATECLOSURE) *Étant donné une instanciation partielle cohérente x, un ensemble de mesures préservantes M, pour tout $i \in x^*$, les règles $(1-3)$ suppriment les valeurs inconsistantes de $dom(x_i)$: (1) si* $\texttt{cl}_{inc}(x^+, i, M) \Rightarrow 0 \notin dom(x_i)$*; (2) si $|\boldsymbol{t}_{\mathcal{D}}(x^+ \cup \{i\})| < \theta \Rightarrow 1 \notin dom(x_i)$; (3) si $\exists j \in x^-$ s.t.* $\texttt{cl}_{inc}(x^+ \cup \{i\}, j, M) \Rightarrow 1 \notin dom(x_i)$.

Notre contrainte globale ADEQUATECLOSURE propage à partir des variables booléennes aux variables entières qui représentent la fréquence d'un motif dans les jeux de données $\mathcal{D}$ et $\mathcal{D}_1$. Ainsi, deux autres règles similaires à (Schaus et al., 2017) sont appliquées pour mettre à jour les bornes de f et f_1 :

1. Un motif non vide P est une extension de fermeture Wang et al. (2003) de Q ssi $\boldsymbol{t}(P \cup Q) = \boldsymbol{t}(Q)$.

Algorithme 1 : Filtrage pour ADEQUATECLOSURE

Input : $\mathcal{D}$: base transactionnelle ; θ : support minimal ; M : ensemble de mesures ;
InOut : $x = \{x_1 ... x_n\}$: Variables d'items booléennes ; f, f_1 : Variables entières ;

begin
  if $|\boldsymbol{t}_{\mathcal{D}}(x^+)| < \theta$ **then return faux** ;
  if $\exists\, i \in x^-$ *s.t.* $\texttt{closureInclusion}(x^+, i, M)$ **then return faux** ;
  foreach $i \in x^*$ **do**
    if $|\boldsymbol{t}_{\mathcal{D}}(x^+ \cup \{i\})| < \theta$ **then**
      $dom(x_i) \leftarrow dom(x_i) - \{1\}$; $x^- \leftarrow x^- - \cup\{i\}$; $x^* \leftarrow x^* \setminus \{i\}$;
    else if $\texttt{closureInclusion}(x^+, i, M)$ **then**
      $dom(x_i) \leftarrow dom(x_i) - \{0\}$; $x^+ \leftarrow x^+ \cup \{i\}$; $x^* \leftarrow x^* \setminus \{i\}$;
  end
  foreach $j \in x^-$ **do**
    foreach $i \in x^*$ **do**
      if $closureInclusion(x^+ \cup \{i\}, j, M)$ **then**
        $dom(x_i) \leftarrow dom(x_i) - \{1\}$; $x^- \leftarrow x^- \cup \{i\}$; $x^* \leftarrow x^* \setminus \{i\}$;
      end
    end
  end
  $\texttt{updateBounds}(f, |\boldsymbol{t}_{\mathcal{D}}(x^+ \cup x^*)|, |\boldsymbol{t}_{\mathcal{D}}(x^+)|)$;
  $\texttt{updateBounds}(f_1, |\boldsymbol{t}_{\mathcal{D}_1}(x^+ \cup x^*)|, |\boldsymbol{t}_{\mathcal{D}_1}(x^+)|)$;
  return vrai ;
end
Function $\texttt{closureInclusion}(x,i,M)$ **: Boolean**
  foreach $m \in M$ **do**
    if $m(x \cup \{i\}) \neq m(x)$ **then return faux**;
  end
  return vrai ;

(4) règles UB : $\begin{cases} if\ |\boldsymbol{t}_{\mathcal{D}}(x^+)| < UB(f) \Rightarrow UB(f) \leq |\boldsymbol{t}_{\mathcal{D}}(x^+)| \\ if\ |\boldsymbol{t}_{\mathcal{D}_1}(x^+)| < UB(f_1) \Rightarrow UB(f_1) \leq |\boldsymbol{t}_{\mathcal{D}_1}(x^+)| \end{cases}$

(5) règles LB : $\begin{cases} if\ |\boldsymbol{t}_{\mathcal{D}}(x^+ \cup x^*)| > LB(f) \Rightarrow LB(f) \geq |\boldsymbol{t}_{\mathcal{D}}(x^+ \cup x^*)| \\ if\ |\boldsymbol{t}_{\mathcal{D}_1}(x^+ \cup x^*)| > LB(f_1) \Rightarrow LB(f_1) \geq |\boldsymbol{t}_{\mathcal{D}_1}(x^+ \cup x^*)| \end{cases}$

Pour connecter ensemble les variables f et f_1, nous définissons une nouvelle contrainte indépendante de la contrainte ADEQUATECLOSURE, qui s'exprime par $f_2 = f - f_1$, où f_2 est une variable entière qui représente la taille de la couverture des motifs qui sont contenus dans le jeu de données $\mathcal{D} \setminus \mathcal{D}_1$. La contrainte sur la variable f_2 est activée seulement si le taux de croissance apparaît dans M. Dans ce cas, les bornes de la variable f_1 sont mis à jour.

Exemple 3 *Soit le jeu de données de la table 1a avec* $M = \{sup, min\}$, $\theta = 2$, *et* $dom(f) = \{0, \ldots, 6\}$. *La contrainte* $f \geq \theta$ *met à jour le minorant de* f *à* 2, *i.e.* $LB(f) = 2$. *Soit l'instanciation partielle* $x^+ = \emptyset$, $x^- = \{E\}$ *and* $x^* = \{A, B, C, D, F\}$. *Graçe à la règle* (3), *la valeur* 1 *est filtrée de* $dom(x_A)$ *car* $E \in \texttt{clos}_M(x^+ \cup \{A\})$ *et* $E \in x^-$. *Soit l'instanciation partielle* $x^+ = \{A, B, C, D\}$, $x^- = \emptyset$, $x^* = \{E, F\}$, $LB(f) = 2$ *et* $UB(f) = 6$, *la règle* (1) *filtre la valeur* 0 *de* $dom(x_E)$ *car* $\texttt{cl}_{inc}(ABCD, E, M)$ *est vrai, i.e.* $E \in \texttt{clos}_M(ABCD)$. *La règle* (2) *filtre la valeur* 1 *de* $dom(x_F)$ *car* $\boldsymbol{t}(ABCDF) = 1 < \theta$. *Finalement, la règle* (4) *met à jour le majorant de* f *à* $|\boldsymbol{t}(ABCDE)|$, *i.e.* $UB(f) = 2$.

L'algorithme 1 montre la propagation de ADEQUATECLOSURE. Il prend en entrée le jeu de données transactionnel $\mathcal{D}$, les variables d'items x, les deux variables entières f et f_1, le support minimum θ et l'ensemble de mesures M. Il commence par calculer la couverture du motif x^+ et vérifie si l'instanciation partielle actuelle est inconsistante (Lemme 1), c'est à

dire, si x^+ est soit infréquent (ligne 2) ou x^+ ne peut pas être étendu à un motif clos par rapport à M sans ajouter i ($i \in x^-$) (ligne 3), dans ce cas la contrainte n'est pas respectée et on retourne un échec. L'algorithme 1 supprime les items $i \in x^*$ qui ne peuvent pas appartenir à une solution qui contient x^+. Pour cela, nous testons en premier si $x^+ \cup \{i\}$ est infréquent par rapport à θ (ligne 5). Si c'est le cas, nous supprimons 1 de $dom(x_i)$ et nous mettons à jour x^- et x^* (ligne 6). Ensuite, pour chaque mesure $m \in M$, la fonction `closureInclusion`(x^+, i, M) vérifie si ajouter l'item i ne modifie pas la valeur de m pour la spécialisation x^+ (lignes 22-23). Si c'est le cas, la fonction retourne **vrai** (ligne 25), supprime 0 de $dom(x_i)$ et met à jour les ensembles x^+ et x^* (ligne 8). Troisièmement, en appliquant la règle (3), nous supprimons 1 du domaine de chaque variable d'item $i \in x^*$ tel que $x^+ \cup \{i\}$ ne peut pas être étendu à un motif clos par rapport à M sans ajouter un item absent $j \in x^-$ (lignes 10-16). Enfin, nous mettons à jour les bornes des variables f et f_1 en appliquant les règles (4) et (5).

Proposition 4 (Consistance et complexité en temps) *Étant donné une base de données transactionnelle $\mathcal{D}$ qui contient n items et m transactions, un support minimal θ et un ensemble de mesures M qui contient c mesures basées sur sup. L'algorithme 1 assure la consistance de domaine en temps $\mathcal{O}(n^2 \times m \times c)$.*

4 Fouille de skypatterns avec ADEQUATECLOSURE

Cette section décrit comment résoudre le problème de fouille de skypatterns à l'aide de la contrainte ADEQUATECLOSURE. L'idée principale est d'utiliser une archive $\mathcal{A}$ de skypatterns pour supprimer les solutions qui sont dominées par au moins une solution de $\mathcal{A}$. Soit $M = \{m_1, \ldots, m_k\}$ un ensemble de mesures à maximiser, et x les variables qui modélisent le motif inconnu. Nous définissons k variables objectifs entières $obj_1, \ldots, obj_k$, et nous contraignons chaque variable obj_i à être égale à la valeur de la i^{eme} mesure par rapport à une instanciation partielle donnée x, i.e. $obj_i = m_i(x)$. Notre modèle PPC est initialisé par la contrainte ADEQUATECLOSURE$_{\mathcal{D},M',\theta}(x, f, f_1)$. Ensuite, pour chaque nouvelle solution s_i, nous ajoutons une contrainte dynamique $\phi(s_i, x) \equiv (s_i \not\succ_{\mathcal{P}} x) \Leftrightarrow (\wedge_{j=1..k} m(s_i) = obj_j) \vee (\vee_{j=1..k} m(s_i) < obj_j)$.

Nous commençons par calculer l'ensemble de mesures M' à partir de M avec l'opérateur $\bar{c}$. Ensuite, nous imposons que x doit être un motif clos selon M' grâce à la contrainte ADEQUATECLOSURE. Deuxièmement, chaque fois qu'une nouvelle solution s_i est découverte, celle-ci est insérée dans l'archive, une nouvelle contrainte dynamique est ajoutée et la recherche se poursuit. Pour maintenir la propriété de non-dominance dans l'ensemble de skypatterns, à chaque ajout de solution s_i, nous supprimons toutes les solutions de $\mathcal{A}$ qui sont dominées par rapport à M par s_i. Ce processus s'arrête lorsque l'ensemble des contraintes du système n'a pas de solution. Il faut noter que contrairement à (Ugarte et al., 2017), il n'est pas nécessaire d'avoir une deuxième étape de traitement des motifs car tous les motifs candidats s_i qui ne sont pas des skypatterns sont supprimés de $\mathcal{A}$ pendant la recherche.

Heuristique de branchement. Comme *heuristique pour ordonner les variables*, nous choisissons l'item libre i (*i.e.* $i \in x^*$) tel que $|\boldsymbol{t}_{\mathcal{D}}(x^+ \cup \{i\})|$ est minimal. Cette heuristique nous permet d'activer le plus tôt possible nos règles de filtrage (voir l'algorithme 1), donc de réduire l'espace de recherche.

5 Expérimentations

Nous avons mené des expérimentations pour répondre aux questions suivantes : (1) Quelles sont les performances (en temps CPU) de notre contrainte globale (noté ADEQUATE-CI) comparé à CP+CLOSED et MICMAC pour la fouille de motifs clos ? CP+CLOSED utilise la première étape de CP+SKY (modèle réifié). (2) Quelles sont les performances (en temps CPU) de notre approche (noté CLOSEDSKY) comparé à CP+SKY et AETHERIS pour la fouille de skypatterns ? (3) Quel est le nombre de skypatterns comparé au nombre de motifs clos ?

Protocole expérimental. Nous avons utilisé les jeux de données UCI (fimi.ua.ac.be/data) et avons choisi des jeux de données de différentes tailles et densités. Certains jeux de données, comme HEPATITIS et CHESS sont très denses (resp. 50% et 49%). D'autres au contraire sont très peu denses, comme T10I4D100K et RETAIL (resp. 1% and 0.06%). L'implémentation a été réalisée avec `Choco` (Prud'homme et al., 2016) version $4.10.5$, une librairie Java pour la programmation par contraintes. Nous avons utilisé les versions d'AETHERIS et MICMAC fournies par les auteurs (implantées en C++). Les expérimentations ont été menées sous un AMD Opteron 6174, 2.2 GHz avec une RAM de 256 Go et une limite de temps de 24 heures. La maximum heap size autorisée par la JVM est 30 Go. Nous considérons les ensembles de mesures suivants pour la fouille de motifs clos : AC_1 : $\{min(X.val), sup(X) \geq \theta\}$, AC_2 : $\{max(X.val), sup(X) \geq \theta\}$ et AC_3 : $\{min(X.val), max(X.val), sup(X) \geq \theta\}$. Les mesures qui utilisent des valeurs numériques, comme min ou max, sont appliquées à des valeurs générées aléatoirement dans l'intervalle $[0, 1]$.

(a) Comparaison entre ADEQUATE-CI, CP+CLOSED et MICMAC. La table 2 compare les performances de ADEQUATE-CI, CP+CLOSED et MICMAC pour différentes valeurs de θ pour différents jeux de données et ensembles de mesures. Pour chaque méthode, nous reportons le temps CPU (en secondes), le nombre de motifs clos et le nombre de noeuds explorés. Concernant les temps d'exécution, ADEQUATE-CI arrive à terminer l'exécution sur toutes les instances contrairement aux autres méthodes qui obtiennent soit *Out Of Memory* ou *Time Out*. Sur 31 instances, MICMAC obtient 10 `OOM` et 1 `TO`, CP+CLOSED 15 `TO` et 2 `OOM`. Comparé à MICMAC, ADEQUATE-CI a le meilleur temps d'éxecution sur 17 instances, avec un facteur d'accélération compris entre 3 et 5. Les seules exceptions sont HEART-CLEVELAND, HEPATITIS et MUSHROOM, où MICMAC est plus efficace. Par ailleurs, ADEQUATE-CI domine très largement CP+CLOSED. En effet, le nombre élevé de transactions et d'items augmente sensiblement le temps de propagation pour le modèle réifié. La nature level-wise de MICMAC explique en partie les *Out Of Memory*, à cause du grand nombre de candidats qui doit être stocké pendant le processus de fouille. Ces résultats démontrent l'intérêt de notre approche pour la fouille de motifs clos. Nous rappelons que notre approche est générique et plus flexible : l'utilisateur peut facilement ajouter de nouvelles contraintes sans avoir à modifier le système sous-jacent.

(b) Comparaison entre CLOSEDSKY, CP+SKY et AETHERIS. Nous avons aussi comparé CLOSEDSKY, CP+SKY et AETHERIS pour la fouille de skypaterns avec différentes combinaisons de mesures parmi l'ensemble $\{sup : f, \max : M, \min : m, area : a, mean : n, growth\text{-}rate : g\}$. Pour chaque méthode et chaque combinaison sélectionnée, nous reportons le temps CPU, le nombre de skypatterns et le nombre de noeuds explorés par les deux méthodes PPC. Rappelons que AETHERIS et CP+SKY calculent en premier un ensemble représentatif de motifs par rapport à M' et appliquent ensuite l'opérateur Sky sur l'ensemble des motifs

Dataset $\lvert\mathcal{I}\rvert \times \lvert\mathcal{T}\rvert$	θ	#Motifs	#Nœud		Temps (s)		
		(1)(2)(3)	(1)	(2)	(1)	(2)	(3)
CHESS 75 × 3,196	0.3	730,6791	**14,613,581**	14,890,173	**1545**	50751	6283
	0.2	30,120,283	**60,240,565**	-	**6536**	TO	26792
	0.1	153,073,913	**306,147,825**	-	**36606**	TO	TO
CONNECT 129 × 67,557	0.18	323,3691	**6,467,381**	-	**6884**	TO	OOM
	0.15	5,084,539	**10,169,077**	-	**11159**	TO	OOM
	0.1	11,903,644	**23,807,287**	-	**26979**	TO	OOM
HEART-CLEVELAND 95 × 296	0.1	14,126,585	**28,253,169**	31,283,345	3297	10264	**1352**
	0.08	26,812,645	**53,625,289**	58,338,937	6251	19423	**3096**
	0.06	53,854,923	**107,709,845**	114,916,691	12738	39720	**7473**
HEPATITIS 68 × 137	0.2	38,6831	**773,661**	847,343	60	159	**24**
	0.1	1,949,759	**3,899,517**	4,115,027	333	799	**127**
	0.05	4,196,027	**8,392,053**	8,713,651	723	1675	**315**
KR-VS-KP 73 × 3,196	0.3	4,501,990	**9,003,979**	9,191,063	**997**	32293	5391
	0.2	17,825,411	**35,650,821**	-	**3848**	TO	22417
MUSHROOM 112 × 8,124	0.01	39,676	**79,351**	130,235	25	2227	**11**
	0.008	48,601	**97,201**	156,131	29	2595	**12**
	0.005	63,914	**127,827**	204,137	42	3263	**13**
PUMSB 2,113 × 49,046	0.8	46,495	**92,989**	-	**440**	TO	660
	0.7	358,767	**717,533**	-	**3485**	TO	12015
SPLICE1 287 × 3,190	0.1	1,580	**3,159**	60,871	5	3043	2
	0.05	30,473	**60,945**	2,541,737	155	51671	**41**
	0.02	565,780	**1,131,559**	-	1283	TO	**308**
T10I4D100K 870 × 100,000	0.005	1,074	**2,147**	-	**657**	TO	OOM
	0.0025	7654	**15,307**	-	**1490**	TO	OOM
T40I10D100K 942 × 100,000	0.08	138	**275**	-	**13**	TO	OOM
	0.05	317	**633**	-	**101**	TO	OOM
	0.01	65,237	**130,473**	-	**8801**	TO	OOM
BMS1 497 × 59,602	0.001	3,977	**7,953**	141,199	167	56872	**131**
	0.0005	178,468	**356,935**	-	2067	TO	**559**
RETAIL 16470 × 88,162	0.004	832	**1,663**	-	**185**	OOM	OOM
	0.002	2,692	**5,383**	-	**3428**	OOM	OOM

(a) $AC_1 : \{min(X.val), sup(X) \geq \theta\}$

Dataset $\lvert\mathcal{I}\rvert \times \lvert\mathcal{T}\rvert$	θ	#Motifs	#Nœud		Temps (s)		
		(1)(2)(3)	(1)	(2)	(1)	(2)	(3)
CHESS 75 × 3,196	0.3	6,788,640	**13,577,279**	13,853,453	**2758**	52298	5782
	0.2	30,521,170	**61,042,339**	-	**12196**	TO	27085
	0.1	175,901,422	**351,802,843**	-	**78599**	TO	TO
CONNECT 129 × 67,557	0.18	6,581,293	**13,162,585**	-	**16886**	TO	OOM
	0.15	10,320,509	**20,641,017**	-	**26497**	TO	OOM
	0.1	24,476,915	**48,953,829**	-	**66785**	TO	OOM
HEART-CLEVELAND 95 × 296	0.1	22,598,666	**45,197,331**	48,301,365	10489	22502	**3198**
	0.08	44,286,784	**88,573,567**	93,459,361	20900	43804	**4981**
	0.06	92,532,208	**185,064,415**	-	45554	TO	**12909**
HEPATITIS 68 × 137	0.2	534,121	**1,068,241**	1,140,525	181	330	**31**
	0.1	3,149,673	**6,299,345**	6,508,211	1071	1839	**195**
	0.05	7,762,648	**15,525,295**	15,847,857	2619	4427	**576**
KR-VS-KP 73 × 3,196	0.3	4,369,825	**8,739,649**	8,925,567	**1731**	32289	4400
	0.2	17,997,787	**35,995,573**	-	**7237**	TO	19426
MUSHROOM 112 × 8,124	0.01	138,795	**277,589**	329,347	119	5771	**25**
	0.008	174,672	**349,343**	409,301	156	7317	**29**
	0.005	232,315	**464,629**	542,607	186	9424	**34**
PUMSB 2,113 × 49,046	0.8	46755	93509	-	964	TO	655
	0.7	337740	675479	-	**7252**	TO	10660
SPLICE1 287 × 3,190	0.1	1,580	**3,159**	60,871	8	2968	2
	0.05	30,473	**60,945**	2,541,737	170	52622	**38**
	0.02	565,846	**1,131,691**	-	1699	TO	**167**
T10I4D100K 870 × 100,000	0.005	1,074	**2,147**	-	**669**	TO	OOM
	0.0025	7,698	**15,395**	-	**1610**	TO	OOM
T40I10D100K 942 × 100,000	0.08	138	**275**	-	**13**	TO	OOM
	0.05	317	**633**	-	**85**	TO	OOM
	0.01	65,237	**130,473**	-	**7094**	TO	OOM
BMS1 497 × 59,602	0.001	3,982	**7963**	141,207	187	60778	**134**
	0.0005	294,585	**589169**	-	3685	TO	**617**
RETAIL 16470 × 88,162	0.004	832	**1663**	-	**306**	OOM	OOM
	0.002	2,692	**5383**	-	**3626**	OOM	OOM

(b) $AC_3 : \{min(X.val), max(X.val), sup(X) \geq \theta\}$

Dataset $\lvert\mathcal{I}\rvert \times \lvert\mathcal{T}\rvert$	θ	#Motifs	#Nœud		Temps (s)		
		(1)(2)(3)	(1)	(2)	(1)	(2)	(3)
CHESS 75 × 3,196	0.3	6,199,288	**12,398,575**	12,674,595	**1416**	46366	5815
	0.2	27,341,083	**54,682,165**	-	**6216**	TO	25545
	0.1	150,302,539	**300,605,077**	-	**37159**	TO	TO
CONNECT 129 × 67,557	0.18	36,897,79	**7,379,557**	-	**8026**	TO	OOM
	0.15	5,741,671	**1,1483,341**	-	**13258**	TO	OOM
	0.1	13,360,851	**267,21,701**	-	**31827**	TO	OOM
HEART-CLEVELAND 95 × 296	0.1	13,652,085	**27,304,169**	30,360,067	3377	9717	**1264**
	0.08	25,776,190	**51,552,379**	56,330,623	6171	18830	**3055**
	0.06	51,168,896	**102,337,791**	109,707,247	12590	38049	**6692**
HEPATITIS 68 × 137	0.2	429,368	**858,735**	932,627	70	166	**25**
	0.1	2,263,487	**4,526,973**	4,739,553	391	887	**143**
	0.05	5,031,535	**10,063,069**	10,378,429	874	1979	**377**
KR-VS-KP 73 × 3,196	0.3	4,345,059	**8,,690,117**	8,875,613	**925**	30780	5232
	0.2	17,881,775	**35,763,549**	-	**3813**	TO	20553
MUSHROOM 112 × 8,124	0.01	45,766	**91,531**	141,203	28	2407	**18**
	0.008	55,721	**111,441**	168,807	37	2795	**20**
	0.005	71,996	**143,991**	218,075	47	3525	**25**
PUMSB 2,113 × 49,046	0.8	36,450	**72,899**	-	**357**	TO	601
	0.7	254,892	**509,783**	-	**2657**	TO	10862
SPLICE1 287 × 3,190	0.1	1,580	**3,159**	60,871	5	2986	2
	0.05	30,473	**60,945**	2,541,737	146	53964	**41**
	0.02	56,5792	**1,131,583**	-	1286	TO	**202**
T10I4D100K 870 × 100,000	0.005	1074	**2,147**	-	**669**	TO	OOM
	0.0025	7697	**153,93**	-	**1503**	TO	OOM
T40I10D100K 942 × 100,000	0.08	138	**275**	-	**11**	TO	OOM
	0.05	317	**633**	-	**81**	TO	OOM
	0.01	65,237	**130,473**	-	**7431**	TO	OOM
BMS1 497 × 59,602	0.001	3,979	**7,957**	141,201	152	54370	**134**
	0.0005	192,125	**384,249**	-	2164	TO	**578**
RETAIL 16470 × 88,162	0.004	832	**1,663**	-	**196**	OOM	OOM
	0.002	2,692	**5,383**	-	**3852**	OOM	OOM

(c) $AC_2 : \{max(X.val), sup(X) \geq \theta\}$

TAB. 2: Analyse comparative pour l'extraction de motifs clos. " − " : résultats non disponibles. TO : Time Out ; OOM : Out Of Memory. (1) : ADEQUATE-CI (2) : CP+CLOSED (3) : MICMAC.

extraits, alors que notre méthode basée sur ADEQUATE-CI ne nécessite qu'une seule étape. Nous avons utilisé un *seuil de fréquence* de 1. La table 3 montre les résultats obtenus.

Dataset $\lvert\mathcal{I}\rvert \times \lvert\mathcal{T}\rvert$	M	|Sky(M)|	# Nœud		Temps (s)			Dataset $\lvert\mathcal{I}\rvert \times \lvert\mathcal{T}\rvert$	M	|Sky(M)|	# Nœud		Temps (s)		
		(1)(2)(3)	(1)	(2)	(1)	(2)	(3)			(1)(2)(3)	(1)	(2)	(1)	(2)	(3)
CHESS 75 × 3,196	ag	18	428,795	428,831	**142**	966	OOM	HEPATITIS 68 × 137	ag	22	98,879	94,071	**20**	23	432
	fa	13	4,085	4,189	**4**	41	OOM		fa	17	31,175	31,599	**4**	3	429
	fg	19	343,221	338,921	**78**	687	OOM		fg	33	78,027	66,875	**14**	16	435
	fag	160	379,675	378,391	**146**	957	OOM		fag	89	95,375	91,073	**30**	35	439
	agmM	577	1,582,829	1,541,931	**3559**	6183	OOM		agmM	280	125,091	133,999	**230**	264	814
	agnM	674	1,671,847	1,630,627	**3838**	6488	OOM		agnM	277	124,399	133,971	**203**	233	860
	agnm	1,459	1,700,829	1,648,195	**5778**	8630	OOM		agnm	1,070	176,333	193,791	**1123**	1184	1351
	fagM	293	576,125	572,445	**241**	1152	OOM		fagM	200	105,053	103,539	**59**	67	460
	fagn	846	1,079,609	1,044,509	**944**	2721	OOM		fagn	424	134,453	147,083	**291**	345	844
	famM	245	24,213	24,815	**19**	155	OOM		famM	73	39,253	39,865	**11**	**11**	804
	fanM	281	27,773	28,717	**24**	151	OOM		fanM	78	40,687	41,339	**10**	**11**	861
	fanm	219	12,169	12,389	**13**	97	OOM		fanm	163	58,433	59,313	**26**	27	1342
	fgmM	249	1,406,647	1,346,019	**377**	2277	OOM		fgmM	127	82,947	76,411	**44**	51	807
	fgnM	356	1,483,179	1,418,595	**434**	2312	OOM		fgnM	130	80,619	76,343	**45**	51	855
	fgnm	446	1,560,195	1493,333	**577**	2516	OOM		fgnm	300	89,991	89,295	**110**	122	1340
	fagmM	2,464	1,606,675	1,564,011	**4862**	7595	OOM		fagmM	454	135,717	147,497	**375**	418	817
	fagnM	3,014	1,689,931	1,645,885	**5287**	8297	OOM		fagnM	467	136,069	149,035	**341**	393	875
	fagnm	3,630	1,715,569	1,663,603	**7286**	10150	OOM		fagnm	1,366	187,435	207,941	1564	1772	**1362**
	fagnMm	7,845	3,591,471	3,551,559	**59720**	66388	OOM		fagnMm	4,242	458,733	573,029	17951	18922	2092
CONNECT 129 × 67,557	ag	45	4,775,201	-	**25641**	TO	OOM	KR-VS-KP 73 × 3,196	ag	18	464,439	495,993	**267**	810	OOM
	fa	17	3,715	3,745	**47**	1722	OOM		fa	13	3,095	3,189	**3**	32	OOM
	fg	26	3,623,055	-	**20483**	TO	OOM		fg	42	30,3767	321,847	**104**	413	OOM
	fag	359	4,186,703	-	**25458**	TO	OOM		fag	145	419,133	448,447	**460**	1079	OOM
	agmM	903	6,618,189	-	**69963**	TO	OOM		agmM	426	550,395	569,155	**1752**	2681	OOM
	agnM	857	6,674,167	-	**68990**	TO	OOM		agnM	491	611,563	631,283	**2756**	3849	OOM
	agnm	-	-	-	TO	TO	OOM		agnm	1,885	910,823	954,803	**10301**	12115	OOM
	fagM	941	6,234,675	-	**43943**	TO	OOM		fagM	435	513,537	547,759	**1477**	2438	OOM
	fagn	-	-	-	TO	TO	OOM		fagn	1,423	586,571	617,695	**3086**	4376	OOM
	famM	142	6,403	6,519	**68**	2186	OOM		famM	125	7,187	7,271	**7**	59	OOM
	fanM	155	6,565	6,681	**61**	2240	OOM		fanM	179	7,795	7,889	**7**	61	OOM
	fanm	367	25,161	25,407	**205**	4162	OOM		fanm	221	7,957	8,049	**10**	67	OOM
	fgmM	291	5,245,879	-	**37120**	TO	OOM		fgmM	241	352,221	357,985	**230**	692	OOM
	fgnM	436	5,319,999	-	**36955**	TO	OOM		fgnM	429	396,307	402,753	**384**	934	OOM
	fgnm	905	10,168,103	-	**83788**	TO	OOM		fgnm	589	627,629	648,899	**918**	1709	OOM
	fagmM	-	-	-	TO	TO	OOM		fagmM	1,639	602,813	633,661	**3368**	4575	OOM
	fagnM	-	-	-	TO	TO	OOM		fagnM	2,285	674,531	706,523	**5576**	7171	OOM
	fagnm	-	-	-	TO	TO	OOM		fagnm	5,854	957,721	1,007,825	**16484**	19022	OOM
	fagnMm	-	-	-	TO	TO	OOM		fagnMm	9,683	1,115,879	1,165,039	**68169**	75077	OOM
HEART-CLEVELAND 95 × 296	ag	28	1,187,475	1,094,719	**326**	463	OOM	MUSHROOM 112 × 8,124	ag	15	15,881	30,299	**10**	355	15
	fa	14	82,407	83,597	**14**	**14**	OOM		fa	5	711	1,169	**3**	97	15
	fg	33	1,108,677	957,337	**237**	374	OOM		fg	17	15,405	29,585	**9**	328	15
	fag	115	1,179,427	1,086,389	**468**	632	OOM		fag	59	15,761	30,175	**10**	356	15
	agmM	479	1,476,827	1,530,369	**7327**	8576	OOM		agmM	245	19,363	35,711	**29**	508	**28**
	agnM	421	1,487,729	1,508,703	**7610**	9204	OOM		agnM	240	19,339	35,803	**30**	513	**30**
	agnm	2,893	1,505,917	1,688,261	**43747**	48893	OOM		agnm	513	23,131	40,195	**72**	663	109
	fagM	181	1,246,275	1,148,699	**1008**	1229	OOM		fagM	110	16,031	30,865	**10**	393	16
	fagn	752	149,0497	1,509,219	**9254**	10331	OOM		fagn	268	18,049	33,565	**20**	444	40
	famM	78	105,135	106,613	**39**	45	OOM		famM	97	1,653	2,761	**5**	152	28
	fanM	130	142,153	143,821	**56**	61	OOM		fanM	137	1,781	3,141	**4**	151	40
	fanm	233	210,245	212,289	**130**	141	OOM		fanm	92	2,157	3,829	**6**	191	110
	fgmM	136	1,249,387	1,100,997	**844**	985	OOM		fgmM	143	18,385	34,163	**19**	452	28
	fgnM	145	1,243,635	1,081,463	**976**	1116	OOM		fgnM	182	18,387	34,223	**20**	447	**30**
	fgnm	417	867,935	835,749	**1468**	1687	OOM		fgnm	234	19,781	35,815	**30**	445	108
	fagmM	830	1,502,943	1,559,133	**10033**	11464	OOM		fagmM	506	19,471	35,851	**34**	534	**29**
	fagnM	795	1,512,021	1,536,235	**10401**	11872	OOM		fagnM	551	19,503	36,001	**33**	548	**30**
	fagnm	4,204	1,708,165	1,966,619	**64454**	70359	OOM		fagnm	974	23,365	40,433	**88**	682	109
	fagnMm	-	-	-	TO	TO	OOM		fagnMm	1,021	25,501	44,911	107	780	**59**

TAB. 3: Analyse comparative pour l'extraction de skypatterns. " − " : resultats non disponibles (TO or OOM). TO : Time Out ; OOM : Out Of Memory. (1) : CLOSEDSKY (2) : CP+SKY (3) : AETHERIS.

Premièrement, les résultats montrent qu'il y a une grande différence entre le nombre de motifs clos (en millions, voir table 2) en comparaison avec le nombre de motifs Sky (en milliers). Cela démontre l'intérêt de la *dominance Pareto* pour réduire le nombre de motifs. Deuxièmement, en observant le temps d'exécution, on constate que CLOSEDSKY surpasse CP+SKY et AETHERIS sur toutes les instances considérées. CLOSEDSKY permet d'extraire les skypatterns là où les deux autres approches échouent. En effet, AETHERIS obtient des *Out Of Memory* sur 76 instances (sur un total de 114), alors que CP+SKY ne parvient pas à finir l'extraction sur 17 instances. Par comparaison, CLOSEDSKY échoue sur 6 instances (5 instances pour CONNECT

et 1 instance pour HEART-CLEVLAND). Pour les jeux de données où CP+SKY et AETHERIS arrivent à terminer l'extraction, CLOSEDSKY obtient le meilleur temps d'exécution, sauf pour 2 instances où AETHERIS est plus efficace. Pour CHESS, CLOSEDSKY est 8 fois plus rapide que CP+SKY ; pour MUSHROOM, le facteur d'accélération est en moyenne de 23.86. Pour HEPATITIS, CLOSEDSKY est plus rapide que AETHERIS (en moyenne 25.56 fois plus rapide).

(c) Impact de la règle (3) sur les performances de ADEQUATE-CI. ADEQUATE-CI assure la cohérence de domaine avec une complexité cubique mais avec un temps d'exécution plus long. Nous avons implémenté une nouvelle version qui ne prend en compte que les règles (1) et (2) (complexité quadratique). Nous avons testé cette nouvelle version (notée CLOSEDSKY-WC) sur CONNECT et SPLICE1. Les résultats sont disponibles dans (Vernerey et al., 2021). WC domine clairement DC en temps de calcul. Pour CONNECT, WC arrivent à trouver les motifs Sky pour 3 instances où DC n'arrivent pas à terminer l'extraction, WC étant en moyenne 9.5 fois plus rapide que DC. Pour SPLICE1, WC réussit à terminer l'extraction sur 7 instances (sur un total de 19). Comme seconde observation, le nombre de noeuds exploré par DC est à chaque fois plus petit que celui de WC mais la différence n'est pas significative contrairement au gain en temps d'éxecution que procure WC. Par conséquent, un filtrage plus faible constitue un bon compromis pour les instances qui sont très difficiles à résoudre.

6 Conclusions

Nous avons proposé une nouvelle contrainte globale pour la fouille de motifs clos par rapport à un ensemble de mesures. Nous avons montré l'utilisation de notre contrainte pour l'extraction de skypatterns. Nous avons mené des expérimentations sur plusieurs jeux de données de l'UCI qui ont démontré l'efficacité et le passage à l'échelle de notre approche pour les deux tâches de fouille comparé au modèle PPC réifié et aux méthodes spécialisées.

Références

Börzsönyi, S., D. Kossmann, et K. Stocker (2001). The skyline operator. In *ICDE*, pp. 421–430.

Calders, T., C. Rigotti, et J. Boulicaut (2004). A survey on condensed representations for frequent sets. In *Constraint-Based Mining and Inductive Databases*, pp. 64–80. Springer.

Giacometti, A., D. Laurent, et C. T. Diop (2002). Condensed representations for sets of mining queries. In *Proceedings of the 1st Int. Workshop on Inductive Databases*, pp. 5–19.

Guns, T., S. Nijssen, et L. De Raedt (2011). Itemset mining : A constraint programming perspective. *Artificial Intelligence 175*(12), 1951–1983.

Hien, A., S. Loudni, N. Aribi, Y. Lebbah, M. Laghzaoui, A. Ouali, et A. Zimmermann (2020). A relaxation-based approach for mining diverse closed patterns. In *Proceedings of PKDD*, Volume 12457 of *Lecture Notes in Computer Science*, pp. 36–54.

Ke, Y., J. Cheng, et J. X. Yu (2009). Top-k correlative graph mining. In *SDM*, pp. 1038–1049. SIAM.

Lazaar, N., Y. Lebbah, S. Loudni, M. Maamar, V. Lemière, C. Bessiere, et P. Boizumault (2016). A global constraint for closed frequent pattern mining. In *Proceedings of the 22nd CP*, pp. 333–349.

Novak, P. K., N. Lavrac, et G. I. Webb (2009). Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research 10*, 377–403.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th ICDT*, pp. 398–416.

Prud'homme, C., J.-G. Fages, et X. Lorca (2016). Choco Solver Documentation.

Raedt, L. D., T. Guns, et S. Nijssen (2008). Constraint programming for itemset mining. In *SIGKDD*, pp. 204–212. ACM.

Schaus, P., J. O. R. Aoga, et T. Guns (2017). Coversize : A global constraint for frequency-based itemset mining. In *Proceedings of the 23rd CP 2017*, pp. 529–546.

Soulet, A. et B. Crémilleux (2008). Adequate condensed representations of patterns. *Data Min. Knowl. Discov. 17*(1), 94–110.

Soulet, A., B. Crémilleux, et F. Rioult (2004). Condensed representation of emerging patterns. In *Proceedings of the 8th PAKDD*, pp. 127–132. Springer.

Soulet, A., C. Raïssi, M. Plantevit, et B. Crémilleux (2011). Mining dominant patterns in the sky. In *Proceedings of the ICDM 2011*, pp. 655–664. IEEE Computer Society.

Ugarte, W., P. Boizumault, B. Crémilleux, A. Lepailleur, S. Loudni, M. Plantevit, C. Raïssi, et A. Soulet (2017). Skypattern mining : From pattern condensed representations to dynamic constraint satisfaction problems. *Artif. Intell. 244*, 48–69.

Ugarte, W., P. Boizumault, S. Loudni, B. Crémilleux, et A. Lepailleur (2014). Mining (soft-) skypatterns using dynamic CSP. In *Proceedings of CPAIOR 2014*, pp. 71–87.

Vernerey, C., S. Loudni, N. Aribi, et Y. Lebbah (April 2021). Supplementary Material : https://drive.google.com/file/d/1LwzEojaTCzMuVs4HFwGn7-JPlHjCWvmC/view?usp=sharing.

Wang, J., J. Han, Y. Lu, et P. Tzvetkov (2005). TFP : an efficient algorithm for mining top-k frequent closed itemsets. *IEEE Trans. Knowl. Data Eng. 17*(5), 652–664.

Wang, J., J. Han, et J. Pei (2003). CLOSET+ : searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the Ninth KDD*, pp. 236–245. ACM.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *PKDD*, Volume 1263 of *LNCS*, pp. 78–87. Springer.

Yang, G. (2004). The complexity of mining maximal frequent itemsets and maximal frequent patterns. In *In KDD '04 :*, pp. 344–353. ACM Press.

Summary

Condensed representations of patterns offer an elegant way to represent solution sets compactly, while minimizing the redundancy and the number of patterns. This approach has been mainly developed in the context of the frequency measure and there are very few works addressing other measures. We propose a generic framework based on constraint programming to efficiently mine adequate condensed representations of patterns w.r.t. a set of measures. For this, we introduce a new global constraint with a complete polynomial filtering. We show how this constraint can be exploited in association with Pareto dominance constraints to mine skypatterns. Experiments performed on standard datasets show the efficiency of our approach and its significant advantages over existing approaches.

Keywords: Condensed representation, Pareto dominance, Constraint programming, Pattern mining, Skyline, Skypatterns.

Comparaisons des mesures de centralité classiques et communautaires : une étude empirique

Stephany Rajeh *, Marinette Savonnet *, Éric Leclercq *, Hocine Cherifi *

*Laboratoire d'Informatique de Bourgogne EA 7534 - Université de Bourgogne
stephany.rajeh@u-bourgogne.fr

Résumé. Contrairement aux mesures de centralité classiques, les mesures de centralité communautaire récemment développées exploitent la structure communautaire du réseau pour identifier les nœuds influents. Dans cet article, on étudie les relations entres ces deux types de mesures de façon expérimentale en s'appuyant sur un ensemble de cinquante réseaux issus de divers domaines du monde réel. Les résultats montrent qu'elles sont, en général, faiblement corrélées. Ce qui démontre tout l'intérêt d'intégrer les propriétés mésoscopiques des réseaux dans la conception de mesures de centralité. On montre que les propriétés topologiques macroscopiques des réseaux qui influencent le plus les variations de corrélation observées sont la transitivité et l'efficacité. Quant aux propriétés mésoscopiques des réseaux, la modularité, le "*mixing parameter*" et le Max-ODF exercent l'influence la plus forte.

Mots-clés : Réseaux complexes, Centralité, Nœuds influents, Structure communautaire

1 Introduction

L'identification des nœuds influents est cruciale pour accélérer ou atténuer les processus de propagation dans les réseaux complexes. À cette fin, de nombreuses mesures classiques de centralité reposant sur diverses propriétés topologiques ont été proposées. On peut distinguer deux grandes catégories : les mesures locales et les mesures globales (Lü et al., 2016). Les mesures locales utilisent les informations dans le voisinage du nœud tandis que les mesures globales rassemblent les informations de l'ensemble du réseau. Notons que certains travaux combinent les informations locales et globales (Sciarra et al., 2018; Ibnoulouafi et al., 2018).

Un autre ensemble de mesures de centralité utilise des informations sur la structure communautaire (Orman et al., 2013; Kumar et al., 2018; Chakraborty et al., 2016) du réseau pour quantifier l'influence des nœuds (Cherifi et al., 2019). Dans cet article, nous les appelons des mesures de centralité "communautaires". Contrairement aux mesures de centralité classiques, celles-ci distinguent les liens intra-comm unautaires des liens inter-communautaires. Les liens intra-communautaires relient des nœuds appartenant à la même communauté. Ils sont liés à l'influence locale du nœud au sein de sa communauté. Les liens inter-communautaires relient

des nœuds appartenant à des communautés différentes. Ils quantifient donc l'impact du nœud au niveau global.

Les mesures de centralité communautaires diffèrent par la façon dont elles intègrent les liens intra-communautaires et inter-communautaires. La mesure "*Community Hub-Bridge*" proposée par (Ghalmane et al., 2019) sélectionne simultanément les "hubs" au sein des communautés de grandes tailles et les ponts entre communautés. "*Comm centrality*" (Gupta et al., 2015) combine les liens intra-communautaires et inter-communautaires d'un nœud en donnant la priorité à ces derniers. "*Community-based Centrality*" (Zhao et al., 2015) pondère un lien intra-communautaire par la taille de sa communauté et un lien inter-communautaire par la taille des communautés qu'il rejoint. "*K-shell with Community*" (Luo et al., 2016) est basé sur la combinaison linéaire du k-shell d'un nœud en considérant séparément les réseaux de liens intra-communautaires et de liens inter-communautaires. Le "*Participation Coefficient*" (Guimera et Amaral, 2005) et le "*Community-based Mediator*" (Tulu et al., 2018) tendent à sélectionner les nœuds importants en fonction de l'hétérogénéité de leurs liens intra-communautaires et inter-communautaires. Le *Participation Coefficient* d'un nœud diminue s'il ne participe à aucune autre communauté que la sienne. Le *Community-based Mediator* se réduit à la centralité de degré normalisée si la proportion de liens intra-communautaires et inter-communautaires d'un nœud sont égales. "*Modularity Vitality*" (Magelinski et al., 2021) est une mesure de centralité signée. Elle est basée sur la variation de la modularité lors de la suppression d'un nœud dans le réseau. Puisque les ponts relient différentes communautés, leur présence diminue la modularité. Par conséquent, les nœuds avec des valeurs négatives de *Modularity Vitality* sont des ponts. A l'inverse, puisque les hubs ont tendance à augmenter la modularité d'un réseau, les nœuds avec des valeurs de *Modularity Vitality* positives sont des hubs locaux.

De nombreuses études sont consacrées aux interactions entre les mesures de centralité classiques (Ronqui et Travieso, 2015; Schoch et al., 2017; Rajeh et al., 2020a; Oldham et al., 2019). Cependant, la relation entre les mesures de centralité classiques et communautaire est peu explorée (Rajeh et al., 2020b). Notre objectif dans cet article est de mieux appréhender ces relations. Ainsi, nous nous efforçons de répondre aux interrogations suivantes :

— Quelles sont les interactions entre les mesures de centralité classiques et les mesures de centralité communautaire ?
— Quelle est l'influence des propriétés topologiques des réseaux sur ces interactions ?

L'article est organisé comme suit. Tout d'abord, nous présentons les mesures de centralité classiques et les mesures de centralité communautaires. Dans les deux sections suivantes, les analyses portant sur la corrélation et l'influence de la topologie du réseau sont présentées. Enfin, nous résumons nos principales conclusions.

2 Mesures de centralité classiques et communautaires

Cette étude porte sur dix mesures de centralité classiques, dont cinq sont locales (Degree, Leverage, Laplacian, Diffusion et Maximum Neighborhood Component) et cinq sont globales (Betweenness, Closeness, Katz, PageRank et Subgraph). Le tableau 2 donne leur définition. Elles sont comparées à sept mesures communautaires introduites précédemment et décrites dans le tableau 3. Le tableau 1 présente les cinquante réseaux du monde réel utilisés dans cette étude. Ils proviennent de divers domaines (animaux, biologiques, collaborations, réseaux so-

Domaine	Nom et numéro du réseau
Réseaux d'animaux	Dolphins (1), Reptiles (2)
Réseaux biologiques	Budapest Connectome (3), Blumenau Drug (4), E. coli Transcription (5), Human Protein (6), Interactome Vidal (7), Kegg Metabolic (8), Malaria Genes (9), Mouse Visual Cortex (10), Yeast Collins (11), Yeast Protein (12)
Réseaux de collaboration	DBLP (13), AstroPh (14), C.S. PhD (15), GrQc (16), NetSci (17), New Zealand Collaboration (18)
Réseaux sociaux hors ligne	Adolescent health (19), Jazz (20), Zachary Karate Club (21), Madrid Train Bombings (22)
Réseaux d'infrastructure	EU Airlines (23), EuroRoad (24), Internet Autonomous Systems (25), Internet Topology Cogentco (26), London Transport (27), U.S. Power Grid (28), U.S. Airports (29), U.S. States (30)
Réseaux d'acteurs	Game of Thrones (31), Les Misérables (32), Marvel Partnerships (33), Movie Galaxies (34)
Réseaux divers	911AllWords (35), Bible Nouns (36), Board of Directors (37), DNC Emails (38), Football (39), Polbooks (40)
Réseaux sociaux en ligne	DeezerEU (41), Ego Facebook (42), Facebook Friends (43), Facebook Organizations (44), Caltech (45), Facebook Politician Pages (46), Hamsterster (47), PGP (48), Princeton (49), Retweets Copenhagen (50)

TAB. 1 – *Les cinquante réseaux du monde réel utilisés dans cette étude se divisent en huit domaines différents. Ces données peuvent être obtenues à partir des ressources citées (Rossi et Ahmed, 2015; Clauset et al., 2016; Latora et al., 2017; Peixoto, 2020; Kunegis, 2014).*

ciaux en ligne/hors ligne, infrastructure et divers). Comme la structure communautaire dépend de l'algorithme de détection de communauté utilisé, Louvain (Blondel et al., 2008) et Infomap (Rosvall et Bergstrom, 2008) sont utilisés pour extraire les liens intra-communautaires et inter-communautaires. En raison de contraintes d'espace, les caractéristiques topologiques des réseaux et les résultats basés sur Louvain sont fournis dans les documents complémentaires [1]. De plus, comme il n'y a pas de différences fondamentales, nous limitons notre attention à l'analyse des résultats basés sur la structure communautaire révélée par Infomap.

3 Analyse de la corrélation

La première expérimentation concerne la corrélation entre les mesures de centralité classique et communautaire pour un réseau donné. Ainsi, pour chacun des cinquante réseaux, la corrélation du Kendall's Tau est calculée pour toutes les combinaisons possibles entre les dix mesures de centralité classiques (α_i) et les sept mesures de centralité communautaires (β_j). La figure 1 montre les distributions des valeurs de corrélation pour chaque réseau. On peut

1. https ://github.com/StephanyRajeh/MixedCommunityAwareCentralityAnalysis

Description de la mesure de centralité	Définition
1) Degree : la somme totale des voisins d'un nœud	$\alpha_d(i) = \sum_{j=1}^{N} a_{ij}$
2) Leverage : quantité de connexions par rapport à ses voisins	$\alpha_{lev}(i) = \frac{1}{k_i} \sum_{j=1}^{N} \frac{k_i - k_j}{k_i + k_j}$
3) Laplacian : combien de dommages un nœud cause après sa suppression	$\alpha_{lap}(i) = k_i^2 + k_i + 2 \sum_{j \in \mathcal{N}_1(i)} k_j$
4) Diffusion : le pouvoir de diffusion d'un nœud et celui de ses voisins	$\alpha_{dif}(i) = \varpi_i \times \alpha_d(i) + \sum_{j \in \mathcal{N}_1(i)} \varpi_j \times \alpha_d(j)$
5) Max. Neighbor. Component : la taille de la plus grande composante connecté (LCC) du nœud établie par son voisinage	$\alpha_m(i) = \lvert LCC \in \mathcal{N}_1(i) \rvert$
6) Betweenness : le nombre de chemin le plus court dans lequel se trouve un nœud entre deux autres nœuds	$\alpha_b(i) = \sum_{s,t \neq i} \frac{\sigma_i(s,t)}{\sigma(s,t)}$
7) Closeness : à quelle distance, en moyenne, un nœud est-il de tous les autres nœuds du réseau	$\alpha_c(i) = \frac{N-1}{\sum_{j=1}^{N-1} d(i,j)}$
8) Katz : la quantité, la qualité et les distances ultérieures des autres nœuds connectés à un nœud	$\alpha_k(i) = \sum_{p=1} \sum_{j=1} s^p a_{ij}^p$
9) PageRank : la quantité et la qualité des nœuds connectés à un nœud sous un processus de chaîne de Markov	$\alpha_p(i) = \frac{1-d}{N} + d \sum_{j \in \mathcal{N}_1(i)} \frac{\alpha_p(j)}{k_j}$
10) Subgraph : la participation d'un nœud à des voisins avec des chemins commençant et se terminant par le même nœud	$\alpha_s(i) = \sum_{j=1}^{N} (v_j^i)^2 e^{\lambda_j}$

TAB. 2 – *Définition des mesures de centralité (α_j). $a_{i,j}$ désigne la connectivité d'un nœud i au nœud j de la matrice d'adjacence A. N est le nombre total de nœuds. k_i et k_j sont les degrés des nœuds i et j, respectivement. $\mathcal{N}_1(i)$ est l'ensemble des voisins directs du nœud i. ϖ_i et ϖ_j sont les probabilités de propagation des nœuds i et j, respectivement ϖ est fixé à 1 pour tous les nœuds dans cette étude). $\sigma(s,t)$ est le nombre de chemins les plus courts entre les nœuds s et t. et $\sigma_i(s,t)$ est le nombre de chemins les plus courts entre les nœuds s et t qui passent par le nœud i. $d(i,j)$ est la distance du chemin le plus court entre le nœud i et j. a_{ij}^p est la connectivité du nœud i par rapport à tous les autres nœuds à un ordre donné de la matrice d'adjacence A^p. s^p est le facteur d'atténuation où $s \in [0,1]$. $\alpha_p(i)$ et $\alpha_p(j)$ sont les centralités PageRank du nœud i et du nœud j, respectivement. d est le paramètre d'amortissement (fixé à 0,85 dans cette étude). v_j désigne un vecteur propre de la matrice d'adjacence A, associé à sa valeur propre λ_j.*

noter une faible cohérence des distributions pour les réseaux d'un même domaine. En effet, elles peuvent être très différentes. Par exemple, bien que EU Airlines (23) et EuroRoad (24) appartiennent au domaine des réseaux d'infrastructure (couleur grise), EU Airlines (23) a une distribution étalée alors que EuroRoad (24) est beaucoup plus concentrée. On peut remarquer que, quelque soit le réseau considéré, la valeur la plus fréquente de la distribution se situe

Description de la mesure de centralité	Définition
1) Community Hub-Bridge (Ghalmane et al., 2019) : pondération des liens intra-communautaires par la taille de la communauté du nœud et des liens inter-communautaires par le nombre de communautés voisines du nœud	$\beta_{CHB}(i) = \lvert c_k \rvert \times k_i^{intra} + \lvert NNC_i \rvert \times k_i^{inter}$
2) Participation Coefficient (Guimera et Amaral, 2005) : l'hétérogénéité des liens d'un nœud	$\beta_{PC}(i) = 1 - \sum_{c=1}^{N_c} \left(\frac{k_{i,c}}{k_i}\right)^2$
3) Community-based Mediator (Tulu et al., 2018) : l'entropie des liens intra-communautaires et inter-communautaires d'un nœud	$\beta_{CBM}(i) = H_i \times \frac{k_i}{\sum_{i=1}^{N} k_i}$
4) Comm Centrality (Gupta et al., 2016) : pondération des liens intra-communautaires et inter-communautaires par la proportion de liens externes et prioriser les ponts	$\beta_{Comm}(i) = (1+\mu_{c_k}) \times \chi + (1-\mu_{c_k}) \times \varphi^2$
5) Modularity Vitality (Magelinski et al., 2021) : le changement de modularité qu'un nœud provoque après sa suppression du réseau	$\beta_{MV}(i) = M(G_i) - M(G)$
6) Community-based Centrality (Zhao et al., 2015) : pondéreration des liens intra-communautaires et inter-communautaires en fonction de la taille de leurs communautés d'appartenance	$\beta_{CBC}(i) = \sum_{c=1}^{N_c} k_{i,c} \left(\frac{n_c}{N}\right)$
7) K-shell with Community (Luo et al., 2016) : la décomposition hiérarchique k-shell du réseau local (des liens intra-communautaires) et du réseau global (des liens inter-communautaires)	$\beta_{ks}(i) = \delta \times \beta^{intra}(i) + (1-\delta) \times \beta^{inter}(i)$

TAB. 3 – *Définitions des mesures de centralité communautaires ($\beta(i)$). c_k représente la k-ième communauté. k_i^{intra} et k_i^{inter} représentent les liens intra-communauté et inter-communauté d'un nœud. N_c est le nombre total de communautés. $k_{i,c}$ est le nombre de liens du nœud i dans une communauté c. k_i est le degré total du nœud i. N est le nombre total de nœuds. $H_i = [-\sum \rho_i^{intra} log(\rho_i^{intra})] + [-\sum \rho_i^{inter} log(\rho_i^{inter})]$ est l'entropie du nœud i basée sur ses ρ^{intra} et ρ^{inter} qui représentent la densité des communautés auxquelles un nœud est lié. $\chi = \frac{k_i^{intra}}{max_{(j \in c)} k_j^{intra}} \times R$ et $\varphi = \frac{k_i^{inter}}{max_{(j \in c)} k_j^{inter}} \times R$. μ_{c_k} est la proportion de liens inter-communautaires sur le total des liens communautaires dans la communauté c_k. R, une constante, pour mettre à l'échelle les valeurs des liens dans le même intervalle. M est la modularité d'un réseau et $M(G_i)$ est la modularité du réseau après la suppression du nœud i. n_c est le nombre de nœuds dans la communauté c. $\beta^{intra}(i)$ et $\beta^{inter}(i)$ représentent la valeur k-shell du nœud i en considérant uniquement les liens intra-communautaires et les liens inter-communautaires, respectivement. Dans cette étude, δ est fixé à 0,5.*

autour de 0,5. La médiane moyenne de toutes les distributions est de 0,43±0,1. L'écart interquartile moyen est de 0,37±0,1. Enfin, la moyenne de la distribution pour tous les réseaux est de 0,37±0,07. En d'autres termes, la majeure partie des mesures de centralité classiques et

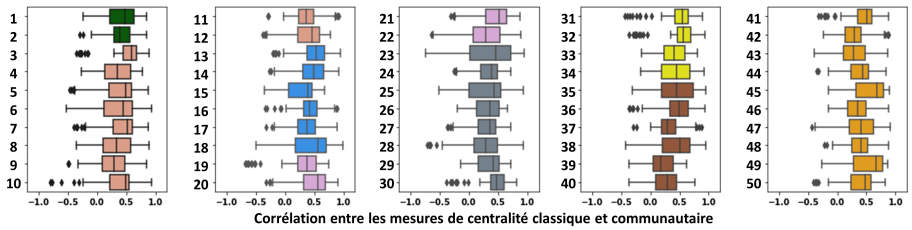


FIG. 1 – *Boxplots de la corrélation de Kendall's Tau entre les mesures de centralité classique et communautaire pour chaque réseau. Les couleurs représentent le domaine du réseau. Les réseaux animaux sont en vert, les réseaux biologiques en rose. Les réseaux de collaboration sont en bleu. Les réseaux sociaux hors ligne sont en violet. Les réseaux d'infrastructure sont en gris. Les réseaux d'acteurs sont jaunes. Les réseaux divers sont marron. Les réseaux sociaux en ligne sont en orange.*

communautaires ont tendance à présenter des valeurs de corrélation moyennes ou faibles.

Pour vérifier la cohérence des valeurs de corrélation du Kendall's Tau pour les différentes paires de centralités communautaires et classiques dans les réseaux, nous procédons comme suit. Chaque réseau est représenté par un échantillon composé de trente-cinq valeurs de paires de corrélation. Les valeurs de corrélation de Pearson entre les échantillons deux à deux sont ensuite calculées pour quantifier la proximité statistique des deux réseaux. La figure 2 (A) illustre sa distribution. Globalement, les résultats sur les réseaux sont bien corrélés. En effet, les valeurs de corrélation de Pearson varient entre 0,6 et 1. Plus précisément, leur valeur moyenne est égale à 0,80, et leur médiane à 0,82. Notez que 911AllWords, Football et Ego Facebook s'écartent de la tendance générale. C'est la raison pour laquelle la distribution a une grosse queue à gauche. On peut donc conclure que la corrélation entre les mesures de centralité classiques et communautaire dans les réseaux est plutôt cohérente.

Enfin, après avoir vérifié que les valeurs de corrélation du Kendall's Tau sont cohérentes entre les réseaux, nous calculons la moyenne et l'écart type pour chaque combinaison (α_i, β_j) sur les cinquante réseaux. Cela permet d'étudier si les mesures de centralité sensibles à la communauté se comportent différemment. Les résultats présentés dans les figure 2 (B) et (C) montrent que les schémas de corrélation des diverses mesures de centralité communautaire sont très différents. *Modularity Vitality* (β_{MV}) est la seule mesure de centralité communautaire qui présente une corrélation négative avec les mesures de centralité classiques. En outre, son écart type moyen est élevée. Étant donné qu'il s'agit d'une mesure de centralité communautaire signée, ce résultat n'est pas inattendu. Les autres mesures de centralité communautaire peuvent être classées en fonction de leurs valeurs de corrélation. Le *Community Hub-Bridge* (β_{CHB}) et le coefficient de participation (β_{PC}) ont tendance à présenter une corrélation moyenne positive faible avec toutes les mesures de centralité classiques ($\leq$ 0,4), à l'exception de (α_b, β_{PC}) qui s'élève à 0,46. Leur écart-type moyen est généralement proche de 0,15. *Comm Centrality* (β_{Comm}) présente une corrélation moyenne minimale de 0,27 et une corrélation moyenne maximale de 0,54. L'écart type de β_{Comm} varie de 0,11 à 0,21. Vient ensuite *Community-based Mediator* (β_{CBM}), dont la corrélation moyenne est comprise entre 0,43 et 0,6. Son écart-type moyen est proche de 0,15 pour toutes les combinaisons, sauf pour (α_m, β_{CBM}) qui

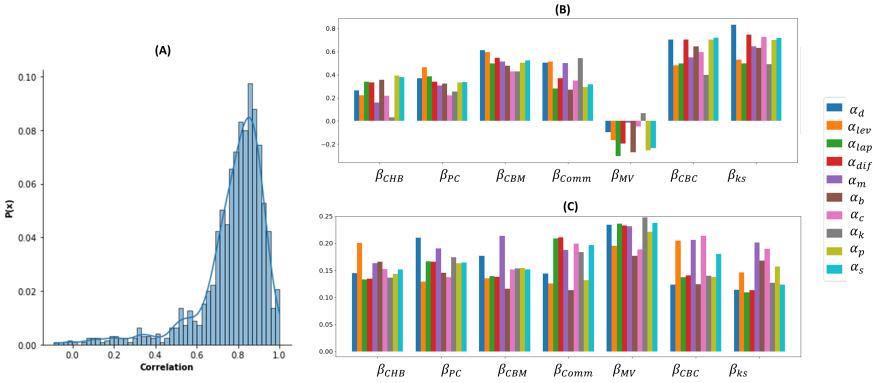


FIG. 2 – *(A) Distribution de la corrélation de Pearson pour la corrélation de Kendall's Tau entre la centralité classique et communautaire de tous les réseaux. (B) Moyenne et (C) écart-type de la corrélation Kendall's Tau pour chaque paire de mesures de centralité classique et communautaire (α_i, β_j) sur les cinquante réseaux.*

s'élève à 0,21. Enfin, *Community-based Centrality* (β_{CBC}) et le *K-shell with Community* (β_{ks}) présentent une corrélation plus élevée avec les mesures de centralité classiques que les autres mesures de centralité sensibles à la communauté. En effet, la corrélation moyenne peut même atteindre 0,83 comme maximum (α_d, β_{ks}). Leur écart-type est compris entre 0,14 et 0,21. Ces résultats corroborent l'observation de valeurs élevées de la corrélation dans la distribution de chaque réseau rapportée dans la figure 1. En effet, ces valeurs correspondent à β_{CBC} et β_{ks}.

4 Analyse de la topologie du réseau

Les valeurs de corrélation entre les mesures de centralité classiques et de centralité communautaires de chaque réseau sont étudiées. Pour un réseau donné, chaque mesure de centralité communautaire est réduite à la valeur moyenne des valeurs de corrélation du Kendall's Tau calculées pour les dix mesures de centralité classiques. Une régression linéaire simple est effectuée pour étudier la relation avec les différentes propriétés topologiques des réseaux. Les valeurs de corrélation moyennes sont les variables dépendantes, tandis que les propriétés topologiques sont les variables indépendantes. Les caractéristiques macroscopiques utilisées sont la densité, la transitivité, l'assortativité, la distance moyenne, le diamètre, l'efficacité et l'exposant de la distribution des degrés. Les caractéristiques mésoscopiques utilisées sont la modularité, le mixing parameter, la distance interne, la densité interne, *Max-ODF*, *Average-ODF*, *Flake-ODF*, *embeddedness* et *hub dominance*. Si la valeur p est inférieure à 0,05, la relation entre les variables dépendantes et indépendantes est considérée comme statistiquement significative. La figure 3 présente les deux cas extrêmes de dépendance statistique entre la moyenne et les caractéristiques topologiques : la transitivité et le mixing parameter. Le premier cas concerne *Community-based Mediator* (la valeur moyenne présente des relations linéaires significatives avec neuf caractéristiques topologiques). Le dernier cas concerne *Modularity Vi-*

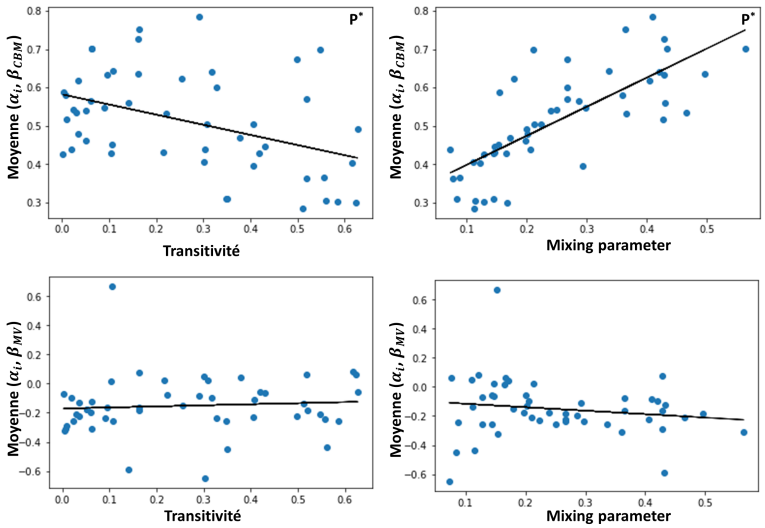


FIG. 3 – *Relation de la moyenne de la corrélation entre les centralités communautaires "Community-based Mediator (β_{CBM})" et "Modularity Vitality (β_{MV})" combinées à toutes les mesures de centralité classiques en fonction des propriétés topologiques des réseaux du monde réel. La droite est ajustée par régression linéaire en utilisant les moindres carrés ordinaires. "P" indique $p \leq 0.05$. "P" et * indiquent $p \leq 0.01$.*

tality. La valeur moyenne ne montre aucune relation linéaire significative avec une quelconque propriété topologique. Les autres figures et l'estimation des paramètres de régression linéaire pour chaque mesure de centralité communautaire sont fournies dans les documents supplémentaires.

En ce qui concerne les propriétés topologiques macroscopiques, nous observons trois situations. Les caractéristiques du réseau présentent une relation linéaire significative avec la moyenne de la centralité communautaire trois, deux ou aucune fois. Dans ce sens, la transitivité et l'efficacité sont les caractéristiques topologiques macroscopiques les plus influentes. Elles présentent des relations significatives avec la moyenne de trois mesures différentes de centralité communautaire. Viennent ensuite la densité, l'assortativité, le diamètre et la distance moyenne qui affectent deux mesures de centralité communautaire. Enfin, l'exposant de la distribution des degrés est la seule caractéristique topologique parmi les caractéristiques macroscopiques qui ne présente aucune relation significative. La transitivité a une association négative significative avec la moyenne du *Community-based Mediator* (β_{CBM}) et du *Participation Coefficient* (β_{PC}). En effet, une augmentation de la transitivité entraîne un plus grand nombre de triangles dans le réseau. Comme le *Community-based Mediator* est basé sur l'entropie des liens intra-communautaires et inter-communautaires d'un nœud, la transitivité peut augmenter la différence entre les deux, ce qui entraîne une corrélation plus faible. Comme le *Participation Coefficient* exploite également la marge de la proportion des liens inter-communautaires et intra-communautaires, il se comporte de manière similaire. On observe une association positive

avec la transitivité et la moyenne de *Community-based Centrality* (β_{CBC}). Si l'ensemble du réseau forme une seule communauté, β_{CBC} se réduit à la centralité du degré (Zhao et al., 2015). Par conséquent, la corrélation entre β_{CBC} et les mesures classiques a tendance à augmenter lorsque la transitivité augmente. L'efficacité a une association positive significative sur *Comm centrality* (β_{Comm}), *Community-based Centrality* (β_{CBC}), et le *K-shell with Community* (β_{ks}). Une augmentation de l'efficacité signifie que la distance moyenne du plus court chemin dans un réseau devient plus petite. En d'autres termes, le réseau est plus efficace lorsque les nœuds sont étroitement connectés. Par conséquent, les mesures de centralité communautaire ont tendance à être plus corrélées avec les mesures classiques. La densité présente une association positive significative avec *Comm Centrality* (β_{Comm}) et *Community-based Centrality* (β_{CBC}). Une augmentation de la densité signifie plus de liens entre les nœuds. Par conséquent, β_{Comm} et β_{CBC} deviennent plus analogues avec les mesures de centralité classiques. L'assortativité a une association négative significative avec la moyenne de la *Community-based Mediator* (β_{CBM}) et *Participation Coefficient* (β_{PC}). Une augmentation de l'assortativité signifie qu'il y a plus d'interactions entre les pairs dans les réseaux. Elle peut également augmenter la marge de différence entre les liens intra-communautaires et inter-communautaires. Les réseaux assortatifs ont tendance à former des communautés avec des nœuds de degré "similaire". Par conséquent, les densités de liens intra-communautaires et inter-communautaires peuvent différer d'avantage d'une communauté à l'autre. On observe donc une corrélation plus faible entre β_{CBM}/β_{PC} et les mesures de centralité classiques. Le diamètre et la distance moyenne ont tous deux une association négative significative avec la moyenne de *Community-based Centrality* (β_{CBC}) et du *K-shell with Community* (β_{ks}). Une augmentation des deux mesures signifie que les nœuds sont plus éloignés les uns des autres. Ces deux mesures de centralités communautaires sont les plus sensibles aux mesures liées à la distance.

En ce qui concerne les caractéristiques topologiques mésoscopiques, on peut distinguer deux cas. Le mixing parameter, la modularité et le Max-ODF sont statistiquement liés de manière linéaire à la moyenne de trois mesures de centralité communautaire. Pour les autres caractéristiques, il existe une dépendance linéaire avec la moyenne de deux mesures de centralité sensibles à la communauté. Le mixing parameter a une association positive significative avec la moyenne sur le *Community Hub-Bridge* (β_{CHB}), le *Participation Coefficient* (β_{PC}) et le *Community-based Mediator* (β_{CBM}). Une augmentation du mixing parameter se traduit par une structure communautaire plus faible. Par conséquent, ces mesures de centralité sensibles à la communauté ont tendance à extraire des informations similaires à celles des mesures de centralité classiques. La modularité a une association négative significative avec *Community-based Mediator* (β_{CBM}), *Community-based Centrality* (β_{CBC}), et *K-shell with Community* (β_{ks}). Une augmentation de la modularité signifie que les communautés sont étroitement liées. Par conséquent, ces mesures extraient des informations différentes de celles des mesures de centralité classiques lorsque le réseau est hautement modulaire. Le Max-ODF présente une association positive significative avec *Community-based Mediator* (β_{CBM}), *Community-based Centrality* (β_{CBC}) et *K-shell with Community* (β_{ks}). En se basant sur les nœuds ayant les liens inter-communautaires les plus élevés dans leur communauté, son augmentation conduit à plus de connexions entre les nœuds hautement connectés dans différentes communautés, ce qui affaiblit la structure communautaire. Par conséquent, la corrélation de β_{CBM}, β_{CBC} et β_{ks} avec les mesures de centralité classiques augmente. La distance interne présente une relation linéaire positive significative avec le *Particiaption Coefficient* (β_{PC}) et une relation néga-

tive avec *Community-based Centrality* (β_{CBC}). Comme β_{PC} exploite l'hétérogénéité entre les liens intra-communautaires et inter-communautaires d'un nœud, une augmentation de l'interne diminue la marge entre les liens intra-communautaires et inter-communautaires. Par conséquent, la corrélation entre β_{PC} et les mesures de centralité classiques augmente. L'effet inverse se produit avec β_{CBC}. La densité interne a une influence négative sur *Community-based Mediator* (β_{CBM}) et *Participation Coefficient* (β_{PC}). Une augmentation de la densité interne signifie que les communautés sont condensées avec des connexions internes. Comme β_{CBM} et β_{PC} exploitent la marge de différence entre les liens intra-communautaires d'un nœud et ses liens inter-communautaires, les deux favoriseront une augmentation de la densité interne. Le Average-ODF a une relation positive significative avec *Community-based Mediator* (β_{CBM}) et du *K-shell with Community* (β_{ks}). Puisqu'elle est basée sur la proportion de liens inter-communautaires, plus la structure de la communauté est faible, plus la corrélation avec les mesures classiques de centralité est élevée. De même, le Flake-ODF présente une relation linéaire positive similaire avec la moyenne de β_{CBM} et de β_{ks}. En effet, il s'agit d'une autre façon de quantifier la force de la structure de la communauté. Embeddedness a une relation négative avec *Community-based Mediator* (β_{CBM}) et du *K-shell with Community* (β_{ks}). En effet, en se basant sur la proportion de liens intra-communautaires, elle est l'inverse de Average-ODF. Enfin, le hub dominance a une relation positive significative avec *Community-based Centrality* (β_{CBC}) et *K-shell with Community* (β_{ks}). Un hub dominance plus élevée signifie que les communautés étroitement connectées sont moins nombreuses. Par conséquent, le comportement de β_{CBC} se rapproche de la centralité de degré, et la corrélation de β_{CBC} avec les mesures de centralité classiques augmente. En ce qui concerne β_{ks}, une plus grande dominance du hub induit des liens intra-communautaires et inter-communautaires plus similaires et une corrélation plus élevée avec les mesures de centralité classiques.

5 Conclusion

L'étude que nous avons présentée examine la relation entre les mesures de centralité classiques et les mesures de centralité sensibles à la communauté. Les résultats montrent que la corrélation du Kendall's Tau entre les mesures de centralité classiques et celles qui tiennent compte de la structure communautaires est généralement moyenne à faible. Ensuite, les modèles de corrélation sont assez cohérents entre les réseaux. De plus, les mesures de centralité communautaires peuvent être classées en quatre groupes selon le type de corrélation avec les mesures de centralité classiques. Plus précisément, *Modularity Vitality* montre une faible corrélation négative. Une corrélation positive faible caractérise *Community Hub-Bridge*, *Participation Coefficient* et *Comm Centrality*. Une corrélation positive moyenne est observée pour *Community-based Mediator*. Enfin, *Community-based Centrality* et *K-shell with Community* présentent une corrélation positive élevée. La transitivité et l'efficacité sont les caractéristiques macroscopiques les plus influentes, tandis que le mixing parameter, la modularité et le Max-ODF sont les caractéristiques mésoscopiques prédominantes. Les résultats de cette étude ouvrent la voie au développement de mesures efficaces de centralité communautaires. En effet, elle démontre que l'intégration des connaissances sur la structure de la communauté du réseau apporte une nouvelle perspective de l'influence des nœuds.

Références

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment 2008*(10), P10008.

Chakraborty, D., A. Singh, et H. Cherifi (2016). Immunization strategies based on the overlapping nodes in networks with community structure. In *International conference on computational social networks*, pp. 62–73. Springer, Cham.

Cherifi, H., G. Palla, B. K. Szymanski, et X. Lu (2019). On community structure in complex networks : challenges and opportunities. *Applied Network Science 4*(1), 1–35.

Clauset, A., E. Tucker, et M. Sainz (2016). The colorado index of complex networks. [Online]. Available : `https://icon.colorado.edu/`.

Ghalmane, Z., M. El Hassouni, et H. Cherifi (2019). Immunization of networks with non-overlapping community structure. *SNAM 9*(1), 1–22.

Guimera, R. et L. A. N. Amaral (2005). Functional cartography of complex metabolic networks. *nature 433*(7028), 895–900.

Gupta, N., A. Singh, et H. Cherifi (2015). Community-based immunization strategies for epidemic control. In *2015 7th international conference on communication systems and networks (COMSNETS)*, pp. 1–6. IEEE.

Gupta, N., A. Singh, et H. Cherifi (2016). Centrality measures for networks with community structure. *Physica A : Statistical Mechanics and its Applications 452*, 46–59.

Ibnoulouafi, A., M. El Haziti, et H. Cherifi (2018). M-centrality : identifying key nodes based on global position and local degree variation. *Journal of Statistical Mechanics : Theory and Experiment 2018*(7), 073407.

Kumar, M., A. Singh, et H. Cherifi (2018). An efficient immunization strategy using overlapping nodes and its neighborhoods. In *Companion Proceedings of the The Web Conference 2018*, pp. 1269–1275.

Kunegis, J. (2014). Handbook of network analysis [konect project]. *arXiv preprint arXiv :1402.5500*.

Latora, V., V. Nicosia, et G. Russo (2017). *Complex networks : principles, methods and applications*. Cambridge University Press. [Online]. Available : `https://www.complex-networks.net/datasets.html`.

Lü, L., D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, et T. Zhou (2016). Vital nodes identification in complex networks. *Physics Reports 650*, 1–63.

Luo, S.-L., K. Gong, et L. Kang (2016). Identifying influential spreaders of epidemics on community networks. *arXiv preprint arXiv :1601.07700*.

Magelinski, T., M. Bartulovic, et K. M. Carley (2021). Measuring node contribution to community structure with modularity vitality. *IEEE Transactions on Network Science and Engineering 8*(1), 707–723.

Oldham, S., B. Fulcher, L. Parkes, A. Arnatkevicute, C. Suo, et A. Fornito (2019). Consistency and differences between centrality measures across distinct classes of networks. *PloS one 14*(7).

Orman, K., V. Labatut, et H. Cherifi (2013). An empirical study of the relation between community structure and transitivity. In *Complex Networks*, pp. 99–110. Springer, Berlin, Heidelberg.

Peixoto, T. P. (2020). The netzschleuder network catalogue and repository. [Online]. Available : `https://networks.skewed.de/`.

Rajeh, S., M. Savonnet, E. Leclercq, et H. Cherifi (2020a). Interplay between hierarchy and centrality in complex networks. *IEEE Access 8*, 129717–129742.

Rajeh, S., M. Savonnet, E. Leclercq, et H. Cherifi (2020b). Investigating centrality measures in social networks with community structure. In *International Conference on Complex Networks and Their Applications*, pp. 211–222. Springer.

Ronqui, J. R. F. et G. Travieso (2015). Analyzing complex networks through correlations in centrality measurements. *Journal of Statistical Mechanics : Theory and Experiment 2015*(5), P05030.

Rossi, R. A. et N. K. Ahmed (2015). The network data repository with interactive graph analytics and visualization. In *AAAI*. [Online]. Available : `http://networkrepository.com`.

Rosvall, M. et C. T. Bergstrom (2008). Maps of random walks on complex networks reveal community structure. *PNAS 105*(4), 1118–1123.

Schoch, D., T. W. Valente, et U. Brandes (2017). Correlations among centrality indices and a class of uniquely ranked graphs. *Social Networks 50*, 46–54.

Sciarra, C., G. Chiarotti, F. Laio, et L. Ridolfi (2018). A change of perspective in network centrality. *Scientific reports 8*(1), 1–9.

Tulu, M. M., R. Hou, et T. Younas (2018). Identifying influential nodes based on community structure to speed up the dissemination of information in complex network. *IEEE Access 6*, 7390–7401.

Zhao, Z., X. Wang, W. Zhang, et Z. Zhu (2015). A community-based approach to identifying influential spreaders. *Entropy 17*(4), 2228–2252.

Summary

Identifying influential nodes in social networks is a fundamental issue. Indeed, it has many applications, such as inhibiting epidemic spreading, accelerating information diffusion, and preventing terrorist attacks. Classical and community-aware centrality measures are two main approaches for identifying influential nodes in complex networks. Nonetheless, both contrast in the way they locate these nodes. The first exploits the overall network topology while the second exploits the community structure of the network. This study investigates the relationship between classical and community-aware centrality measures on a set of 50 real-world networks of diverse domains. Results show that the correlation between representative measures of these two approaches ranges from low to medium values. Additionally, transitivity and mixing parameter are critical network topological properties driving their interactions.

Keywords: Complex networks, Centrality, Influential nodes, Community structure

Algorithme quantique pour trouver les séparateurs d'un graphe orienté

Ahmed ZAIOU[*,**], Younès BENNANI[**]
Mohamed HIBTI[*] , Basarab MATEI[**]

[*]EDF R&D, Palaiseau, France
LIPN & LaMSN, USPN, Villetaneuse, France
prénom.nom@edf.fr
[**]LIPN & LaMSN. Université Sorbonne Paris Nord
Villetaneuse, France
prénom.nom@sorbonne-paris-nord.fr

Résumé. Le problème de séparateur des sommets d'un graphe orienté consiste à trouver toutes les combinaisons de sommets qui peuvent déconnectés la source et le terminal du graphe, ces combinaisons sont minimales si ne contiennent que le nombre minimal de sommets. Dans ce papier, nous introduisons un nouveau algorithme quantique basé sur une stratégie de mouvement pour trouver ces séparateurs d'un coup dans une superposition quantique avec une complexité linéaire.

Mots-clés : Calcul quantique, Séparateur d'un graphe orienté, Coupes minimales

1 Introduction

L'informatique quantique (Nielsen et Chuang, 2002) a suscité un énorme intérêt au cours des dernières années et attiré de nombreux chercheurs dans différentes disciplines. Cet engouement a fait suite aux deux principaux algorithmes révolutionnaires introduits par Grover et Shor. Le premier algorithme introduit par Grover (Grover, 1996), arrive à réduire la complexité à $O(\sqrt{N})$ pour trouver un élément dans une base de données non structurée de taille N, le second par Shor (Shor, 1999), qui peut casser le code RSA en un temps polynomial. L'un des principaux objectifs de ce nouveau domaine de recherche est de résoudre des problémes qui ne peuvent être résolus dans le cadre classique et de briser la complexité informatique de nombreux problèmes difficiles et parfois de trouver de bons raccourcis et de nouvelles approches pour les résoudre. En 2012, John Preskill a introduit le terme "quantum supremacy", un concept pour décrire le point que les ordinateurs quantiques peuvent faire des choses que les ordinateurs classiques ne peuvent pas (Preskill, 2012). Différents algorithmes ont ensuite été proposés pour atteindre cet objectif (Sycamore (Pednault et al., 2019), Chinois (Zhong et al., 2020)). D'autres ont montré une accélération importante par rapport aux meilleurs résultats des ordinateurs classiques. Dans le domaine de la théorie des graphes, nous pouvons

mentionner l'article de Shawn at al (Cui et al., 2016), qui montre que, contrairement au cas classique, la conjecture max-flow/min-cut quantique n'est pas vraie en général. Sous certaines conditions, par exemple, lorsque la capacité de chaque arête est une puissance d'un nombre entier fixe, il est prouvé que le flux maximal quantique est égal à la coupe minimale quantique. Et également ils ont trouvé des connexions du max-flow/min-cut quantique avec l'entropie de l'intrication et le problème de la satisfiabilité quantique. Et aussi le papier de Stacey Jeffery et Shelby Kimmel (Jeffery et Kimmel, 2017) qui donnent une nouvelle limite supérieure à la complexité quantique des requêtes pour décider la st-connectivité sur certaines classes de graphes planaires, et montre que cette limite est parfois exponentiel et aussi que l'évaluation des formules booléennes se réduit à décider la connectivité sur une telle classe de graphes. Cet article donne pour certaines classes de formules booléennes une accélération quadratique par rapport à la complexité classique. Kazuya et Ryuhei (Shimizu et Mori, 2021) ont fourni un algorithme quantique en temps exponentiel pour calculer le nombre chromatique, et plusieurs autres travaux ont été publiés dans ces domaines, comme (Dörn, 2007) sur les problémes de parcours de graphe, qui montre des résultats qui améliore les meilleurs algorithmes classiques pour les problémes de circuit Eulérien ou Hamiltonien, le probléme du voyageur de commerce et l'ordonnancement de projets. (Magniez et al., 2007) présente deux algorithmes quantique soit pour trouver un triangle dans un graphe non orienté soit rejeté si le graphe sans triangle. Le premier algorithme utilise des idées combinatoires avec Grover Search et le second algorithme basé sur un concept d'Ambainis (Ambainis, 2007).

Dans cet article, nous abordons le problème du séparateur de sommets d'un graphe orienté qui n'a pas encore été résolu par un algorithme quantique précis. Nous introduisons donc un algorithme quantique pour résoudre ce problème et aussi nous montrons que cet algorithme est facilement applicable dans les cadres existants de partitionnement de graphes et qu'il est également faisable sur le plan informatique.

L'article est organisé comme suit : Dans la section suivante, nous présentons une description formelle du problème du séparateur de sommets. Dans la section 3 nous introduisons comment nous représentons les séparateurs (les coupes) avec des états quantiques et nous définissons c'est quoi un oracle de mouvement et nous décrivons notre algorithme et sa complexité.

Dans la section 4, une analyse comparative est effectuée sur des bases qualitatives et sur la base de quelques cas de test. L'article est finalement conclu dans la section 5.

2 Problématique et l'état de l'art

Nous modélisons notre problème par des graphes orientés (réseaux). En théorie des graphes, le problème du séparateur de sommets d'un graphe (VSP) consiste à trouver un sous-ensemble de sommets (appelé séparateur de sommets) qui permet de diviser l'ensemble des sommets du graphe en deux composantes connexes. Le VSP est NP-hard (Bui et Jones, 1992). Il existe un certain nombre d'algorithmes qui peuvent trouver ces séparateurs de sommets. Nous mentionnons (Kernighan et Lin, 1970) qui présente une méthode heuristique de partitionnement de graphes arbitraires qui est à la fois efficace pour trouver des partitions optimales et suffisamment rapide pour être pratique dans la résolution de gros problèmes. Le papier de Fiduccia-Mattheyses (Fiduccia et Mattheyses, 1982) présente une méthode heuristique en temps linéaire pour améliorer les partitions de réseau. Ces deux articles (Kernighan et Lin, 1970) et (Fiduccia et Mattheyses, 1982) sont adaptés par Ashcraft C. et Liu J. (Ashcraft et Liu, 1994) et Hen-

drickson B. et Rothberg E. (Hendrickson et Rothberg, 1998) pour généraliser les méthodes et Améliorer le temps d'exécution.

Selon Hager W. et Hungerford J. (Hager et Hungerford, 2015), le problème de séparateur des sommets peut être formulé par un problème de programmation quadratique bilinéaire. Et, récemment, plusieurs travaux (Davis et al., 2019), (Hager et al., 2018), (Kolodziej et Davis, 2016) et (Kolodziej, 2019) ont combiné la méthode combinatoire traditionnelle et la méthode basée sur l'optimisation pour améliorer la performance et la qualité du séparateur. Nous mentionnons le résultat de Kolodziej S. et Davis T. (Kolodziej et Davis, 2020) qui a introduit un nouvel algorithme hybride pour calculer les séparateurs de sommets dans les graphes arbitraires en utilisant l'optimisation computationnelle.

Dans l'approche classique du problème de séparateur des sommets pour un graphe planaire avec n sommets, Lipton et Tarjan (Lipton et Tarjan, 1979) ont fourni un algorithme en temps polynomial pour trouver un séparateur de sommets. Cet algorithme a été amélioré dans (Lipton et Tarjan, 1980) pour d'autres familles de graphes comme les graphes de genre fixe. Ces familles de graphes incluent les arbres, les grilles 3D et les mailles qui ont de petits séparateurs. Pour obtenir tous les séparateurs de sommets minimaux d'un graphe, Kloks et Kratsch (Kloks et Kratsch, 1998) ont fourni un algorithme efficace listant tous les séparateurs de sommets minimaux d'un graphe non orienté . L'algorithme nécessite un temps polynomial par séparateur trouvé.

Dans cet article, nous nous intéressons au probléme du séparateur de sommets (VSP) dans un graphe orienté qui a une source s et un terminal t. commençons par la définition d'un graphe orienté et un séparateur de sommets :

Définition : Un graphe orienté ou un réseau est un graphe dans lequel les arêtes ont des orientations. Plus précisément, un graphe orienté est une paire ordonnée (V, E) comprenant : (i) V un ensemble de sommets et (ii) $E \subset \{(x, y)|(x, y) \in V \times V, x \neq y\}$ un ensemble d'arêtes ou d'arcs orientés qui sont des paires de sommets ordonnées et distincts.

Définition : Un séparateur de sommets $s - t$ noté (S, C, T) est une partition de V telle que $s \in S$ et $t \in T$. Alors, la coupe $s-t$ pour nous est une division des sommets du graphe en trois sous-ensembles indépendants S, C et T, avec la source s dans le sous-ensemble S, le terminal t dans T et le sous-ensemble C représente la coupe. La coupe C est minimale, cela signifie que le nombre de sommets existant dans C est minimal, c'est-à-dire que si l'on enlève un seul sommet de C le reste n'est plus suffisant pour une coupe.

Pour un seul graphe orienté, nous pouvons trouver plusieurs coupes minimales entre la source et le terminal. Dans la suite de cet article, nous proposons notre algorithme quantique pour déterminer toutes les coupes minimales d'un graphe orienté.

3 Algorithme

Dans cette section, on va décrire notre algorithme quantique pour trouver toutes les coupes minimales qui peuvent arrêter le flux entre la source et le terminale d'un graphe orienté. Cet algorithme est basé sur une stratégie de mouvement où il utilise des oracles de mouvements pour construire une superposition quantique contient toutes ces coupes minimales.

La première question qui arrive ici est de savoir comment représenter tous les ensembles de sommets avec des qubits quantiques ? D'autre part, pour un graphe de n sommets, nous trouverons 2^n sous ensemble différents de ces sommets. Dans le cadre quantique, avec n qubits, nous pouvons représenter 2^n états possibles (nous citons ce livre pour les bases de l'informatique quantique (Le Bellac, 2005)). Dans les deux cas avec n éléments on a 2^n possibilités, donc on représente les sous ensembles par les états quantiques de ces n qubits. Pour ce faire, on utilise pour chaque sommet du graphe un qubit, et chaque état de ces qubits représente un sous ensemble des sommets.

3.1 Oracle de mouvement

Avant de commencer à expliquer le fonctionnement de notre algorithme, nous commençons par la définition d'un mouvement est comment ces mouvements sont appliqués par des circuits quantiques.

Définition : Le mouvement d'un sommet v c'est le passage de v vers ses successeurs $Succ(v)$.

$$Mov(v) = Succ(v)$$

Le mouvement d'un sommet $v \in \theta$ c'est le remplacement de v par ses successeurs $Succ(v)$ dans l'ensemble θ.

$$Mov_\theta(v) = \{S \setminus v\} \cup Succ(v)$$

Dans le contexte quantique, pour appliquer ces mouvements, nous utilisons des oracles quantiques appelés oracles de mouvements. Chaque sommet à un oracle de mouvement, qui permet d'appliquer le mouvement si le sommet de mouvement existe dans le sous ensemble d'entrer et fournir en sortie l'entrer plus les sous ensemble après le mouvements.

Pour un sommet v et un certain sous-ensemble de sommets θ, tel que $v \in \theta$. on suppose que le sous-ensemble θ est représenter par l'état quantique $|\psi_\theta\rangle$. Pour appliquer le mouvement de v, nous donnons l'état quantique $|\psi_\theta\rangle$ à l'oracle comme entrée, et dans la sortie de l'oracle nous trouverons une superposition contient deux états $|\psi_\theta\rangle$ et l'état $|\psi_{\theta'}\rangle$, avec $\theta' = (\theta \backslash v) \cup Succ(v)$.

Plus générale, supposons que l'ensemble d'entrÃ©e est l'union de deux sous-ensembles $\theta = \theta_1 \cup \theta_2$ représentée par la superposition quantique $|\psi_\theta\rangle = \alpha_1 |\theta_1\rangle + \alpha_2 |\theta_2\rangle$ où l'état $|\theta_1\rangle$ représente le sous-ensemble θ_1 et l'état $|\theta_2\rangle$ représente le sous-ensemble θ_2. Soit un sommet $w \in \theta_1$ et $w \notin \theta_2$. La sortie de l'oracle de mouvement de $w \in \theta$ est $|\psi_{out}\rangle = \frac{\alpha_1}{\sqrt{2}} |\theta_1\rangle + \alpha_2 |\theta_2\rangle + \frac{\alpha_1}{\sqrt{2}} |\theta_3\rangle$, où l'état $|\theta_3\rangle$ représente le sous ensemble $\theta_3 = Mov_{\theta_1}(w) = \{\theta_1 \setminus w\} \cup Succ(w)$.

Afin de donner une formule générale d'un oracle de mouvement, prenons v un sommet, O_v l'oracle de mouvement de v et $|\psi_S\rangle$ est une superposition quantique représente un certain nombre des sous-ensembles de sommets. La formule générale de l'oracle est :

$$|\psi'_\theta\rangle = O_v |\psi_\theta\rangle \text{ et } |\psi'_\theta\rangle = \sum_{e \in \theta} \alpha_e f_v(e) |e\rangle$$

$$\text{avec f(e)} = \begin{cases} 1 & \text{si } e \text{ est l'état de départ} \\ 1 & \text{si } e \text{ est l'état de mouvement} \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad \begin{cases} \alpha_e = 0 & \text{si } f(e) = 0 \\ \sum_e \alpha_e = 1 \end{cases}$$

Pour représenter un oracle avec un circuit quantique simple, nous devons ajouter deux qubits de contrôle supplémentaires $|c_0\rangle$ et $|c_1\rangle$. Le qubit $|c_0\rangle$ est utilisé pour vérifier si le sommet du mouvement existe dans l'ensemble d'entrée et il sera dans l'état $|1\rangle$, si le qubit correspondant au sommet du mouvement est dans l'état $|1\rangle$ (existe dans l'ensemble d'entrée), et dans $|0\rangle$ sinon. Pour cela, on utilise la porte C-X avec le qubit correspondant au sommet du mouvement comme contrôle et le qubit $|c_0\rangle$ comme cible. Si le sommet est dans l'ensemble d'entrée, nous ajoutons un autre ensemble à l'ensemble des coupes. Ou, nous pouvons dire que si le qubit du sommet est dans l'état $|1\rangle$, nous ajoutons un nouvel état à la superposition d'entrée. Pour ce faire, nous utilisons la porte C-H avec le qubit $|c_0\rangle$ comme contrôle et le qubit du sommet du mouvement comme cible. Ensuite, on utilise une porte C-X pour appliquer le mouvement sur le nouvel ensemble.

Pour ajouter tous les successeurs du sommet de mouvement au nouvel ensemble, nous utilisons le circuit $movP$ du figure 1 qui permet d'inverser le qubit du successeur dans l'état $|1\rangle$ s'il est dans l'état $|0\rangle$ et de ne rien faire s'il est dans l'état $|1\rangle$.

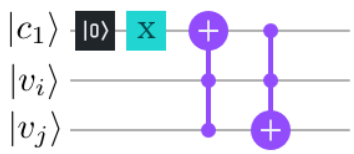


FIG. 1: Le circuit du mouvement de v_i vers v_j, ce circuit utilise trois qubits : $|v_i\rangle$, $|v_j\rangle$ et $|c_1\rangle$ pour le contrôle.

3.2 Description de l'algorithme

Soit $G = (V, E)$ un graphe orienté, où V est l'ensemble des sommets tels que $|V| = n$ et E l'ensemble des arêtes, la source du graphe est le sommet S et le terminal est le sommet T. La première étape de l'algorithme consiste à préparer le nombre nécessaire de qubits pour représenter tous les sous-ensembles possibles de sommets du graphe. le graphe à n sommets, nous utilisons donc n quits.

Au départ, tous les n qubits sont initialisés dans l'état $|0\rangle$, donc l'état quantique est initialisé à $|\psi\rangle = |0 \dots 0\rangle$. Avec le premier qubit représente la source S, le deuxième qubit pour le deuxième sommet, jusqu'au dernier qubit pour le dernier sommet (le terminal T). Il n'y a que des zéros dans l'état $|\psi\rangle$, ce qui signifie que l'ensemble représenté par $|\psi\rangle$ ne contient aucune sommet.

$$|\psi\rangle = |T v_n \dots v_1 S\rangle = |00 \dots 00\rangle \iff \psi = \{\}$$

Afin de commencer avec un ensemble ψ contenant uniquement le sommet d'entrée S, on applique l'opération `not` sur le premier qubit, ce qui nous donne comme résultat $|\psi\rangle = |0 \dots 01\rangle$, le qubit $|s\rangle$ étant dans l'état $|1\rangle$.

Dans la deuxième étape, pour chaque sommet, on a un oracle de mouvement, on appelle donc tous ces oracles de mouvement.

Le premier oracle correspondant au mouvement du sommet S vers ses successeurs :

$$|\psi_1\rangle = O_1 |\psi\rangle = \alpha_1 |\psi\rangle + \alpha_2 |Mov_\psi(S)\rangle$$
$$|\psi_1\rangle = \alpha_1 |\psi\rangle + \alpha_2 |Succ(S)\rangle$$

Après cette itération, l'oracle O_1 ajoute à la superposition l'état $|Succ(S)\rangle$ qui représente la première coupe du graphe.

Après cela, nous appliquons tous les oracles restants :

$$|\psi_{fin}\rangle = O_n O_{n-1} \dots O_2 |\psi\rangle$$

Chacun de ces oracles ajoute un certain nombre d'états à la superposition, ce qui signifie qu'il ajoute un certain nombre de coupes à l'ensemble des coupes représentées par la superposition.

$$|\psi_{fin}\rangle = \alpha_1 |cut_1\rangle + \dots + \alpha_k |cut_k\rangle$$

Après les n oracles, dans la superposition de sortie $|\psi_{fin}\rangle = \sum_i \alpha_i |Mincut_i\rangle$, on trouve toutes les coupes minimales possibles représentées par les états $|Mincut_i\rangle$.

Enfin, nous utilisons un filtre classique simple pour éliminer les coupes non minimales.

L'algorithme 1 représente les étapes pour générer le circuit quantique pour trouver toutes les coupes minimales possibles d'un graphe orienté.

Algorithme 1 : Toutes les coupes minimales d'un graphe orienté

Entrées : Graphe $G = (V, E)$, avec $n = |V|$ est le nombre de sommets du graphe, la source S, le terminal T

Résultats : L'ensemble des coupes minimales Cs

Début :

Réserver n qubits,

$$|\psi_0\rangle = |0, \dots, 0\rangle$$

Appliquer la porte X sur le premier qubit qui représente la source S.

$$|\psi_1\rangle = |0, \dots, 0, 1\rangle$$

Appliquer l'oracle O_s pour faire le mouvement de S

$$|\psi_2\rangle = O_s |\psi_1\rangle$$

for *chaque* $v \in V$ *et* v *!=* S *et* $T \notin Succ(v)$ **do**

$$|\psi_{i+1}\rangle = O_v |\psi_i\rangle$$

$Cs =$ mesurer $|\psi_{n-1}\rangle$ et éliminer les coupes non minimales.

return Cs

3.3 Analyse de la complexité

Supposons que nous avons un graphe $G = (V, E)$, avec V l'ensemble des sommets et E l'ensemble des arêtes telles que $|V| = n$ et $|E| = m$. Pour construire le circuit, nous avons

besoin de n qubits pour représenter tous les états possibles du graphe et 2 qubits auxiliaires pour le contrôle. Pour n oracles de mouvement nous avons besoin de n portes C-H, $2n-2$ portes C-X, $m+2$ portes X et $2m$ portes CC-X. On peut donc dire que notre algorithme a une complexité linéaire.

4 Étude de cas

Dans cette section, nous présentons la version détaillée d'une étude de cas de notre algorithme. Pour cela, prenons le graphe orienté $G = (V, E)$ représenté dans la figure 2, où V est l'ensemble des sommets de taille 9 ($n = |V| = 9$), qui est étiqueté de v_0 à v_8 comme suit : $V = \{v_0, v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$. Et l'ensemble des arêtes entre ces sommets est noté E et présenté comme ceci : $E = \{(v_0, v_1), (v_0, v_2), (v_1, v_5), (v_1, v_7), (v_2, v_3), (v_2, v_4), (v_3, v_5), (v_4, v_6), (v_4, v_8), (v_5, v_6), (v_6, v_7), (v_7, v_8)\}$. Nous avons également fixé la source S dans le sommet v_0 et le terminal T dans le sommet v_8.

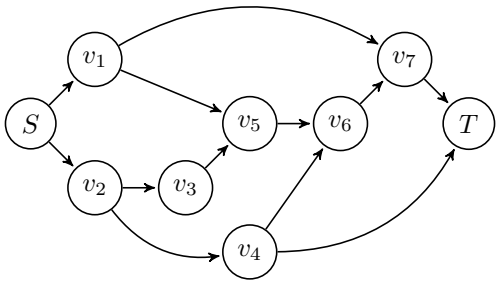


FIG. 2: Graphe orienté de 9 sommets

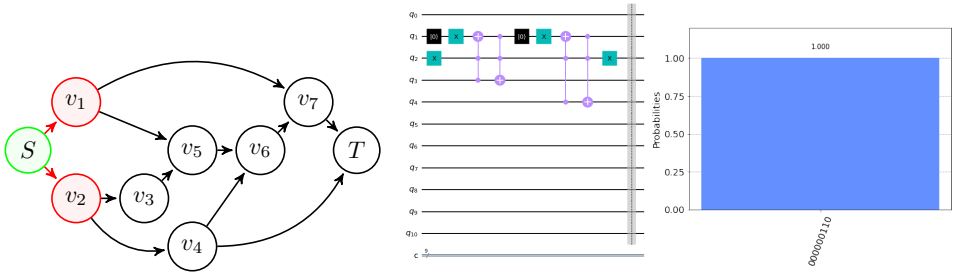
Visuellement, nous pouvons identifier l'ensemble des séparateurs de sommets définissant les coupes minimales : $Cs = \{C_1 = \{v_1, v_2\}, C_2 = \{v_2, v_7\}, C_3 = \{v_4, v_7\}, C_4 = \{v_1, v_4, v_5\}, C_5 = \{v_1, v_3, v_4\}, C_6 = \{v_1, v_4, v_6\}\}$, chaque sous-ensemble $C_i, i = 0, \ldots, 6$ étant un séparateur de sommets du graphe avec un nombre minimal de sommets.

Dans la suite, on cherche à trouver cet ensemble des coupes minimales Cs par notre algorithme quantique. Pour ce faire, on utilise le simulateur quantique d'IBM afin de montrer les résultats intermédiaires de notre algorithme. Dans le graphe G, nous avons $n = 9$ sommets, ce qui nous pousse à utiliser 9 qubits pour représenter toutes les combinaisons possibles de ces sommets. La source S est représentée par le premier qubit $|v_0\rangle$ et chaque sommet $v_i, i = 1, \ldots, 7$ est associé au qubit $|v_i\rangle, i = 1, \ldots, 7$ et le terminal T est associé au dernier qubit $|v_8\rangle$. Ces 9 qubits $|v_i\rangle, i = 0, \ldots, i = 8$ peuvent représenter 2^9 états possibles, donc, ces qubits peuvent générer la superposition $|v_8 v_7 v_6 v_5 v_4 v_3 v_2 v_1 v_0\rangle = \sum_{i=0}^{2^9-1} \alpha_i |i\rangle$, qui représente tous les sous-ensembles possibles de sommets du graphe G. Chaque état $|i\rangle$ dans la superposition $|v_8 v_7 v_6 v_5 v_4 v_3 v_2 v_1 v_0\rangle$ représente un seul séparateur de sommets, on dit que le sommet v_i appartient au séparateur de sommets $|C_i\rangle$ (ou aux coupes minimales $|C_i\rangle$) si le qubit correspondant $|v_i\rangle$ dans l'état $|C_i\rangle$ est dans l'état $|1\rangle$. Par exemple, le séparateur de sommets $C = \{v_1, v_2\}$ peut être codé par l'état quantique $|000000110\rangle$.

Au début de l'algorithme, chaque qubit est initialisé dans l'état $|0\rangle$, ainsi, le registre quantique est initialisé dans l'état $|\psi\rangle = |00000000000\rangle$. Pour commencer avec la source du graphe, nous voulons générer l'état $|\psi_0\rangle = |000000001\rangle$ (qui représente la présence de la source dans l'état) à partir de l'état initial $|\psi\rangle$. Pour cela, on applique la porte not dans le qubit $|v_0\rangle$.

Supposons maintenant que tous les successeurs de la source $S = v_0$ soient tombés en panne, alors il n'y a pas d'autre moyen d'aller aux sommets suivants, donc le sous-ensemble des successeurs du sommet $S = v_0$ est un séparateur de sommets, en plus de cela, si l'un de ces successeurs est en bon état, nous trouverons un moyen d'aller aux sommets suivants. Par conséquent, le sous-ensemble des successeurs de $S = v_0$ est un séparateur de sommets avec un nombre minimal de sommets, autrement dit une coupe minimale. Ensuite, pour trouver cette première coupe minimale, on applique le premier oracle du mouvement O_S sur l'état $|\psi_0\rangle$. C'est-à-dire qu'on prend $|\psi_0\rangle$ comme état d'entrée et on applique le mouvement du sommet $v_0 = S$ à ses successeurs v_1 et v_2, ce qui nous donne l'état $|\psi_1\rangle = |000000110\rangle$ en sortie. Le graphe du mouvement de l'oracle S et le résultat en sortie sont représentés dans la figure 3.

Le premier oracle génère le premier état, qui représente une coupe minimale. Ici l'état $|\psi_0\rangle$ n'est pas une coupe, alors nous utilisons seulement la porte not pour transformer cet état $|\psi_0\rangle$ à l'état $|\psi_1\rangle$ qui représente la première coupe minimale.



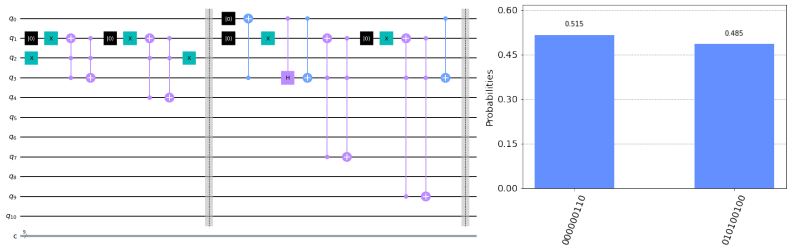
(a) (b) (c)

FIG. 3: Le mouvement de S vers les deux successeurs v_1 et v_2 représentés en (a) peut être généré par l'oracle (b). (c) Le résultat de l'exécution donne l'état $|000000110\rangle$ qui représente la première coupe $\{v_1, v_2\}$.

Maintenant, supposons qu'un des successeurs v_1 et v_2 soit dans un bon état, alors il y a un moyen d'aller aux sommets suivants. Par exemple, si le sommet v_1 est dans un bon état nous pouvons trouver un chemin vers le terminal à travers les successeurs de v_1. Si ces successeurs de v_1 sont en panne, on ne trouve pas un chemin vers le terminal. Donc, le sous-ensemble $Succ(v_1) \cup C_1 \setminus \{v_1\}$ est une coupe. Donc, si on applique le mouvement de v_1 dans l'état $|\psi_1\rangle$, on trouve une nouvelle coupe minimale contenant les successeurs de v_1 et le sommet v_2.

Ici, l'oracle utilise la porte Hadamard pour conserver la première coupe et ajouter le nouvel état correspondant à la nouvelle coupe. La figure 4 présente le circuit avec le second oracle et les résultats d'exécution dans le simulateur.mouvements

A l'étape k, on applique l'oracle O_k sur la superposition de sortie de l'étape $k-1$, ainsi, on applique le déplacement du sommet v_k correspondant à l'oracle O_k, qui ajoute de nouveaux



(a) (b)

FIG. 4: (a) Le deuxième oracle commence après la première barrière dans le circuit. (b) Le résultat de l'exécution donne deux états : $|000000110\rangle$ représente l'entrée et $|010100100\rangle$ représente la nouvelle coupe aprés le mouvement. Notez que le second oracle dans (a) utilise la porte de Hadamard pour ajouter le nouvel état dans la superposition.

états dans la superposition $|\psi_k\rangle$, si le qubit correspondant au sommet v_k est dans l'état $|1\rangle$ pour chaque état de la superposition $|\psi_{k-1}\rangle$.

$$
\begin{aligned}
|\psi_k\rangle &= O_k \, |\psi_{k-1}\rangle \\
|\psi_k\rangle &= O_k \sum_j \alpha_j \, |c_j\rangle = \sum_j \alpha_j O_k \, |c_j\rangle \\
|\psi_k\rangle &= \sum_j \beta_j \, |c_j\rangle + \sum_j \beta_j Mov_{v_k}(|c_j\rangle)
\end{aligned}
$$

Après tous les mouvements possibles, nous avons trouvé la superposition $|\psi_{final}\rangle$:

$$
|\psi_{final}\rangle = \sum_i \alpha_i \, |v_i\rangle = \sum_i \alpha_i \, |v_{i_8} v_{i_7} v_{i_6} v_{i_5} v_{i_4} v_{i_3} v_{i_2} v_{i_1} v_{i_0}\rangle
$$

où $\sum_i \alpha_i = 1$ et chaque état $|i\rangle = |v_{i8} v_{i7} v_{i6} v_{i5} v_{i4} v_{i3} v_{i2} v_{i1} v_{i0} c_{i_1} c_{i_0}\rangle$ de l'état $|\psi_{final}\rangle$ représente une coupe C_i, avec $|v_{ij}\rangle = |1\rangle$ si le sommet j est dans la coupe C_i et $|v_{ij}\rangle = |0\rangle$ dans le cas contraire. Dans notre exemple de graphe, après l'exécution du circuit 5 dans le simulateur d'IBM et l'ordinateur quantique IBM Q 16 Melbourne nous présentons les résultats dans le figure 6.

Enfin, nous éliminons les coupes non minimales. Pour cela, pour chaque (i, j) élimine la coupe C_j si $C_i \in C_j$.

Pour vérifier les résultats nous avons visualisé chaque état de la superposition 7 dans un graphe indépendant dans le figure 2, avec la couleur rouge si le qubit de sommet dans l'état $|1\rangle$ (présent dans la coupe minimale) et noire s'il est dans l'état $|0\rangle$.

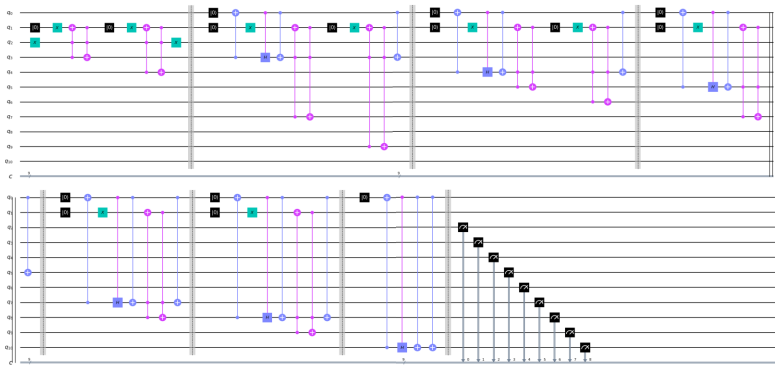


FIG. 5: Le circuit réserve 11 qubits : 9 pour représenter tous les états de défaillance possibles, 2 pour le contrôle. Et aussi il utilise 7 oracles de mouvements séparés par des séparateurs verticaux. Chaque oracle représente le mouvement d'un sommet. À la fin du circuit, on mesure les 9 qubits pour trouver la superposition qui représente toutes les coupes minimales.

5 Conclusion

Nous avons proposé un algorithme quantique pour trouver toutes les coupes minimales d'un graphe orienté. Plus précisément, nous proposons un algorithme quantique qui utilise des oracles de mouvements pour générer en sortie une superposition de toutes les états qui représentent les coupes minimales. Dans cet article, les coupes sont représentées par un ensemble de sommets, qui peuvent séparé la source et le terminal du graphe, et elles sont minimales si elles contiennent juste le nombre minimal de sommets pour être une coupe. Aussi, cet algorithme a une complexité linéaire, car : il utilise seulement $n + 2$ qubits, n pour représenter toutes les combinaisons possibles de sommets et 2 pour le contrôle, et il utilise n oracles de mouvements, n étant le nombre de sommets du graphe.

Références

Ambainis, A. (2007). Quantum walk algorithm for element distinctness. *SIAM Journal on Computing 37*(1), 210–239.

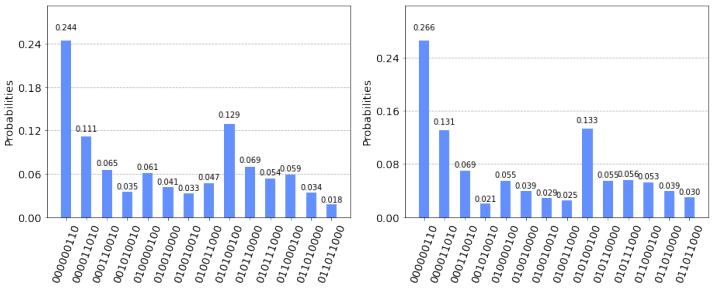
Ashcraft, C. et J. W. Liu (1994). A partition improvement algorithm for generalized nested dissection. *Boeing Computer Services, Seattle, WA, Tech. Rep. BCSTECH-94-020*.

Bui, T. N. et C. Jones (1992). Finding good approximate vertex and edge partitions is np-hard. *Information Processing Letters 42*(3), 153–159.

Cui, S. X., M. H. Freedman, O. Sattath, R. Stong, et G. Minton (2016). Quantum max-flow/min-cut. *Journal of Mathematical Physics 57*(6), 062206.

Davis, T. A., W. W. Hager, S. P. Kolodziej, et S. N. Yeralan (2019). Algorithm xxx : Mongoose, a graph coarsening and partitioning library. *ACM Trans. Math. Software*.

Dörn, S. (2007). Quantum algorithms for graph traversals and related problems. In *Proceedings of CIE*, Volume 7, pp. 123–131. Citeseer.



(a) (b)

FIG. 6: Les histogrammes représente la superposition de sortie $|\psi_{final}\rangle$. (a) Le résultat de l'exécution dans le simulateur Qasm d'IBM. (b) Le résultat de l'exécution dans l'ordinateur quantique IBM Q 16 Melbourne

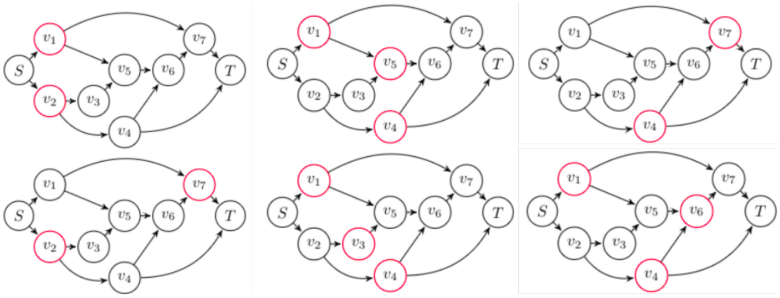


FIG. 7: Résultats

Fiduccia, C. M. et R. M. Mattheyses (1982). A linear-time heuristic for improving network partitions. In *19th design automation conference*, pp. 175–181. IEEE.

Grover, L. K. (1996). A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 212–219.

Hager, W. W. et J. T. Hungerford (2015). Continuous quadratic programming formulations of optimization problems on graphs. *European Journal of Operational Research 240*(2), 328–337.

Hager, W. W., J. T. Hungerford, et I. Safro (2018). A multilevel bilinear programming algorithm for the vertex separator problem. *Computational Optimization and Applications 69*(1), 189–223.

Hendrickson, B. et E. Rothberg (1998). Improving the run time and quality of nested dissection ordering. *SIAM Journal on Scientific Computing 20*(2), 468–489.

Jeffery, S. et S. Kimmel (2017). Quantum algorithms for graph connectivity and formula evaluation. *Quantum 1*, 26.

Kernighan, B. W. et S. Lin (1970). An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal 49*(2), 291–307.

Kloks, T. et D. Kratsch (1998). Listing all minimal separators of a graph. *SIAM Journal on Computing 27*(3), 605–613.

Kolodziej, S. et T. Davis (2016). Vertex separators with mixed-integer linear optimization. *17th SIAM Conference on Parallel Processing for Scientific Computing*.

Kolodziej, S. P. (2019). *Computational Optimization Techniques for Graph Partitioning*. Ph. D. thesis.

Kolodziej, S. P. et T. A. Davis (2020). Generalized gains for hybrid vertex separator algorithms. In *2020 Proceedings of the SIAM Workshop on Combinatorial Scientific Computing*, pp. 96–105. SIAM.

Le Bellac, M. (2005). *Introduction à l'information quantique*. Belin.

Lipton, R. J. et R. E. Tarjan (1979). A separator theorem for planar graphs. *SIAM Journal on Applied Mathematics 36*(2), 177–189.

Lipton, R. J. et R. E. Tarjan (1980). Applications of a planar separator theorem. *SIAM journal on computing 9*(3), 615–627.

Magniez, F., M. Santha, et M. Szegedy (2007). Quantum algorithms for the triangle problem. *SIAM Journal on Computing 37*(2), 413–424.

Nielsen, M. A. et I. Chuang (2002). Quantum computation and quantum information.

Pednault, E., J. A. Gunnels, G. Nannicini, L. Horesh, et R. Wisnieff (2019). Leveraging secondary storage to simulate deep 54-qubit sycamore circuits. *arXiv preprint arXiv :1910.09534*.

Preskill, J. (2012). Quantum computing and the entanglement frontier. *arXiv preprint arXiv :1203.5813*.

Shimizu, K. et R. Mori (2021). Exponential-time quantum algorithms for graph coloring problems. In *Latin American Symposium on Theoretical Informatics*, pp. 387–398. Springer.

Shor, P. W. (1999). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review 41*(2), 303–332.

Zhong, H.-S., H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, et al. (2020). Quantum computational advantage using photons. *Science 370*(6523), 1460–1463.

Summary

The Vertex Separator Problem of a directed graph consists in finding all combinations of vertices which can be disconnected the source and the terminal of the graph, these combinations are minimal if they contain only the minimal number of vertices. In this paper, we introduce a new quantum algorithm based on a movement strategy to find these separators in a quantum superposition with linear complexity.

Keywords: Quantum Computation, Vertex separators, Minimal cuts

Les tweets vocaux entre humanisation et modération: conséquences, défis et opportunités

Didier Henry

LAMIA, Université des Antilles, Guadeloupe
didier.henry@univ-antilles.fr

Résumé. De nos jours, Twitter est largement utilisé pour accéder ou partager des informations liées à une grande variété de sujets. Ces dernières années, des chercheurs ont montré que les données Twitter sont très utiles dans plusieurs domaines et peuvent avoir de nombreuses applications. En juin 2020, Twitter a annoncé les tweets vocaux pour créer une expérience plus humaine. Cet article donne un aperçu des conséquences possibles de cette nouvelle fonctionnalité. De plus, nous présentons certains défis auxquels les chercheurs seront confrontés pour exploiter ce type de données. Enfin, nous présentons également plusieurs opportunités de recherches liées aux tweets vocaux.

Mots-clés : réseaux sociaux, collecte de données, analyse de données, recherche sur Twitter

1 Introduction

En quelques années, Twitter a transformé la façon dont nous créons, partageons et accédons aux informations. Cette plateforme s'appuie sur un vaste réseau permettant le libre échange d'informations entre des centaines de millions de personnes (célébrités, organisations, particuliers, syndicats, etc.) à travers le monde instantanément. Chaque jour, les gens diffusent des centaines de millions de messages sur une grande variété de sujets (politique, sport, santé, actualités, technologie, etc.). Qu'ils soient liés à un événement global, c'est-à-dire spécifique à l'ensemble des individus, ou à un événement local, c'est-à-dire spécifique à un individu, ces messages peuvent influencer une société et peuvent contenir des informations utiles. En effet, ces dernières années, Twitter a fait l'objet de nombreuses études. Par exemple, Asur et Huberman (2010) ont introduit une méthode basée sur des messages diffusés pour prédire les revenus générés par les films. De même, Bollen et al. (2011) ont utilisé des messages partagés sur Twitter pour prédire l'évolution des marchés boursiers. Dans leurs travaux, Tumasjan et al. (2010) ont analysé les messages diffusés sur les réseaux sociaux Twitter pour prédire les résultats d'une élection politique. Ainsi, les messages publiés sur Twitter peuvent fournir des informations utiles pour prévoir ou détecter des événements et sauver des vies. Le 17 juin 2020, Twitter a lancé les tweets vocaux visant à créer une expérience plus humaine [1]. En effet, 280 caractères peuvent être insuffisants pour exprimer fidèlement les émotions ou le ton

1. https://blog.twitter.com/en_us/topics/product/2020/your-tweet-your-voice.html

même s'ils peuvent être accompagnés de photos, vidéos, gifs ou émoticônes. Malgré les efforts pour améliorer l'expérience utilisateur, cette nouvelle fonctionnalité pourrait avoir des conséquences négatives. De plus, l'utilisation de tweets vocaux apporte de nouveaux défis aux chercheurs et offre de nouvelles opportunités ou applications. Dans cet article, nous présentons un ensemble d'aspects positifs et négatifs des tweets vocaux. Ensuite, nous introduisons plusieurs problèmes auxquels les chercheurs peuvent être confrontés pour exploiter ce nouveau type de données. Enfin, nous proposons plusieurs pistes pour les futures recherches.

2 Conséquences

Quatorze ans après sa création, Twitter a annoncé le déploiement de messages vocaux. En effet, depuis le 17 juin 2020, un nombre limité d'utilisateurs peut enregistrer jusqu'à deux minutes et vingt secondes de son pour chaque tweet posté. Si l'audio enregistré dure plus longtemps, un nouveau tweet sera publié. La durée de 140 secondes faisant référence aux 140 caractères qui étaient la limite initiale des messages écrits, avant qu'elle ne passe à 280, en novembre 2017. De plus, la voix ne peut être utilisée que pour les tweets originaux, c'est-à-dire que les utilisateurs ne peuvent pas l'inclure dans des réponses ou des retweets avec un commentaire. Un autre point à noter est que, quelle que soit l'image de profil de l'utilisateur, lorsque l'utilisateur enregistre un message audio, elle sera toujours jointe à ce tweet. Cependant, si Twitter semble enthousiasmé par cette nouvelle fonctionnalité qui pourrait avoir des effets positifs sur l'expérience des utilisateurs, cette nouveauté peut aussi avoir des répercussions négatives. Dans cette section, nous proposons une liste non exhaustive des effets positifs et négatifs que les tweets vocaux pourraient avoir.

2.1 Effets positifs des tweets vocaux

2.1.1 Être pratique

Premièrement, les tweets vocaux peuvent être utiles, car ils dépassent la limite de 280 caractères, ce qui oblige souvent les utilisateurs à s'exprimer brièvement. En effet, cette nouvelle fonctionnalité pourrait être utilisée afin de partager des informations plus complètement en incluant des détails qu'ils n'auraient pas pu transcrire avec la limite du nombre de caractères. Deuxièmement, l'utilisation de tweets vocaux pourrait considérablement améliorer la vitesse de soumission de nouvelles informations ainsi que leur lecture. En effet, les utilisateurs n'auront plus besoin de composer leurs messages et pourront écouter les messages des autres tout en effectuant une autre tâche simultanément. Ainsi, les tweets vocaux pourraient être très utiles dans les situations d'urgence ou de stress. Troisièmement, ce nouveau type de message pourrait être utile pour les personnes souffrant de problèmes de santé tels que les malvoyants ou les personnes dyslexiques. Cela pourrait également améliorer l'expérience des utilisateurs ayant un handicap physique ou mental. Enfin, on peut supposer que cette nouvelle fonctionnalité pourrait permettre d'économiser de l'énergie notamment pour les appareils portables, car l'utilisation de l'écran pourrait être réduite.

2.1.2 Humanisation

Les tweets peuvent être accompagnés de photos, vidéos, gifs ou émoticônes, ils résultent d'une période de réflexion, car les gens doivent résumer leurs idées ou réflexions en raison de la limite du nombre de caractères du tweet. En d'autres termes, ils ne sont pas spontanés comme peut l'être une conversation entre des personnes. Au contraire, les tweets vocaux seront plus dynamiques et peuvent rendre les informations plus captivantes. De plus, la fonction de tweet vocal peut apporter une touche plus intime aux messages postés. En effet, les utilisateurs pourront entendre la voix de l'auteur du message ce qui le rendra plus personnel. De plus, grâce aux tweets vocaux, les gens pourront partager un large éventail d'émotions, de sorte que les utilisateurs comprendront les sentiments de l'auteur plus précisément et plus rapidement. Ainsi, les gens pourront dire si l'auteur du message est heureux, triste ou même en colère rien qu'en écoutant. De même, les utilisateurs auront accès à une justesse de ton qui manque encore dans les tweets. Par exemple, les messages ironiques pourraient être compris plus rapidement que lorsqu'ils sont écrits.

2.1.3 Diverses applications

Les tweets vocaux peuvent avoir une grande variété d'applications. Par exemple, alors que les tweets écrits sont limités à un seul auteur, les tweets vocaux peuvent contenir une conversation avec plusieurs orateurs. Ainsi, les tweets vocaux offrent plus de liberté quant au nombre de locuteurs dans un même message. De plus, les tweets vocaux peuvent contenir des effets sonores ou des jingles pour mettre en évidence une partie du message ou une idée. Par exemple, les tweets vocaux pourraient être un nouveau moyen de publicité pour annoncer de nouveaux produits ou promotions. De plus, les influenceurs pourraient utiliser cette nouvelle fonctionnalité pour promouvoir une marque et gagner plus d'audience, car le contenu du tweet ne s'affiche pas. Une autre application possible serait d'utiliser des tweets vocaux pour susciter l'intérêt des utilisateurs en proposant des annonces audio de contenus qu'il soit lié à des séries, des films, des jeux vidéo ou de la musique. En outre, certains utilisateurs peuvent utiliser des tweets vocaux pour se divertir, par exemple en suggérant à d'autres utilisateurs de deviner le nom du contenu ou son origine.

2.2 Répercussions négatives des tweets vocaux

Bien que les tweets vocaux puissent avoir des effets positifs lorsqu'ils sont utilisés de manière bienveillante, ils peuvent également être nocifs s'ils sont publiés par des personnes malveillantes. Outre le fait que cette nouvelle fonctionnalité ne pourra pas être utilisée par des personnes sourdes ou malentendantes, elles peuvent avoir des répercussions inquiétantes.

2.2.1 Harcèlement et contenu répréhensible

Depuis plusieurs années, Twitter a des problèmes de modération en matière de harcèlement (Guberman et Hemphill (2017)). En effet, certains utilisateurs se sont suicidés ou ont tenté de se suicider [2] en raison du harcèlement dont ils ont souffert sur la plate-forme. Les tweets vocaux peuvent faciliter le harcèlement des utilisateurs et aggraver l'état émotionnel des victimes.

2. https://www.fastcompany.com/40547818/did-we-create-this-monster-how-twitter-turned-toxic

En effet, contrairement aux tweets écrits, les tweets vocaux permettent des cris et des bruits ou des sons. Cette nouvelle fonctionnalité pourrait apporter un travail supplémentaire aux modérateurs de Twitter car contrairement aux photos ou aux textes, ils les obligent à écouter tout ce qui est dit pour passer à l'action. Habituellement, les modérateurs n'ont que quelques instants pour déterminer si un élément de contenu enfreint les règles de la plate-forme. Avec les tweets vocaux, cela peut être une tâche longue et fastidieuse. Les tweets vocaux pourraient être détournés pour diffuser des contenus interdits au jeune public, des contenus violents ou choquants. En effet, des utilisateurs malveillants ou inconscients pourraient enregistrer des audios contenant de la pornographie ou des séquences violentes. De tels tweets pourraient heurter la sensibilité d'un jeune public ou provoquer de mauvaises surprises sur un lieu de travail ou dans un lieu public. Enfin, les tweets vocaux utilisés à des fins de canular ou non pourraient contenir des sons indésirables ou stridents de manière temporaire ou prolongée qui auraient pour conséquence de nuire à l'audition des utilisateurs.

2.2.2 Messages illégaux, cryptés ou spam

Les tweets vocaux pourraient contenir du contenu illégal. Par exemple, les utilisateurs pourraient enregistrer puis publier des oeuvres musicales complètes, partielles ou parodiques qui constitueraient une violation du droit d'auteur. De même, certains utilisateurs pourraient enregistrer d'autres tweets vocaux, puis les publier sur leur propre compte, ce qui constituerait également une violation du droit d'auteur. En outre, les tweets vocaux peuvent être utilisés pour publier des messages haineux ou racistes, par exemple en utilisant des subtilités de langage, des accents ou des sons fortement connotés. De plus, que ce soit pour le plaisir ou à des fins sérieuses, les utilisateurs peuvent publier des audios contenant des informations cryptées. D'une part, cela pourrait s'avérer problématique pour les modérateurs ou les algorithmes de détection de contenu suspect. D'un autre côté, ce type de tweets vocaux serait potentiellement dangereux s'il était utilisé par des groupes d'individus malveillants. De plus, comme les tweets écrits, les tweets vocaux peuvent être utilisés par des groupes ou des organisations pour spammer les utilisateurs. Premièrement, ce type de contenu pourrait être utilisé pour dénigrer ou discriminer un individu ou un groupe d'individus et aussi lors d'une élection politique pour influencer les votes. Deuxièmement, l'expérience utilisateur peut être affectée, car les personnes peuvent être ennuyées en raison d'informations vocales répétitives et indésirables.

2.2.3 Désinformation, fake news, deepfake

La désinformation est sur Twitter depuis des années (Ghenai et Mejova (2017)). L'utilisation de tweets vocaux pourrait accentuer ce phénomène. En effet, ce nouveau moyen d'expression pourrait contenir un extrait d'informations réelles, qui sortit de leurs contextes pourrait être considérée comme vraie. En outre, les tweets vocaux contenant des informations satiriques ou parodiques pourraient se transformer en désinformation s'ils sont pris au sérieux par des personnes étourdies ou qui ne savent pas qu'il s'agit de fausses informations. De même, les tweets vocaux pourraient être à l'origine de fausses nouvelles. Par exemple, un journaliste ou une chaîne d'information officielle pourrait accidentellement publier un fichier audio contenant une erreur de prononciation, une erreur de numéro, de nom ou de lieu. De plus, une erreur de traduction ou d'interprétation pourrait survenir lors de la diffusion d'informations par un média d'un pays étranger ou de simples utilisateurs. De plus, certaines personnes peuvent uti-

liser des tweets vocaux pour publier des canulars plus crédibles qui pourraient avoir un public plus large que les tweets écrits. Ainsi, on peut supposer que cette nouvelle fonctionnalité sera fortement utilisée pendant le poisson d'avril.

3 Challenges

Malgré le fait que les tweets vocaux semblent être une fonctionnalité intéressante qui peut dans certains cas enrichir l'expérience utilisateur, cette nouveauté peut également avoir des conséquences inquiétantes. Avant de proposer des solutions efficaces pour détecter le détournement ou l'utilisation malveillante des tweets vocaux, les chercheurs peuvent se heurter à plusieurs problématiques concernant ce nouveau type de données sociales. En effet, les chercheurs dont les travaux reposent sur Twitter ne sont pas habitués à travailler avec des données vocales. Dans cette section, nous décrivons de manière non exhaustive certains défis que les chercheurs devront relever pour exploiter les tweets vocaux.

3.1 Qui ?

Savoir qui est impliqué dans le tweet vocal est l'une des principales questions auxquelles les chercheurs devront répondre avant de commencer des recherches plus spécifiques. En effet, plusieurs axes de recherche nécessitent de connaître la ou les personnes à l'origine de l'information. Par exemple, certains travaux (Bakshy et al. (2011); Castillo et al. (2011)) reposent sur la crédibilité ou l'influence de la source. Un travail préalable sera nécessaire pour savoir si l'audio posté ne contient aucune, une ou plusieurs voix. Le vrai défi sera d'identifier la ou les voix, car l'auteur du tweet n'intervient pas forcément dans l'audio. Bien que cela semble être une tâche non triviale, il serait intéressant de déterminer les relations entre les différents interlocuteurs. Un autre défi à relever serait de classer les voix selon le sexe des individus. Enfin, s'il y a plusieurs voix, un travail important sera de les taguer tout au long de la conversation pour savoir qui dit quoi.

3.2 Quoi ?

Déterminer précisément ce qu'il y a dans un tweet vocal sera également un défi. C'est une question importante à laquelle il faut répondre, car plusieurs études (Jansen et al. (2009); Adamic et al. (2016)) ont montré que le contenu du message joue un rôle dans sa diffusion. Dans un premier temps, il s'agira de déterminer la nature du contenu de l'audio. En effet, le tweet vocal ne contient pas forcément une ou plusieurs voix, il peut aussi contenir des sons, une mélodie, du bruit, des effets sonores, etc. Deuxièmement, si le tweet vocal contient des voix, il serait fondamental d'identifier ce que ces voix disent. Une première étape devrait être de déterminer la ou les langues dans lesquelles les voix sont exprimées. Dans un second temps, le contenu du message sera extrait et devra être traduit si nécessaire. Dans un troisième temps, le sujet du tweet pourrait être déterminé à l'aide de différentes techniques d'analyse de texte. Il serait intéressant dans une autre étape d'évaluer le niveau sonore des voix, c'est-à-dire d'identifier s'il s'agit d'un chuchotement, d'un hurlement ou d'une voix normale. Que le tweet vocal contienne ou non des voix, il semble important de fournir une contextualisation de tous les

artefacts qu'il contient. Enfin, après avoir identifié le contenu du tweet, il serait intéressant de le situer dans l'espace et dans le temps.

3.3 Quand ?

Bien que les tweets vocaux soient publiés à une date précise, ils peuvent être liés à des événements présents ou relativement passés ou futurs. C'est la raison pour laquelle les chercheurs devraient s'intéresser à cette question. Premièrement, si le tweet ne contient que des voix, le contenu extrait doit déterminer si une date ou une heure est mentionnée. Dans certains cas, indirectement, le contenu du tweet lui-même lui permettrait de se situer dans le temps. Par exemple, si l'utilisateur évoque clairement la prochaine élection présidentielle des Etats-Unis ou si l'auteur publie un extrait audio d'une symphonie de Beethoven. Deuxièmement, si le tweet contient des éléments de nature différente en plus des voix, ceux-ci pourraient compléter l'analyse de datation du contenu audio. Enfin, dans certains cas, la date du contenu du tweet pourra être celle à laquelle il a été posté.

3.4 Où ?

Semblable à la date, l'emplacement contenu dans les informations publiées peut être différent de celui de l'auteur du message audio. L'extraction de ces données à partir de tweets vocaux pourrait être l'un des défis les plus difficiles pour les chercheurs. Premièrement, les informations extraites de la voix si elles sont présentes dans le message pourraient apporter des réponses concernant la localisation mentionnée dans le tweet vocal. Deuxièmement, le bruit ou le son ambiant pourrait également être utile pour identifier où le message a été publié. Par exemple, si un utilisateur enregistre un message vocal alors qu'il assiste à un match de football, le son de fond peut être un indice. Troisièmement, le lieu pourrait être déterminé dans certains cas, avec plus ou moins de précision si le tweet vocal contient la voix de plusieurs personnes clairement identifiées et dont le lien entre les interlocuteurs est établi. Par exemple, si un tweet vocal contient un échange entre deux personnes discutant d'un événement se déroulant actuellement dans une salle de sport et que ces personnes suivent le compte Twitter de cette même salle de sport, il est fort probable que le tweet ait été publié depuis cette salle de sport.

4 Opportunités de recherche

La nouvelle fonctionnalité de tweet vocal devrait améliorer considérablement l'expérience des utilisateurs de Twitter. Cependant, cette nouveauté peut être détournée par des utilisateurs malveillants et conduire à des abus inquiétants. Par conséquent, pour détecter ce genre de tweet vocal et anticiper les situations dramatiques, les chercheurs devraient s'intéresser à ce nouveau type de données. Bien que les chercheurs devront surmonter de nombreux défis pour exploiter les tweets vocaux, cette nouveauté offre des perspectives de recherche nouvelles et intéressantes. Sur la base de l'enquête sur une grande variété d'études précédentes et également de nos expériences de recherche sur Twitter, nous présentons dans cette section plusieurs directions pour les recherches futures reposant sur la fonctionnalité de tweet vocal.

4.1 Classification et recommandation du contenu

Au cours des dix dernières années, plusieurs chercheurs (Yang et al. (2014); Posch et al. (2013)) se sont intéressés à la classification ou à la reconnaissance de catégorie des messages publiés sur les réseaux sociaux. Par exemple, Michelson et Macskassy (2010) ont proposé une méthode pour associer un tweet à une catégorie de Wikipédia. Ainsi, une approche de classification du contenu des tweets contenant des voix pourrait également être basée sur les données de Wikipédia. Par ailleurs, concernant les tweets contenant des sons ou des effets sonores, de nouvelles approches devraient être proposées afin de classer ce nouveau type de données. En effet, nous sommes convaincus que les effets sonores ou les sons contenus dans les messages pourraient également jouer un rôle dans la diffusion de l'information. Le problème de classification des messages est étroitement lié à la recommandation des messages. Ces dernières années, Gong et Zhang (2016) ont utilisé des réseaux de neurones convolutifs pour résoudre le problème de recommandation de hashtag. Par conséquent, une ligne de recherche intéressante serait de proposer un nouveau système de recommandation d'information ou de contenu basé sur les précédents tweets vocaux postés. De telles recommandations pourraient s'avérer plus pertinentes, car les tweets vocaux sont de nature plus intime.

4.2 Extraction de tons, d'émotions et de sentiments

Premièrement, mieux comprendre les émotions exprimées dans les médias sociaux est pertinent pour proposer des modèles de propagation plus réalistes. En effet, les émotions peuvent influer sur la décision de partager des informations (Gruzd (2013)). De plus, les émotions peuvent être contagieuses (Kramer et al. (2014)) et ainsi permettre la diffusion plus large de l'information. Deuxièmement, l'identification des émotions présentes dans les tweets vocaux des utilisateurs pourrait permettre de dresser un état émotionnel général de chaque individu afin de déterminer lesquels sont déprimés ou suicidaires. Ainsi, de tels travaux seraient très utiles pour faire des rapports préventifs et sauver des vies. Enfin, l'extraction des émotions peut être nécessaire pour détecter les situations de stress ou de panique chez les utilisateurs et aider à localiser un événement dramatique ou inquiétant. La polarité des messages diffusés sur les réseaux sociaux a fait l'objet de nombreuses études (Birjali et al. (2017); Hassan et al. (2017)). En effet, les sentiments des utilisateurs peuvent être utilisés par exemple pour prédire l'évolution des marchés financiers (Li et Meesad (2016)) ou aussi dans la sphère politique (Elghazaly et al. (2016)). Extraire la polarité et la subjectivité des tweets vocaux pourrait être utile dans une grande variété de domaines, en particulier dans l'arène du marketing, où les marques ou les entreprises sont souvent intéressées par les avis des utilisateurs sur un produit, ou par le signalement de ceux qui communiquent un défaut de conception ou même ceux qui sont déçus par la qualité d'un service. L'extraction du ton contenu dans les tweets vocaux est également un domaine de recherche à explorer. Bien que ce thème semble relativement récent pour les tweets écrits (Frenda (2017)), il pourrait se développer considérablement avec des contributions concernant les tweets vocaux. Premièrement, il serait intéressant d'identifier les messages qui sont humoristiques, ironiques ou sarcastiques, par exemple pour signaler aux utilisateurs étourdis de ne pas prendre ces informations au sérieux. Deuxièmement, l'extraction du ton du message permettrait également d'identifier plus précisément l'état émotionnel de l'utilisateur et en fonction de cet état d'informer les services compétents pour venir en aide à l'utilisateur.

4.3 Profil psychosociologique

En psychologie, le modèle des « Big Five Traits » (McCrae et John (1992)) est le plus largement utilisé pour représenter la personnalité des individus. Dans ce modèle, la personnalité est définie sur la base de cinq dimensions : l'ouverture à l'expérience, la consciencieusité, l'extraversion, l'agréabilité et le névrosisme. Les chercheurs ont montré les relations entre ces traits et le comportement des internautes en termes de langage utilisé (Schwartz et al. (2013)) ou de notes attribuées aux vidéos (Scott et al. (2016)). Dans leur approche, Guntuku et al. (2017) montrent que l'extraversion des utilisateurs influence les images publiées sur Twitter. De plus, Wald et al. (2012) utilisent les informations trouvées dans les profils d'utilisateurs Twitter et leurs tweets pour identifier les personnes atteintes de psychopathie. Les tweets vocaux pourraient permettre d'extraire des profils psychosociologiques plus précis des utilisateurs car il semblerait que cette nouvelle fonctionnalité apporte une touche plus humaine au message et donc il s'enrichit de nouveaux paramètres comme le ton. Ainsi, les utilisateurs pourraient être décrits par leur style social, leur style émotionnel et leur style de pensée. Un tel profil psychosociologique des utilisateurs pourrait être utilisé dans plusieurs domaines de recherche. Par exemple, cela permettrait d'identifier des personnes susceptibles de diffuser des informations liées à un thème spécifique.

4.4 Détection de harcèlement, de contenu répréhensible et illégal

Différentes formes de harcèlement (Chowdhury et al. (2019)) peuvent se produire sur Twitter et conduire à des situations désastreuses pour les victimes. Les modérateurs semblent déjà avoir du mal à filtrer les tweets écrits. Ainsi, les tweets vocaux peuvent augmenter considérablement leur travail et être difficiles à gérer. En effet, ce nouveau type de média nécessite d'écouter l'intégralité du contenu du message et d'en comprendre la signification pour décider s'il est conforme aux conditions d'utilisation de Twitter. De plus, cette nouvelle fonctionnalité peut aggraver le harcèlement subi par les victimes en raison de la lenteur de la modération d'un tel message. En ce sens, comme la détection du harcèlement verbal semble encore très peu étudiée, nous encourageons fortement les chercheurs à mener des études sur ce sujet. La détection de messages vocaux contenant des propos illicites ou répréhensibles serait également un domaine de recherche à explorer. En effet, il serait intéressant d'apporter de nouvelles contributions sur la détection de messages haineux, racistes (Hasanuzzaman et al. (2017)) ou misogynes, la détection de messages homophobes ou la détection de messages criminels (Granizo et al. (2020)). Certains chercheurs pourraient se concentrer sur la détection des messages vocaux qui pourraient nuire à l'intégrité de l'audition. De plus, des travaux devraient être menés pour la détection et le filtrage des messages vocaux pornographiques, car les mineurs ont accès à Twitter et également parce que l'écoute de ce type de message peut être la cause de malentendus ou de scandales.

4.5 Détection de fake news et de deepfake

De nos jours, les fausses nouvelles sont l'un des principaux fléaux des médias sociaux et Twitter ne fait pas exception. En effet, au sein de ces plateformes sociales, tout le monde peut être source d'information et celle-ci n'est pas systématiquement contrôlée. La diffusion de fausses informations ou de rumeurs peut conduire à des situations stressantes (Mendoza

et al. (2010)) ou à une mauvaise réputation et peut également avoir des répercussions économiques (Matthews (2013)). En conséquence, au cours des dernières années, la détection des fausses nouvelles sur les réseaux sociaux a reçu beaucoup d'attention (Buntain et Golbeck (2017); Atodiresei et al. (2018); Hamdi et al. (2020)). Par exemple, certains chercheurs (Henry et Stattner (2019)) ont proposé des modèles prédictifs pour détecter les fausses nouvelles relatives à l'annonce du décès d'une célébrité. Récemment, lors de la crise sanitaire des coronavirus de l'année 2020, il y a eu un flot de fausses informations (Kouzy et al. (2020); Shahi et al. (2020)). En effet, alors que le monde médical tente de réduire la propagation du virus et de trouver des solutions pour soigner les malades, plusieurs rumeurs et fausses informations se répandent rapidement sur diverses plateformes de médias sociaux. En raison du manque d'informations et de la peur du virus, l'Organisation mondiale de la santé a observé que les gens partagent massivement la désinformation médicale, y compris les remèdes miracles. De plus, de nombreuses théories du complot ont été largement diffusées, en particulier concernant Bill Gates [3]. Les tweets vocaux pourraient devenir un nouveau moyen de diffuser intentionnellement ou non de fausses informations ou des rumeurs. Les utilisateurs pourraient donner une touche émotionnelle à leur message afin de rendre l'information plus crédible et tromper un plus grand nombre de personnes, générant ainsi une diffusion plus large de la fausse information. En ce sens, les chercheurs devraient prendre en compte ce nouveau média dans les futures études menées sur la détection des fausses nouvelles. L'intelligence artificielle est désormais capable de produire un deepfake audio très réaliste en lui fournissant juste quelques mots. De plus, ce type de technologie pourrait rapidement démocratiser (Siarohin et al. (2019)) et donc être utilisé de façon malveillante. Par exemple, l'usurpation de l'identité d'un chef d'État pourrait conduire à un conflit géopolitique ou avoir des conséquences économiques. En ce sens, les chercheurs s'intéressent déjà à la détection de vidéo deepfake (Güera et Delp (2018)). Cependant, des études supplémentaires seront nécessaires pour fournir des systèmes capables de traiter simultanément d'énormes quantités de données afin de pouvoir analyser les flux de données gigantesques des médias sociaux.

5 Conclusion

Dans cet article, nous nous sommes intéressés à la nouvelle fonctionnalité de Twitter qui a été présentée avec enthousiasme : les tweets vocaux. Bien que la fonction de tweet vocal pourrait améliorer l'expérience utilisateur en apportant une touche plus humaine, elle pourrait également être utilisée de façon malveillante et conduire à des situations inquiétantes. D'une part, nous avons montré à quel point les tweets vocaux pourraient s'avérer pratiques à travers les dimensions humaines qui pourraient être incorporées dans ce type de message et certaines applications positives. D'autre part, nous avons présenté quelques abus possibles avec des tweets vocaux, notamment concernant le contenu répréhensible ou illégal. Ensuite, nous avons présenté différents défis que les chercheurs devraient relever dans les futures études sur les tweets vocaux. Enfin, sur la base d'une enquête sur une grande variété d'études antérieures et également de nos expériences de recherche sur Twitter, nous avons fourni plusieurs opportunités de recherche liées aux tweets vocaux.

3. https://www.nytimes.com/2020/04/17/technology/bill-gates-virus-conspiracy-theories.html

Références

Adamic, L. A., T. M. Lento, E. Adar, et P. C. Ng (2016). Information evolution in social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 473–482. ACM.

Asur, S. et B. A. Huberman (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, Volume 1, pp. 492–499. IEEE.

Atodiresei, C.-S., A. Tănăselea, et A. Iftene (2018). Identifying fake news and fake users on twitter. *Procedia Computer Science 126*, 451–461.

Bakshy, E., J. M. Hofman, W. A. Mason, et D. J. Watts (2011). Identifying influencers on twitter. In *Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*.

Birjali, M., A. Beni-Hssane, et M. Erritali (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science 113*, 65–72.

Bollen, J., H. Mao, et X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science 2*(1), 1–8.

Buntain, C. et J. Golbeck (2017). Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 208–215. IEEE.

Castillo, C., M. Mendoza, et B. Poblete (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pp. 675–684. ACM.

Chowdhury, A. G., R. Sawhney, R. Shah, et D. Mahata (2019). # youtoo ? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2527–2537.

Elghazaly, T., A. Mahmoud, et H. A. Hefny (2016). Political sentiment analysis using twitter data. In *Proceedings of the International Conference on Internet of things and Cloud Computing*, pp. 11. ACM.

Frenda, S. (2017). Ironic gestures and tones in twitter. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, Volume 2006, pp. 1–6. CEUR-WS.

Ghenai, A. et Y. Mejova (2017). Catching zika fever : Application of crowdsourcing and machine learning for tracking health misinformation on twitter. *arXiv preprint arXiv :1707.03778*.

Gong, Y. et Q. Zhang (2016). Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*, pp. 2782–2788.

Granizo, S. L., Á. L. V. Caraguay, L. I. B. López, et M. Hernández-Álvarez (2020). Detection of possible illicit messages using natural language processing and computer vision on twitter and linked websites. *IEEE Access 8*, 44534–44546.

Gruzd, A. (2013). Emotions in the twitterverse and implications for user interface design. *AIS Transactions on Human-Computer Interaction 5*(1), 42–56.

Guberman, J. et L. Hemphill (2017). Challenges in modifying existing scales for detecting harassment in individual tweets. In *Proceedings of 50th Annual Hawaii International Confe-*

rence on System Sciences (HICSS).

Güera, D. et E. J. Delp (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6. IEEE.

Guntuku, S. C., W. Lin, J. Carpenter, W. K. Ng, L. H. Ungar, et D. Preoţiuc-Pietro (2017). Studying personality through the content of posted and liked images on twitter. In *Proceedings of the 2017 ACM on web science conference*, pp. 223–227. ACM.

Hamdi, T., H. Slimi, I. Bounhas, et Y. Slimani (2020). A hybrid approach for fake news detection in twitter based on user features and graph embedding. In *International Conference on Distributed Computing and Internet Technology*, pp. 266–280. Springer.

Hasanuzzaman, M., G. Dias, et A. Way (2017). Demographic word embeddings for racism detection on twitter.

Hassan, A. U., J. Hussain, M. Hussain, M. Sadiq, et S. Lee (2017). Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression. In *Information and Communication Technology Convergence (ICTC), 2017 International Conference on*, pp. 138–140. IEEE.

Henry, D. et E. Stattner (2019). Predictive models for early detection of hoax spread in twitter. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 61–64. IEEE.

Jansen, B. J., M. Zhang, K. Sobel, et A. Chowdury (2009). Twitter power : Tweets as electronic word of mouth. *Journal of the American society for information science and technology 60*(11), 2169–2188.

Kouzy, R., J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, et K. Baddour (2020). Coronavirus goes viral : quantifying the covid-19 misinformation epidemic on twitter. *Cureus 12*(3).

Kramer, A. D., J. E. Guillory, et J. T. Hancock (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 201320040.

Li, J. et P. Meesad (2016). Combining sentiment analysis with socialization bias in social networks for stock market trend prediction. *International Journal of Computational Intelligence and Applications 15*(01), 1650003.

Matthews, C. (2013). How does one fake tweet cause a stock market crash. *Wall Street & Markets : Time*.

McCrae, R. R. et O. P. John (1992). An introduction to the five-factor model and its applications. *Journal of personality 60*(2), 175–215.

Mendoza, M., B. Poblete, et C. Castillo (2010). Twitter under crisis : can we trust what we rt ? In *Proceedings of the first workshop on social media analytics*, pp. 71–79. ACM.

Michelson, M. et S. A. Macskassy (2010). Discovering users' topics of interest on twitter : a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pp. 73–80. ACM.

Posch, L., C. Wagner, P. Singer, et M. Strohmaier (2013). Meaning as collective use : predicting semantic hashtag categories on twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 621–628. ACM.

Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. (2013). Personality, gender, and age in the language of social media : The open-vocabulary approach. *PloS one 8*(9), e73791.

Scott, M. J., S. C. Guntuku, W. Lin, et G. Ghinea (2016). Do personality and culture influence perceived video quality and enjoyment ? *IEEE Transactions on Multimedia 18*(9), 1796–1807.

Shahi, G. K., A. Dirkson, et T. A. Majchrzak (2020). An exploratory study of covid-19 misinformation on twitter. *arXiv preprint arXiv :2005.05710*.

Siarohin, A., S. Lathuilière, S. Tulyakov, E. Ricci, et N. Sebe (2019). First order motion model for image animation. In *Advances in Neural Information Processing Systems*, pp. 7137–7147.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, et I. M. Welpe (2010). Predicting elections with twitter : What 140 characters reveal about political sentiment. *ICWSM 10*, 178–185.

Wald, R., T. M. Khoshgoftaar, A. Napolitano, et C. Sumner (2012). Using twitter content to predict psychopathy. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, Volume 2, pp. 394–401. IEEE.

Yang, S.-H., A. Kolcz, A. Schlaikjer, et P. Gupta (2014). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1907–1916. ACM.

Summary

Nowadays, Twitter is largely used to access or share information related to a large variety of topics. In recent years, researchers have shown that Twitter data are very helpful in several fields and may have many applications. In June 2020, Twitter has announced voice tweets to create a more human experience. This paper gives an overview of possible consequences of this new feature. Furthermore, we present some challenges that researchers will face to exploit this type of data. Finally, we also introduce several opportunities for future research related to voice tweets.

Keywords: social networks, data collection, data analysis, Twitter research

Utilisation de la science des données pour analyser des bases de données d'un observatoire du vignoble français

Elizaveta Logosha**, Solène Malblanc*,
Frédéric Bertrand**, Myriam Maumy-Bertrand **
Céline Abidon *, Sophie Louise-Adèle *

*Institut Français de la Vigne et du Vin
Biopôle – 28 rue de Herrlisheim – 68000 COLMAR
celine.abidon@vignevin.com
**list3n, Université de technologie de Troyes
12, rue Marie Curie CS 42060 - 10004 TROYES CEDEX
elizaveta.logosha@utt.fr, frederic.bertrand@utt.fr, myriam.maumy@utt.fr

Résumé. Depuis l'interdiction de l'arsénite de sodium, il est nécessaire de trouver des moyens de luttes alternatifs contre les maladies du bois de la vigne. C'est dans cette optique que nous étudions les possibles liens entre les pratiques culturales et l'apparition des maladies ESCA et BDA. Pour cela, nous devons isoler les effets dus aux pratiques culturales des effets immuables des caractéristiques de parcelle. Une analyse exploratoire a permis d'identifier des catégories de parcelles, et les variations temporelles ont été prises en compte grâce aux méthodes STATIS et ATP. Au sein de ces parcelles, une analyse d'autocorrélation spatiale a été réalisée afin de vérifier un potentiel effet de contagion. Certaines pratiques ont été identifiées comme préférables, cependant leurs effets sont à relativiser face à l'influence des caractéristiques telles que le cépage et le type de sol. Une étude de plus grande envergure permettrait sûrement de consolider ces résultats.

Mots-clés : analyse exploratoire, autocorrélation spatiale, maladies du bois.

1 Introduction

Depuis l'interdiction de l'arsénite de sodium en France, les pertes de récolte provoquées par les maladies du bois (ESCA, BDA) sont conséquentes : le taux de progression moyen des maladies du bois est entre $0,5$ et 1% par année selon les régions françaises et les cépages (Quéré et Sermier, 2015). Les champignons responsables de ces maladies causent, non seulement, des pertes de rendement les premières années d'infection, mais sont surtout à l'origine de la mort du cep ce qui nécessite un renouvellement des plants pouvant atteindre plus de 10% d'un vignoble (Larignon et al., 2009). L'Alsace n'étant pas épargnée, depuis 2003, un observatoire des maladies du bois regroupant près de 80 parcelles de trois cépages alsaciens (l'*Auxerrois*, le *Gewurztraminer* et le *Riesling*) a vu le jour. Les parcelles sont observées chaque année pour suivre l'expression foliaire des maladies du bois d'une part, et tenter de comprendre l'influence

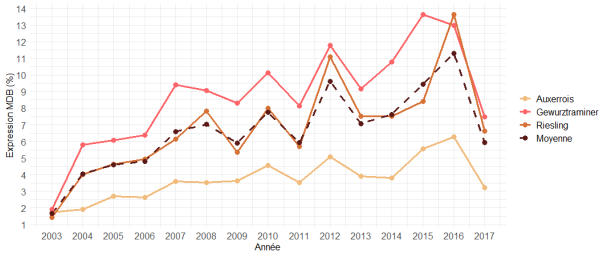


FIG. 1 – *Expression pluriannuelle des maladies du bois de 2003 à 2017.*

de facteurs tels que les caractéristiques parcellaires et les pratiques culturales sur leur apparition et leur développement, d'autre part. La figure 1 illustre l'expression des maladies du bois (équation (1)) entre 2003 et 2017. En 2016, les inter-professions vitivinicoles regroupées au sein du Comité National des Interprofessions des Vins (CNIV) avec le soutien du Ministère de l'Agriculture et de FranceAgriMer se sont engagées dans le Plan National Dépérissement du Vignoble (PNDV). Plusieurs projets ont depuis répondu à l'appel du PNDV dont le projet Eurêka, lauréat en 2017. Eurêka traite du développement de moyens de lutte curatifs et préventifs contre les maladies du bois à travers six actions dont l'action 5 pilotée par l'Association des Viticulteurs d'Alsace (AVA) et l'Institut Français de la Vigne et du Vin (IFV) qui s'appuie sur l'observatoire alsacien. Elle est née du constat suivant : "les pratiques viticoles ont une influence non négligeable sur le développement des maladies du bois", mais aussi de la volonté d'**explorer plus finement la base de données** remplie depuis 2003, au moyen d'outils de la science des données. Cet article présente la méthodologie d'étude de l'observatoire alsacien mis en œuvre au travers de deux bases de données : une base de données dite historique étoffée par des entretiens semi-directifs et une base de données à l'échelle du cep.

2 Méthodologie de la phase exploratoire

2.1 Construction de la base de données

Afin d'affiner et compléter la pré-base de données de l'observatoire, des enquêtes par entretiens semi-directifs individuels des 30 viticulteurs de l'observatoire ont été réalisées. Une base de données détaillée des caractéristiques de chaque parcelle de l'observatoire et des pratiques culturales appliquées sur chacune d'elles par les vignerons de 2003 à 2017 a été constituée. Les données ainsi récoltées concernent, in fine, les viticulteurs de l'observatoire historique alsacien. L'unité élémentaire de l'étude est la parcelle. Ainsi, nous nous plaçons à l'échelle parcellaire, définie comme l'unité homogène de traitement. Placées en première colonne, elles sont codées de la manière suivante : code du viticulteur et nom du cépage. De plus, chacune des colonnes de la base représente une variable relevée. Cette base de données concerne à la fois les pratiques temporelles et fixes. Pour la construire, nous partons du principe que les variables fixes influencent les maladies du bois mais surtout la prédisposition à cette maladie. En effet, les pratiques pérennes ne variant pas, nous concluons que elles n'influencent pas directement la variabilité inter-année.

2.2 Identification de la variable réponse et des variables explicatives

Dans notre analyse, la variable réponse est l'expression des maladies du bois. Cette dernière est une variable quantitative et temporelle, relevée sur 15 ans. L'expression des maladies du bois (MDB), exprimée en pourcentage (Marrou, 2009), est établie à partir des symptômes décelés sur les parcelles parmi les pieds productifs :

$$ExpressionMDB = \frac{(P + T + APO)}{(n - M - A - JP)} \quad (1)$$

avec P = nombre de ceps avec une forme lente partielle, T = avec une forme lente totale, APO = avec une apoplexie, n = nombre total de ceps, M = nombre de morts, A = nombre d'absents et JP = nombre de jeunes plants (les jeunes plants n'étant pas sensibles aux symptômes foliaires). Cette formule donne le rapport entre le nombre de ceps malades et le nombre total de ceux susceptibles de l'être. La figure 2 illustre la répartition de l'expression des maladies du bois, les trois cépages alsaciens confondus. Les variables explicatives correspondent aux

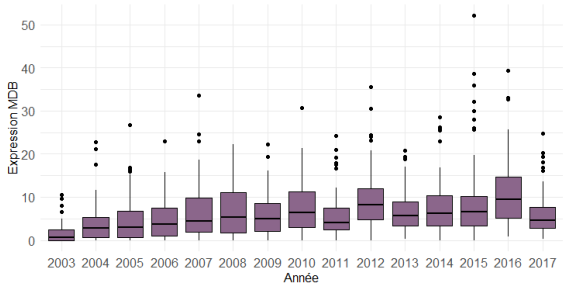


FIG. 2 – *Répartition de l'expression des maladies du bois entre 2003 et 2017.*

caractéristiques des parcelles et des pratiques viticoles. Ces **61 variables** sont relevées lors de la passation du questionnaire. Nous nous intéresserons ici à l'effet des modalités du facteur sur l'expression des maladies du bois. Toutes les variables explicatives sont divisées en deux groupes : pratiques pérennes et pratiques annuelles (Beauchet et al., 2014). Le but de l'analyse est de comprendre les évolutions de la maladie pour mettre en évidence quelles sont les variables impliquées et corriger si possible les pratiques culturales. Une première étape a été la réalisation d'une analyse exploratoire sur les 61 variables. Un *screening* de ces dernières (détection et imputation des valeurs manquantes, détection des valeurs atypiques ou des modalités rares) et des tests d'indépendance ont permis de retenir **47 variables**. Nous avons ainsi pu observer des liaisons significatives entre les variables dont il faudra tenir compte dans la modélisation pour ne pas induire de biais supplémentaire. Afin de comprendre maintenant les liens entre les 47 variables et l'expression des maladies du bois mais aussi les ressemblances entre les parcelles, l'analyse exploratoire a transcrit une vue d'ensemble de la répartition des variables, l'association entre elles ainsi que la répartition des parcelles selon les variables. En effet, ces dernières permettent de comprendre l'organisation des variables et de constituer des typologies de parcelles en observant les ressemblances et les dissemblances. Elles permettent ainsi de résumer l'information contenue dans les données de façon à en dégager les premières tendances (Bertrand et al., 2007, 2008; Fussler et al., 2008).

2.3 Analyse exploratoire

Évolution de l'expression des maladies du bois. Pour étudier l'évolution de l'expression des maladies sur les parcelles et la variation, nous avons utilisé l'analyse en composantes principales (ACP). Le but de l'ACP est de définir H variables, qui sont des combinaisons linéaires des variables initiales. L'ACP permet ainsi d'obtenir un espace construit de composantes principales et de dimension réduite en modifiant le moins possible la configuration initiale et en conservant un maximum d'information. (Voir Fig.3). L'utilisation de la fonction de classement hiérarchique a permis de regrouper les parcelles en fonction de leur ressemblance. (Voir Fig. 8).

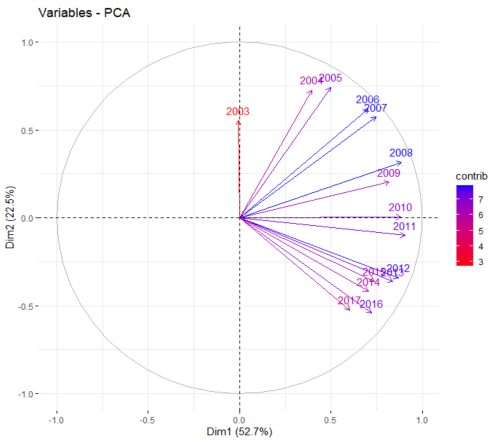


FIG. 3 – *Analyse en composantes principales.*

Identification de typologie de viticulteurs ayant des pratiques culturales annuelles communes. À partir des groupes obtenus par la classification, nous cherchons maintenant à identifier ce qui les caractérise, c'est-à-dire, des stratégies communes concernant les pratiques culturales annuelles. Pour cela, nous avons utilisé l'Analyse des Correspondances Multiples (ACM). L'ACM est une technique visant à résumer l'information contenue dans un grand nombre de variables qualitatives afin de faciliter l'interprétation des liaisons existantes entre ces dernières. Nous cherchions donc à savoir quelles sont les modalités liées entre elles. De plus, l'ACM transforme les observations de départ en des propriétés géométriques qui donnent des représentations graphiques synthétiques. Cette analyse a permis alors de dresser une typologie des viticulteurs. (voir Fig. 9).

3 Étude de la base de données intégrant la temporalité

L'étude de la base de données s'étend de 2003 à 2017, ce qui nous procure un ensemble de **15 tableaux** constitués de **73 parcelles** et de **47 variables**. Ces dernières sont divisées en deux groupes : les fixes et les temporelles (elles correspondent aux pratiques culturales qui peuvent changer durant la période d'observation). L'objectif est d'étudier les trajectoires de dépérissement des parcelles et si possible d'identifier un ensemble de variables explicatives des maladies. En effet, la prise en compte de la temporalité dans l'étude apporte une dimension supplémentaire et un poids aux variables. Aussi, nous nous concentrons non seulement sur l'expression des maladies du bois mais aussi nous incluons une nouvelle réponse, le taux de mortalité :

$$pM = \frac{M}{n}, \quad \text{avec } M = \text{nombre de ceps morts et } n = \text{nombre total de ceps.} \tag{2}$$

L'Analyse Triadique Partielle (ATP) et la Structuration des Tableaux À Trois Indices de la Statistique (STATIS) ont été utilisées ici. Ces deux méthodes (Blanc et al., 1998) permettent une analyse conjointe de plusieurs tableaux. Elles ont été conçues pour des situations de données à plusieurs voies reposant sur l'idée de base du calcul des distances euclidiennes entre configurations de points (Escoufier, 1973). L'idée principale est de comparer les configurations des mêmes observations obtenues dans différentes circonstances. Trois étapes sont nécessaires à la réalisation de ce type de méthodes : (1) l'interstructure (comparaison et analyse des relations entre les différents ensembles), (2) le compromis (fusion des k-tableaux en une structure-compromis analysée en ACP pour révéler les relations entre les observations et pour analyser simultanément les individus et les variables des k-tableaux) (3) l'intrastructure (projection de l'ensemble des données d'origine sur le compromis pour analyser les similitudes et les divergences). L'utilisation des méthodes STATIS et ATP a permis de mettre en évidence l'impact des changements de pratiques culturales des viticulteurs sur l'expression des maladies du bois. Ainsi, l'intérêt des données dynamiques réside principalement dans l'identification des trajectoires de dépérissement c'est-à-dire l'implication et l'enchevêtrement des variables dans le processus de dépérissement.

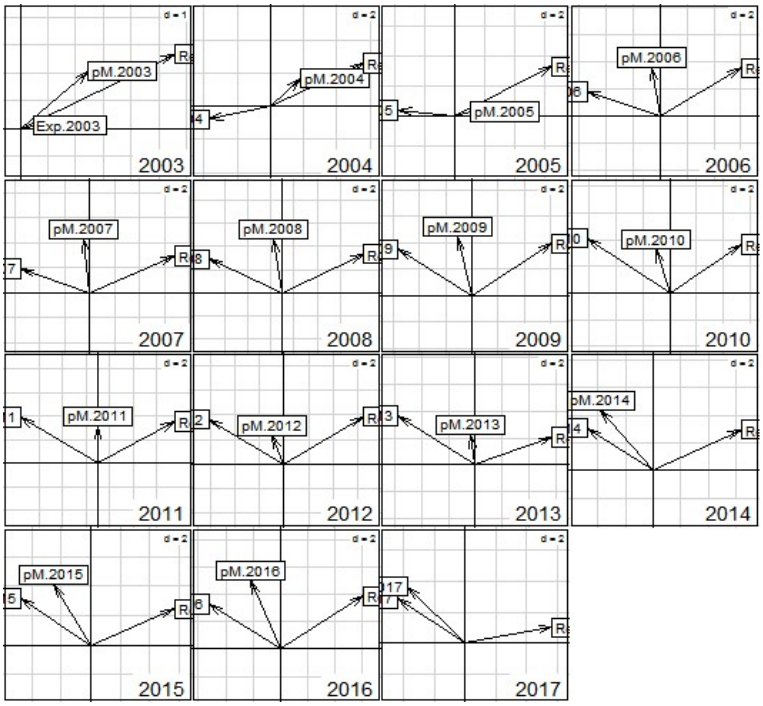


FIG. 4 – *STATIS : Représentation des variables Expression MBD (gauche), taux de mortalité (intitulé pM) et Rendement (droite) par année*

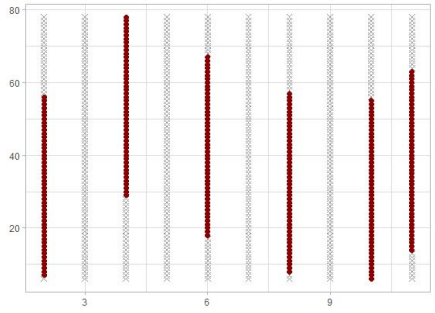
Souvent, les analyses en k-tableaux indiquent que les relations entre une variable explicative et la variable réponse "expression des maladies du bois" n'est valable que pendant une certaine période (Bertrand et Maumy-Bertrand, 2010). Dans ce cas, la variable est considérée importante si cette période dépasse un tiers du temps de l'étude.

4 Étude de la base de données à l'échelle du cep

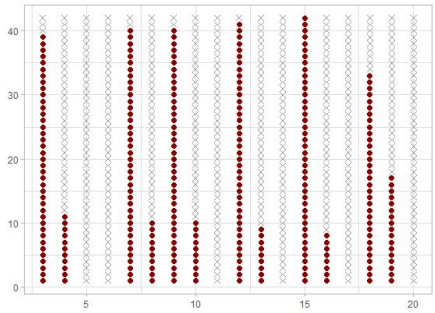
Depuis 2003, sur des fiches de notations, les observateurs de l'IFV et de la Chambre d'Agriculture d'Alsace consignent chaque année l'état des mêmes 300 ceps de chacune des parcelles. À partir de ces fiches, une deuxième base de données a été créée. Elle donne individuellement l'état des 300 ceps des parcelles entre 2004 et 2019, l'année 2003 n'étant pas disponible dans les archives papiers. Dans cette base, chaque cep est classé dans une catégorie parmi les six états : sain, partiellement malade, totalement malade, mort et absent.

4.1 Propagation spatio-temporelle des maladies du bois

Les coordonnées approximatives de géolocalisation des 73 parcelles ainsi que la localisation (numéro du rang et positionnement dans le rang) de chacun des ceps dans les parcelles sont disponibles. Sur chaque parcelle, les 300 ceps sont répartis sur 6 rangs choisis au hasard. Sur ces 6 rangs, 50 ceps sont suivis par rang. La répartition des ceps sur les parcelles suit une distribution homogène. Toutefois, comme la longueur des rangs et leur nombre varient selon les parcelles, il existe une grande hétérogénéité spatiale sur certaines parcelles. Cependant, il est intéressant d'étudier la propagation de la maladie, c'est-à- dire le passage d'un état du cep à un autre état dans l'espace afin de vérifier si la localisation des ceps malades dans le rang est aléatoire. De fait, la plupart des parcelles n'ont pas de rangs d'observation contiguës. Toutefois, au sein d'un rang d'une parcelle, les ceps sont de vrais "voisins". Ainsi, le test (Li, 2015) a été adapté pour analyser les parcelles à l'échelle du rang.



(a) *Parcelle 28. Répartition typique des ceps étudiés sur la parcelle.*



(b) *Parcelle 18. Un exemple de parcelle avec une grande hétérogénéité spatiale.*

FIG. 5 – *Répartition des ceps sélectionnés pour l'étude sur les parcelles* 18 *et* 28*. La couleur rouge indique les ceps choisis pour l'étude, la couleur grise indique les ceps non étudiés, éventuellement situés dans la région, mais pas nécessairement. L'emplacement exact des ceps non étudiés est inconnu.*

4.2 Analyse d'autocorrélation spatiale

L'autocorrélation spatiale est définie comme la corrélation positive ou négative d'une variable avec elle-même en raison de la localisation spatiale des observations (Bouayad-Agha et de Bel-

lefon, 2018). Dans le cas d'une autocorrélation spatiale, la valeur d'une variable pour une observation est liée aux valeurs de la même variable pour les observations voisines. L'autocorrélation spatiale est positive (respectivement négative) lorsque des valeurs similaires de la variable à étudier sont regroupées (respectivement dispersées) géographiquement. Si l'autocorrélation est nulle, la répartition spatiale des observations est aléatoire. La figure 6 illustre schématiquement les trois types d'autocorrélation. Pour mesurer la dépendance spatiale entre

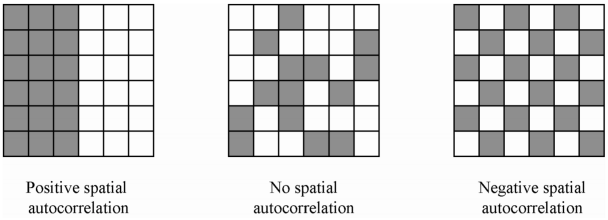


FIG. 6 – *Illustrations de l'autocorrélation spatiale (Radil, 2011).*

les valeurs de la variable « état de santé d'un cep » qui sont binaires en différents endroits de l'espace, le test de Join-Count (Pietrzak et al., 2007, 2014) est utilisé ici. C'est la méthode la plus adaptée compte tenu de l'hétérogénéité spatiale et du fait que les ceps ne soient pas situés côte à côte (de vrais voisins) dans certaines parcelles.

Si les deux zones voisines partagent une frontière commune, elles sont considérées comme jointes. Compte tenu de la disposition des ceps, nous considérons que les zones situées dans un certain rayon les unes des autres comme jointes (Li, 2015) tel que décrit sur la figure 7.

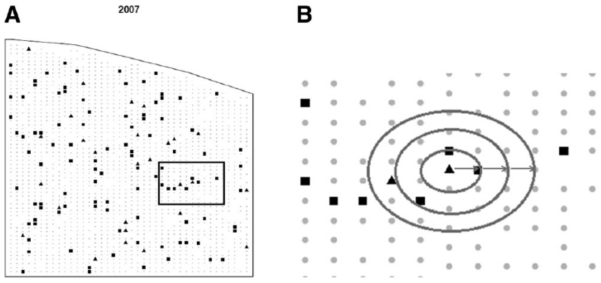


FIG. 7 – *Choix des voisins (Li, 2015)*

Join-Count test est une méthode utilisée pour les variables binaires. Nous avons représenté l'état du cep i sur la parcelle k à un instant t comme proposé par Li (2015) :

$$C^1_{ikt} = \begin{cases} 1 \text{ si le cep } i \text{ de la parcelle } k \text{ a des symptômes de maladies } \textbf{à l'instant } t \textbf{ ou avant} \\ 0 \text{ sinon} \end{cases} \tag{3}$$

ou d'une autre manière, par exemple :

$$C^2_{ikt} = \begin{cases} 1 \text{ si le cep } i \text{ de la parcelle } k \text{ a des symptômes de maladies } \textbf{à l'instant } t \\ 0 \text{ sinon} \end{cases} \tag{4}$$

$$C^3_{ikt} = \begin{cases} 1 \text{ si le cep } i \text{ de la parcelle } k \text{ est mort } \textbf{à l'instant } t \textbf{ ou avant} \\ 0 \text{ sinon.} \end{cases} \tag{5}$$

Pour la définition du concept de zones "jointes" il a été décidé d'utiliser la technique appelée *Row Distance test*, où les ceps sons voisins si ils se trouvent dans le même rang et la distance entre eux se situe dans un certain rayon. Il faut noter que si nous travaillons uniquement sur les parcelles avec les ceps côte à côte, nous allons voir qu'elles appartiennent au même viticulteur. Il n'y a donc pas de changement de pratique.
Nous réalisons le test pour chaque parcelle séparément. Ainsi, nous pouvons voir si sur une parcelle il y a une dépendance spatiale ou pas pour toutes les années et tous les rayons.
La statistique de test JC^{rang}, pour le test sur les rangs, calculée sur la parcelle k à l'année t pour le rayon r est définie :

$$JC^{rang}(k,t,r) = \frac{1}{2}\sum_{i\neq j} C^*_{ikt} C^*_{jkt} w^{rang}_{kij}(r) \tag{6}$$

où

$$w^{rang}_{kij}(r) = \mathbb{1}_{\{r-2<|x_{ik}-x_{jk}|\leqslant r\}}$$

est le poids prenant la valeur 1 si la distance entre deux ceps appartient à la classe $(r-2, r]$, $r = 2, 4 \ldots 16$, et la valeur 0 dans les autres cas, C^*_{ikt} est l'état du cep i déterminé par une des équations (3)-(5) et prend la valeur 1 ou 0, x_{ik} est la coordonnée (dans un rang) du cep i dans la parcelle k.
Nous utilisons la technique de test de permutations avec 1000 simulations pour valider l'hypothèse $\mathcal{H}_0$: les cas des maladies du bois sont distribués au hasard dans la parcelle k à l'année t pour le rayon r contre une hypothèse alternative $\mathcal{H}_1$: les paires de ceps présentant des symptômes et appartenant à la même classe de distance $(r-2, r]$ sont plus fréquentes que dans une distribution aléatoire.
Un test global a également été proposé par Li (2015), permettant d'utiliser simultanément toutes les années et/ou tous les rayons. Une hypothèse alternative dans ce cas serait la suivante : les cas ne sont pas distribués de manière aléatoire pour toutes les années et pour toutes les distances dans la parcelle k. La formule de la statistique JC pour le test Global est donnée par :

$$JC^{glob}(k) = \sum_{2004}^{2019} \sum_{r=2,r+2}^{16} \frac{|JC^{rang}(k,t,r) - \overline{JC^{rang}_{sim}(k,t,r)}|}{\overline{JC^{rang}_{sim}(k,t,r)}}. \tag{7}$$

et représente une somme pondérée des écarts entre les statistiques JC observées calculées à partir de la formule (6) et les moyennes des statistiques JC simulées pour toutes les classes de distance, de 2 m à 16 m, et pour chaque année de 2004 à 2019.

Le test est effectué à la parcelle avec les trois différentes expressions de l'état d'une vigne malade ou morte. Dans ces trois cas, le nombre de parcelles pour lesquelles il existe une autocorrélation spatiale représente moins d'un tiers du nombre total de parcelles étudiées. Nous ne

rejetons donc pas l'hypothèse d'une indépendance spatiale. Il semble donc ne pas y avoir de phénomène de contagion. Il n'y a donc pas besoin d'intégrer dans notre étude la composante spatiale d'un cep au sein de sa parcelle.

4.3 Intérêt d'un modèle spatial

Lors de la constitution de l'observatoire, un petit nombre de ceps a été choisi afin de suivre le pourcentage moyen de ceps touchés par les maladies du bois dans la parcelle. Ainsi, il est tout à fait normal d'observer une hétérogénéité spatiale puisque le choix des rangs et des ceps les uns par rapport aux autres a été décidé en ce sens. La disposition aléatoire était un critère de sélection initial puisque l'objectif de l'observatoire est d'observer le % de ceps touchés par parcelle de façon représentative en gommant le potentiel facteur de contagiosité. La grande hétérogénéité spatiale, le faible nombre d'observations à la parcelle et les résultats du *Joint-count* test qui incitent à ne pas rejeter l'hypothèse d'indépendance spatiale, il a été décidé de laisser de côté le modèle spatial.

5 Résultats obtenus

L'ACP puis la Classification Ascendante Hiérarchique (CAH) ont permis d'identifier cinq groupes d'individus (voir Figure 8). A droite, nous avons une même exploitation qui est très touchée par les maladies du bois pour deux cépages différents, le *Gewurztraminer* et le *Riesling*. Ainsi nous pouvons dire qu'il est important de s'intéresser aux pratiques culturales afin de comprendre quel facteur sous-jacent est impliqué. Par ailleurs, nous constatons que l'*Auxerrois* a une moindre pression.

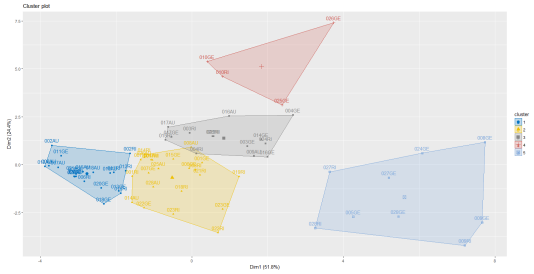


FIG. 8 – *Groupes de parcelles crées par l'ACP et CAH*

L'ACM a permis d'identifier les stratégies de gestion des maladies du bois des viticulteurs de l'observatoire. Nous comptons quatre typologies de viticulteurs avec quatre stratégies différentes représentées équitablement les unes par rapport aux autres. La stratégie la plus efficace combine (qui est liée au groupe ayant l'expression des maladies la plus faible) : une taille respectueuse des flux, un ébourgeonnage et un épamprage manuels, un recépage quasi-systématique, l'emploi de la complantation et de pratiques innovantes de lutte contre les maladies du bois telles que le curetage et le regreffage. La figure 9 présente schématiquement les quatre typologies de viticulteurs de l'observatoire alsacien des maladies du bois ainsi que leurs caractéristiques. Le choix des variables explicatives des maladies du bois découle de trois

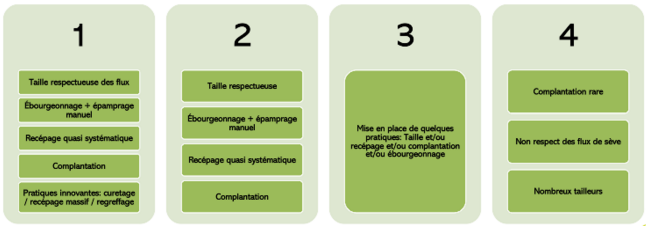


FIG. 9 – *Schéma descriptif des typologiques de viticulteurs de l'observatoire.*

méthodes : analyses multivariées, analyses en k-tableaux prenant en compte la dimension temporelle de l'étude et tests statistiques d'indépendance. D'après ces derniers, parmi les variables qui expliquent l'expression, nous trouvons : le cépage, l'âge de la parcelle, l'unité de sol et la contrainte hydrique.
Le *Gewurztraminer* et le *Riesling* sont les deux cépages les plus sensibles aux maladies du bois à l'inverse de l'*Auxerrois*. Les parcelles les plus touchées par les deux maladies du bois étudiées ont entre 25 et 30 ans et les moins touchées ont moins de 20 ans. Les sols lourds sont les sols les moins sensibles aux maladies du bois. À l'inverse, les sols loessiques sont les plus sensibles. Nous identifions également des pratiques culturales : la taille propre, respectueuse des flux de sève, l'ébourgeonnage et l'épamprage. Ces pratiques culturales empêchent ou retardent le développement et l'expression des maladies du bois.
La prise en compte de la dimension spatiale dans une base de données plus fine, à l'échelle du cep, a permis d'étudier la propagation des maladies du bois. L'analyse de l'autocorrélation spatiale n'a pas permis de valider l'hypothèse d'une dépendance spatiale et donc d'une contamination cep à cep. Néanmoins, sur certaines parcelles, nous avons eu des tests qui permettent de conclure à une répartition non aléatoire des symptômes. Cette constatation reste un point intéressant, et qui pourra être approfondi grâce au rajeunissement de l'observatoire, qui a été réalisé en ajoutant des parcelles aux rangs et ceps contiguës afin de poursuivre l'étude de cette propagation. Les symptômes apparaissent dans la parcelle avec une répartition aléatoire.

6 Conclusion, discussion et perspectives

Les méthodes que nous avons mises en place sur les bases de données du projet **Eureka** ont permis d'identifier un ensemble de variables explicatives des maladies du bois. Ces dernières sont liées à des paramètres fixes propres à la parcelle et inchangés dans le temps comme le cépage et l'unité de sol. Toutefois, nous avons également montré l'impact de pratiques culturales telles que la taille, l'ébourgeonnage et l'épamprage. Par ailleurs, nous avons identifié des typologies regroupant chacune un ensemble de techniques et classé ces typologiques selon leur niveau d'efficacité vis-à-vis des maladies du bois. Ensuite, il a été décidé de nous focaliser sur la dimension temporelle de façon moins complexe. Ainsi, nous avons défini la variable d'intérêt comme la durée entre le début de l'étude et le moment où le cep est atteint par les maladies du bois ou meurt. Un modèle d'analyse de survie utilisant les covariables dépendantes du temps a été réalisé. Ces méthodes ne sont pas présentées ici ainsi que leurs résultats. L'analyse de l'autocorrélation spatiale a permis d'exclure l'impact de la seule dimension spatiale

sur la base de données cep à cep. D'autre part, la mise en place d'un modèle de type Chaînes de Markov prenant en compte la dimension spatio-temporelle et les points de rupture dans les traitements auxquels sont soumis les parcelles permettrait de confirmer l'hypothèse selon laquelle il existe un effet de propagation des maladies du bois de proche en proche. D'autres méthodes pourraient être appliquées comme la réalisation d'un modèle *Partial Least Square Regression* à partir de la base de données de l'observatoire qui pourrait permettre d'identifier toutes les variables ayant un effet sur l'incidence et expliquer plus précisément les maladies du bois. Par ailleurs, l'ajout des données provenant des autres observatoires et leur étude conjointe permettrait de confronter les résultats et peut-être de les consolider. Le projet **Decidep** est né de cette idée et est actuellement à l'œuvre.

Grâce au projet **Decidep** porté par l'UMR Save de l'INRAE de Bordeaux, débuté en 2019 pour une durée de 42 mois, cette analyse se poursuit. **Decidep** propose de mettre en place une logique d'aide à la décision dans le cadre des dépérissements du vignoble. L'IFV est impliqué dans l'action 2 : analyse des interactions pratiques terroirs-dépérissement, dont l'objectif est d'identifier des facteurs techniques explicatifs des dépérissements à l'échelle nationale par le biais de la science des données. Nos bases de données ont été croisées avec d'autres bases issues des observatoires régionaux afin d'élargir le cadre de l'étude à d'autres vignobles français et de valider nos résultats. Une étude sur plusieurs régions françaises est enrichissante et incontournable car elle nous impose de définir des indicateurs communs à différentes situations pour comparer les résultats et sortir du simple résultat obtenu à l'échelle locale. Par ailleurs, ainsi nous apporterons une grande robustesse à nos résultats si l'effet d'une pratique est validé dans un autre contexte. L'observatoire des maladies du bois a une nouvelle fois été mis à profit ce qui prouve la richesse des bases de données générées.

Références

Beauchet, S., C. Renaud-Gentié, V. Cariou, M. Thiollet-Scholtus, R. Siret, et F. Jourjon (2014). Analyses multivariées pour une meilleure compréhension des pratiques viticoles et des facteurs du milieu afin d'expliquer la qualité du raisin. In Array (Ed.), *37th World Congress of Vine and Wine and 12th General Assembly of the OIV (Part 2)*, pp. 05008.

Bertrand, F., M. Maumy, L. Fussler, N. Kobes, S. Savary, et J. Grosman (2007). Using factor analyses to explore data generated by the national grapevine wood diseases survey. *Case Studies in BIGS 1*(2), pp. 183–202.

Bertrand, F., M. Maumy, L. Fussler, N. Kobes, S. Savary, et J. Grosman (2008). Étude statistique des données collectées par l'observatoire national des maladies du bois de la vigne. *Journal de la société française de statistique 149*(4), 73–106.

Bertrand, F. et M. Maumy-Bertrand (2010). Using partial triadic analysis for depicting the temporal evolution of spatial structures : Assessing phytoplankton structure and succession in a water reservoir. *Case Studies in BIGS 4*(1), 23–43.

Blanc, L., D. Chessel, et S. Dolédec (1998). Etude de la stabilité temporelle des structures spatiales par analyses d'une série de tableaux de relevés faunistiques totalement appariés. *Bulletin Français de la Pêche et de la Pisciculture 348*, 1–21.

Bouayad-Agha, S. et M.-P. de Bellefon (2018). Spatial autocorrelation indices. In V. Loonis et J.-L. Tavernier (Eds.), *Handbook of Spatial Analysis. Theory and practical application with R*, Chapter 3, pp. 51–68. Insee - Eurostat.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics 29*(4), 751–760.

Fussler, L., N. Kobes, F. Bertrand, M. Maumy, J. Grosman, et S. Savary (2008). A characterization of grapevine trunk diseases in france from data generated by the national grapevine wood diseases survey. *Phytopathology 98*(5), 571–9.

Larignon, P., F. Fontaine, S. Farine, C. Clément, et C. Bertsch (2009). Esca et black dead arm : deux acteurs majeurs des maladies du bois chez la vigne. *Comptes Rendus Biologies 332*(9), 765–783.

Li, S. (2015). *Spatio-temporal modelling of esca grapevine disease at vineyard scale*. Theses, Université de Bordeaux.

Marrou, H. (2009). Etude des facteurs environnementaux et techniques favorisant l'expression des symptômes des maladies du bois de la vigne (Esca et Black Dead Arm), dans le vignoble bordelais. Montpellier SupAgro.

Pietrzak, M., J. Wilk, R. Bivand, et T. Kossowski (2014). The application of local indicators for categorical data (licd) in the spatial analysis of economic development. *Comparative Economic Research 17*, 203–220.

Pietrzak, M., J. Wilk, T. Kossowski, et R. Bivand (2007). The identification of spatial dependence in the analysis of regional economic development - join-count test application. Institute of Economic Research Working Papers 30/2013, Toruń.

Quéré, C. et J.-M. Sermier (2015). Rapport d'information sur les maladies de la vigne et du bois. Rapport d'information, Assemblée nationale.

Radil, S. M. (2011). *Spatializing social networks : making space for theory in spatial analysis*. Dissertation, University of Illinois at Urbana-Champaign.

Summary

Since sodium arsenite has been banned, winegrowers have to find alternative ways of fighting wood diseases. In this paper we investigate the possible links between cultivation practices and the emergence of ESCA and BDA diseases. To do so, we have to separate the effects of cultivation practices from the immutable effects of land characteristics. An exploratory analysis revealed different categories of fields, and chronological variations were taken into account using STATIS and PTA methods. Within each field, a spatial autocorrelation analysis was used to test for a potential propagation effect. Some practices were found to be preferable, but their effects are to be put into perspective in view of the influence of characteristics such as grape variety or soil type. A broader study would certainly help to consolidate these results.

Keywords: exploratory data analysis, spatial autocorrelation, wine desease.

Recommandations en cas d'urgence : mobilité urbaine des ambulanciers

Ayoub Charef *, Zahi Jarir* and Mohamed Quafafou **

* Laboratoire Ingénierie des Systèmes Informatiques (LISI), Faculté des sciences, Université Cadi Ayyad, Marrakech, Morocco
ayoubcharef@ced.uca.ma
jarir@uca.ma
** Laboratoire LSIS, Université Aix-Marseille, Marseille, France
mohamed.quafafou@univmed.fr

Résumé. Grâce à l'évolution croissante des technologies et des techniques avancées telles que l'Internet des Objets, l'IA et le Big Data, l'industrie des systèmes de gestion du trafic a acquis de nouvelles méthodologies de création de services et d'applications avancées et intelligentes pour la gestion et la sécurité du trafic urbain. Notre contribution actuelle se focalise sur la mise en place d'un service de recommandation de chemin pour ambulanciers en état d'urgence que nous qualifions comme l'une des questions les plus critiques et complexes de la gestion du trafic pour la survie des individus impliqués dans des incidents d'urgence. Nous avons principalement orienté cette étude sur le temps de réponse aux incidents mettant la vie des patients en danger, qui est un indicateur pour les services d'ambulance d'urgence, pour recommander un trajet d'ambulance plus court. Pour cela, nous nous sommes intéressés aux techniques de machine learning pour prédire le parcours le plus rapide pour atteindre une destination. Plus particulièrement, nous proposons en premier lieu une approche locale basée sur la technique KNN pour prédire la congestion des différents tronçons d'une carte depuis une origine à une destination sous OpenStreetMap. En second lieu, nous suggérons une approche globale montrant aux ambulanciers le chemin le plus rapide en temps réel au fur et à mesure de leurs mobilités.

Mots-clés : Systèmes de recommandation, trafic urbain en cas d'urgence, mobilité des ambulances, services d'urgence, service de navigation, apprentissage automatique.

1. Introduction et motivation

De nos jours, les systèmes de transport intelligents (STI) sont de plus en plus demandés en raison de la croissance rapide des technologies de l'information et de la communication. En général, ces systèmes consistent à traiter les données collectées et à prendre des décisions ciblées et intelligentes qui augmentent l'efficacité du transport, réduisent la congestion et, en

particulier, améliorent la sécurité. Les systèmes de transport d'urgence en ambulance représentent l'un des systèmes les plus importants et les plus critiques. Ces systèmes doivent prendre en compte les actions immédiates qui suivent un incident de santé personnel mettant en danger la vie de la victime et qui sont essentielles à sa survie. L'arrivée rapide d'ambulanciers qualifiés, en particulier, fait souvent la différence entre la vie et la mort (Poulton et al., 2018). Cependant, la recherche d'une réponse efficace est considérée comme un défi complexe dans la littérature. Cette complexité s'accentue lorsque d'autres contraintes ou facteurs doivent être respectés, comme l'impossibilité de créer des voies d'urgence en raison de l'étroitesse des routes, etc. L'idée est de proposer un service de mobilité d'urgence ayant l'avantage de recommander en permanence aux conducteurs d'ambulances le chemin le plus rapide pour atteindre la destination. Les données utilisées proviennent des capteurs de l'infrastructure routière, de l'IoT installé dans l'ambulance, etc. ainsi que des données des trajets précédemment parcourus, telles que les incidents, les changements de ligne, la durée du trajet, le flux de trafic, etc. Toutes ces données nous aident à estimer le temps de trajet pour n'importe quel itinéraire et l'intervalle de vitesse d'une ambulance répondant à un événement d'urgence. Dans cette contribution, nous nous concentrons principalement sur la manière de réduire le temps de trajet des ambulances en relevant les défis d'une réponse efficace des ambulanciers. Pour atteindre cet objectif, notre approche est double : d'abord, analyser la mobilité des ambulances d'urgence et identifier les principaux facteurs empêchant leurs déplacements, et ensuite, modéliser une solution de recommandation de trajet le plus rapide pour améliorer le temps de la mobilité des ambulances en se concentrant sur la collecte de données à partir de capteurs de l'infrastructure routière, des données archivées sur l'historique des anciens déplacements confirmés, etc.

Le reste de l'article est organisé comme suit. La section 1 présente un travail connexe sur les solutions de recommandation de chemin pour les ambulanciers en état d'urgence présentées dans la littérature. La section 2 détaille les contraintes à prendre en considération pour proposer une solution de recommandation de chemin en cas d'urgence. La section 3 décrit l'architecture multicouche de notre solution de recommandation. La section 4 présente une approche hybride permettant la prédiction de l'état du trafic basée sur la technique kNN et la proposition du chemin le plus rapide pour atteindre une destination en utilisant l'algorithme de Dijkstra en temps réel. Enfin, la section 5 conclut l'article et présente les travaux futurs.

2. Travaux connexes

Il est difficile pour les services d'ambulances d'urgence d'obtenir des temps de réponse courts lors des interventions sans recourir à des algorithmes efficaces et à des technologies avancées. Dans la littérature, plusieurs approches ont été proposées pour réduire le délai de réponse des ambulances. Ce délai est principalement dû aux feux de circulation qui créent une foule de véhicules et mettent ainsi les ambulances en attente. Il serait très utile pour l'ambulance si les feux de circulation sur le chemin entre la position d'activation de l'ambulance et la scène d'un incident ou d'un patient sont toujours allumés comme présenté dans Weisfeldt et Becker (2002). Malheureusement, cette solution est très contraignante. Dans le même contexte, les auteurs Athavan et al. (2012) proposent une méthode de contrôle des feux de circulation et de réalisation de la tâche susmentionnée afin de permettre à l'ambulance de traverser tous les carrefours sans attendre. Alors que le travail de Vasuki et Ruthvik (2015) utilise les capteurs radiofréquence pour compter des véhicules, qui sont équipés d'éti-

quettes d'identification par radiofréquence, afin de proposer un signal de circulation dynamique comme solution pour éviter le blocage des ambulances.

De plus, Azimi et al (2017) propose un algorithme d'optimisation pour minimiser la différence entre les demandes et les offres du réseau contraint (diagrammes de Voronoï), pour une allocation optimale de l'espace aux centres d'urgence en fonction de la population, de la rue et du nombre d'ambulances. Les auteurs de Ibri et al (2012) proposent un système multi-agents pour intégrer un dispatching dynamique des ambulances afin de limiter les déviations des véhicules, et d'affecter un véhicule à un autre appel si ce dernier est plus prioritaire que le premier.

La contribution de Poulton et al. (2018) introduit une méthodologie basée sur les données pour la prédiction précise de l'itinéraire suivi par une ambulance répondant à un incident d'urgence et se déplaçant avec des feux bleus et des sirènes. Cette méthode utilise les données historiques pour estimer la vitesse moyenne à laquelle une ambulance répondant à une demande d'urgence, puis estime le temps de trajet à partir de ces vitesses pour chaque route. Enfin, elle utilisant un algorithme standard (Hidden Markov Model) de théorie des graphes pour déterminer l'itinéraire le plus rapide entre n'importe quel point de départ et d'arrivée.

En se référant à la littérature, les approches proposées ne sont pas complètes et nécessitent des améliorations pour être pleinement opérationnelles. Cela nous motive à contribuer, d'une part, à la mise en place d'une architecture flexible pour collecter, traiter et analyser les informations utiles recueillies à partir de l'infrastructure routière et du trafic et, d'autre part, à l'utilisation des algorithmes de machine learning combinés avec l'algorithme de Dijkstra amélioré et des techniques de Big Data.

3. Approche de recommandation de chemin en cas d'urgence

L'objectif de cette contribution est de proposer une solution de recommandation de trafic d'urgence ayant l'avantage de proposer des parcours d'une origine à une destination tout en évitant le plus possible que les ambulances ne restant coincées dans les embouteillages lors de leurs déplacements La mise en place d'un tel système de **recommandation de trafic d'urgence** nécessite des données empiriques massives pour construire le chemin le plus approprié et le plus rapide pour les ambulanciers. De ce fait, nous proposons en premier lieu une architecture flexible ayant l'avantage de collecter toutes les données requises issues de l'environnement et également de les traiter en temps réel pour répondre à l'urgence du déplacement.

L'architecture que nous proposons fait appel à des techniques de Big data pour collecter, filtrer, traiter les données en temps réel, et des techniques d'intelligence artificielle telles que les techniques d'apprentissage par renforcement pour analyser et prédire le chemin le plus rapide en temps réel tout au long de la mobilité.

Conformément à notre vision et afin de répondre à ce défi, nous avons commencé par comprendre, identifier et analyser tous les paramètres impliqués dans le processus de recommandation de trafic d'urgence avant la phase de modélisation. Cela s'explique par le fait que le développement de solutions intelligentes sans une architecture appropriée est une tâche très difficile à accomplir.

Les paramètres nécessaires que nous avons identifiés appartiennent principalement à deux grandes catégories : les paramètres statiques et dynamiques. Les paramètres statiques font référence aux données qui ne changent pas pendant la période du trajet, comme les données de l'infrastructure, la signalisation routière, les structures des routes, la planification des feux rouges, etc. Tandis que la catégorie dynamique se focalise sur les variables dont les valeurs évoluent dans le temps et qui concernent par exemple les trajets de circulation, les données collectées par les objets connectés IoT ou les capteurs, les comportements des conducteurs, le flux de trafic, etc. La figure 1 résume les composants et les paramètres pris en compte par notre architecture.

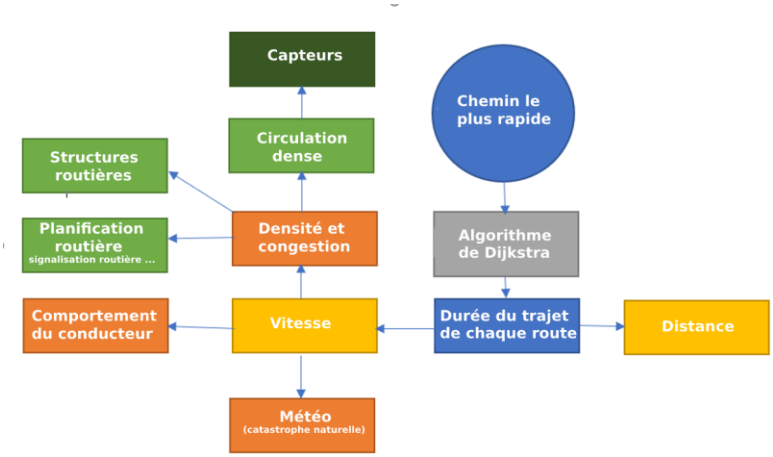


FIG. 1 – *Composants et paramètres liés à la mobilité d'urgence.*

Afin de proposer un modèle flexible pour les services de trafic d'urgence pour la cas de la mobilité des ambulanciers, nous nous sommes intéressés à l'utilisation des systèmes multi-agents. Nous trouvons que ce type de système est plus adapté à notre cas et répond parfaitement à nos besoins. En plus, ces systèmes possèdent des capacités à détenir les propriétés des systèmes distribués, à répondre à la coordination requise qui consiste à organiser la coopération entre les objets connectés en partageant des connaissances et en s'appuyant sur les capacités et les connaissances de chaque entité, et à s'adapter aux techniques de l'intelligence artificielle. Pour toutes ces raisons, dans ce papier, nous avons opté pour SARL[1] (System Multi-agent Programming Language) comme langage de programmation d'agents compte tenu de sa robustesse, de la multitude de fonctionnalités offertes, etc.

[1] http://www.sarl.io/.

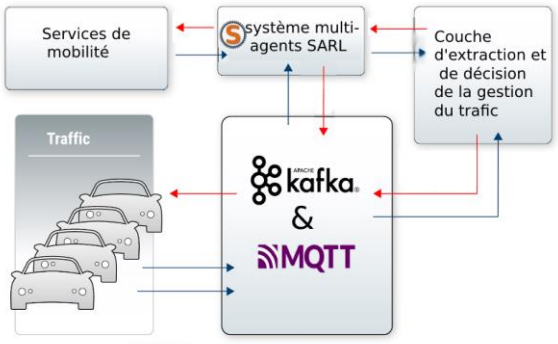


FIG. *2 – Une architecture pour les services de mobilité d'urgence.*

Les principaux modules de notre architecture concernent un module dédié au système multi-agents, un module de gestion des données représenté par une solution Big Data telle que Kafka (Kafka Streaming Platform), un module d'extraction et de décision pour la gestion du trafic, et un module services de mobilité qui représente l'interface du système de recommandation offrant aux ambulanciers des informations de positionnement en temps réel, le chemin le plus rapide entre un point de départ et destination, etc. Ce dernier module est en interaction avec le module système multi-agents, qui par le biais des agents, permet de gérer chaque objet physique (capteurs, IoT, etc.) en les abstrayant, et de contrôler leurs tâches et services souscrits à des topics particuliers en utilisant le protocole Publish/Subscribe MQTT (Message Queuing Telemetry Transport). Ces agents sont connectés en permanence à l'infrastructure routière. En se basant sur les informations ainsi collectées de l'infrastructure physique, ce module est capable de composer l'événement ou la tâche requise en utilisant un processus de coordination et de l'envoyer au module services de navigation. Toutes informations utilisées sont transférées vers la plateforme Apache [2]Kafka. Cette dernière est une plateforme distribuée de diffusion de données en continu, capable de publier, stocker, traiter et souscrire à des flux d'enregistrement en temps réel. Elle est conçue pour gérer des flux de données provenant de plusieurs sources. Nous avons intégré Kafka[2] en raison de son haut débit, de sa faible latence, de son évolutivité, de la gestion en temps réel de son pipeline de données, etc.

Le module d'extraction et de décision de la gestion du trafic fournit la décision globale sur la mesure appropriée de la mobilité, de la fluidité dans les tronçons et les nœuds à traverser pour atteindre une destination à partir d'une origine. Cette décision est basée sur l'historique d'états des trafics urbains, des données collectées en temps réel par le système multi-agents et transférées vers Kafka, et sur les algorithmes appropriés (algorithme de Dijkstra, algorithme de prédiction, etc.)

Pour mettre en place notre solution, nous proposons une approche hybride : (1) une approche locale dédiée à la prédiction du trafic basée sur un algorithme d'apprentissage et

[2] https://kafka.apache.org/

(2) une approche globale pour proposer le chemin le plus rapide d'une origine à une destination.

4. Approche hybride de recommandation

L'approche hybride que nous proposons pour recommander un trajet dans le cas d'urgence d'un point de départ à un point d'arrivé pour un ambulancier dont le temps de parcours doit être réduit au minimum est décrite comme suit.

Lors du départ de l'ambulancier, tous les tronçons qui constituent un chemin possible entre A et B sont coloriés en fonction de leur état de congestion. Ainsi, l'état du trafic est évalué et peut être "très congestionné", "congestionné" ou "fluide". Le conducteur de l'ambulance doit alors prendre une décision à chaque carrefour (point d'intersection) pour choisir la direction qu'il va prendre en tenant compte de l'état du trafic sur les sections de routes permettant d'aller au point d'arrivée. Il peut suivre les recommandations traduites par le coloriage des chemins ou prendre une autre voie alternative et qui peut même être en contradiction avec la recommandation. Les décisions prises sont enregistrées et constituent l'expérience du conducteur. La coloration des tronçons est mise à jour chaque fois que l'ambulance s'approche d'un point d'intersection. La prédiction du trafic dans une section est faite par l'algorithme KNN décrit par l'approche locale, alors que l'approche globale consiste à suggérer un chemin d'urgence à temps minimal d'un point de départ à un point d'arrivé en se basant sur l'algorithme de Dijkstra. Les sous-sections suivantes détailleront ces deux approches.

4.1 Approche locale : prédiction de l'état du trafic basée sur KNN

Les états du trafic routier d'une même route dans les séries temporelles sont assez réguliers et peuvent être reproduits dans une certaine mesure. Ainsi, grâce à une correspondance efficace entre l'historique et les états actuels du trafic routier, les états futurs du trafic routier peuvent être prédits efficacement.

Dans ce papier, nous nous basons sur le modèle K-Nearest Neighbor (KNN) vu les avantages qu'il offre. Pour ce faire, nous allons tout d'abord construire une base de données historique représentative de grande capacité. Ensuite, nous définissons les éléments du modèle, y compris la valeur du vecteur d'état de k et l'algorithme de prédiction. Enfin, en fonction des valeurs observées de l'entrée et du mécanisme de recherche, un voisin proche correspondant au temps réel actuel et les données d'observation de la base de données historique sont récupérés pour prédire le flux de trafic au moment suivant.

4.1.1 Génération de trafic aléatoire

Pour évaluer notre approche et vu l'absence de données empiriques, nous nous sommes intéressés à générer d'une manière aléatoire des données représentant des véhicules, sur les tronçons et les intersections de route, qui se déplacent aléatoirement avec une vitesse constante. Les propriétés principales de cette génération de trafic sont présentées comme suit.

— Initialisation à t=t0 : Attribuer à chaque intersection un nombre de véhicules NV;

— Pour chaque intersection I à l'instant t avec le nombre de véhicules NVi(t), attribuer à chaque tronçon de sortie de l'intersection I un pourcentage aléatoire XTi(t) % de NVi(t) ;

— Les tronçons de sortie d'une intersection I correspondent aux tronçons d'entrée à d'autres intersections J1, J2, ..Jn ;

— Nous présentons les équations suivantes Buisson. (2010) ;

— Le flux de trafic est le nombre N de véhicules passant une période Δt au point x,

$$Q_{\Delta t}(x) = Q(x, t \rightarrow t + \Delta t) = \frac{N}{\Delta t} \quad (1)$$

— La densité K correspond au nombre M de véhicules entre x et x+Δx à l'instant t,

$$K_{\Delta x}(t) = K(x \rightarrow x + \Delta x, t) = \frac{M}{\Delta x} \quad (2)$$

— Ces deux équations (1) et (2) permettent de calculer le temps nécessaire pour passer d'une intersection I à une intersection J;

Un exemple de visualisation du trafic suite à la génération des véhicules est montré par la figure 3.

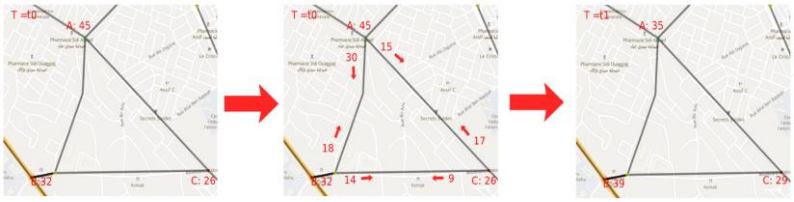


Fig. 3 – *Exemple de génération du trafic entre les intersections et les tronçons.*

4.1.2 Algorithme de coloration de congestion dans les tronçons et les intersections

Les états de circulation des segments routiers d'un réseau routier ont tendance à être influencés par leurs mouvements en amont et en aval. Par exemple, les congestions sont souvent initiées sur un ou plusieurs segments de route et se propagent à d'autres segments de route dans la majorité des cas après un certain temps. La figure 4 illustre la congestion en attribuant des couleurs aux segments de route ; les lignes rouge, jaune et verte représentent respectivement les segments de route congestionnés, légèrement congestionnés et non congestionnés.

Pour visualiser les changements de congestions sur une carte, nous proposons de colorer les tronçons et les intersections en respectant les deux règles suivantes :

Règle 1 — CIj la capacité maximale de l'intersection J , donc si la valeur de NVj(t) s'approche de CIj alors il aura une congestion. Trois classes de congestion sont prises en compte par notre experimentation : classe 0 pour NVj(t)<= CIj/3 (coloration verte), Classe 1 pour CIj/3< NVj(t) <= 2CIj/3 (coloration en jaune) et Classe 2 pour 2Cj/3< NVj(t) <= CIj (coloration en rouge);

Règle 2 — CTk la capacité maximale du tronçon K, donc si la valeur de XVk(t) s'approche de CTk alors il aura une congestion. Trois classes de congestion sont prises en compte par notre experimentation : classe 0 pour NVk(t)<= CTk/3 (coloration verte), Classe 1 pour CTk/3< NVk(t) <= 2CTk/3 (coloration en jaune) et Classe 2 pour 2CTk/3< NVk(t) <= CTk(coloration en rouge);

L'idée est d'exploiter le fait que les embouteillages se déplacent, se dissipent où se forment d'un tronçon à l'autre ou d'une intersection à une autre. En vue de reconnaître ces congestions et les montrer sur une carte, nous nous sommes intéressés à l'algorithme KNN pour

prédire l'état d'un tronçon sorti d'une intersection I à l'instant t à base des historiques du trafic des tronçons d'entrées de la même intersection I à l'instant t-1.

Les données de cette historique simulées sont enregistrées dans une base de données MongoDB et proviennent des flux de trafic (nombre de voitures par 3 minutes) générés aléatoirement sur 40 intersections et 60 tronçons sur la carte de la ville Marrakech au Maroc. Ces données collectées qui correspondent à environ 1000 heures d'exécution de l'algorithme de génération du trafic ont engendré plus de 30000 éléments pour chaque tronçon. L'algorithme KNN appliqué est présenté formellement dans l'algorithme 1.

Algorithme 1 : KNN pour la Prédiction et la coloration de congestion dans les tronçons

Soit ;
D l'ensemble des objets d'apprentissage ;
Z(I)=(NV1,NV2..NVk,t) le vecteur de valeurs en temps réel, tel que NVIj représente le nombre de véhicules du tronçon d'entrée j de l'intersection I à l'instant, t) avec j appartient à 1...k ;
CG l'ensemble d'états du trafic(0 : trafic fluide/1 : semi-congestionné /2 : congestionné);
L classe utilisée pour étiqueter les objets (CG , plus le nombre des véhicules du tronçon de sortie de l'intersection I à l'instant t+$\triangle t$)
Input : Z(I) ;
Output : Cz=(CG, nombre des véhicules du tronçon de sortie) $\in L, classe \quad de \quad Z$
for *objet* $y \in D$ **do**
  calculer d(z,y) , la distance entre z et y;
end
sélectionner N de D, l'ensemble (voisinage) des k objets training les plus proches pour z ;
cap-max est la capacité maximale du tronçon sortie de l'intersection I ;
$Cz = \sum_{y \in N} NV(t)/cap - max$

Prenons à titre d'exemple le tableau 1.

Segment	Départ	Destination	Capacité	Véhicules
A-B	A	B	30	10
A-C	A	C	25	9
B-C	B	C	43	20
B-F	B	D	45	28
C-E	C	E	30	14
D-E	D	E	32	29
D-F	D	F	22	8
J-H	J	H	26	25
H-I	H	I	12	10
N-O	N	O	30	20
K-M	K	M	20	17

TAB. 1 – *Nombre de véhicules sur chaque tronçon à 10h.*

L’utilisation de l’algorithme KNN nous permet de prédire les colorations des tronçons (voir figure 5, carte 10h).

Fig. 4 – Coloration des congestions basée sur KNN.

La figure 5 présente l’évolution dans le temps des congestions dans les tronçons

Fig. 5 – État de congestion pour chaque tronçon en fonction du temps.

L’intégration des résultats obtenus par l’approche locale basée sur KNN dans le système de recommandation permet aux ambulanciers de prendre des décisions partielles sur le parcours à suivre. Pour renforcer cette recommandation, nous intégrons une deuxième approche qualifiée globale qui a pour but de suggérer un chemin en état d’urgence depuis une origine à une destination. Cette approche, détaillée dans la sous-section suivante, repose sur l’algorithme de Dijkstra vu sa simplicité et son efficacité.

4.2 Approche globale : prédiction basé sur le coefficient de fluidité du trafic

Pour déterminer le chemin le plus rapide entre un point de départ et toutes les autres intersections d'un graphe (carte), nous proposons comme première contribution d'adapter l'algorithme de Dijkstra pour prendre en compte le paramètre de fluidité du trafic. La valeur de ce paramètre est calculée en fonction de la densité routière en temps réel, l’historique du

trafic, ainsi l'état de la route (travaux publics, ...), etc. Cette valeur est également calculée en continu à partir des données collectées à l'instant par les capteurs de l'infrastructure routière afin de trouver le chemin le plus rapide à partir d'un point de départ et d'exclure la route présentant un degré élevé de congestion. Avant de présenter l'équation (6) utilisée pour déterminer le temps nécessaire pour traverser chaque route, nous présentons les équations auxquelles elle fait référence.
La densité K est exprimée en nombre de véhicules par unité de longueur (véh/km ou véh/m).

$$k = \frac{TO}{L+l} \qquad (3)$$

où TO est le taux d'occupation, L est la longueur des véhicules supposés et l si la longueur de la boucle de détection, tell que Q est Le flux de trafic.

$$Q = K \times V \qquad (4)$$

Ainsi, la vitesse v s'exprime comme suit.

$$V = \frac{Q}{TO} \times (L + l) \qquad (5)$$

Enfin, le temps de trajet Ti pour l'arête i est calculé comme suit.

$$Ti = \frac{Vi}{Di} \qquad (6)$$

L'algorithme 2 utilisé suit les étapes suivantes :

1. Initialiser toutes les routes avec un haut degré de congestion "infini", en d'autres termes les véhicules ne peuvent pas se déplacer. De plus, l'ambulance est initialement au point de départ à l'instant T0.
2. Activer le point de départ
3. Calculer le temps du parcours temporaire de toutes les intersections voisines de l'intersection courante en additionnant leur temps
4. Si le temps calculé d'une intersection est inférieur au temps actuel, mettre à jour le temps et définir l'intersection actuelle comme antécédent. Cette étape est également appelée mise à jour et constitue l'idée centrale de Dijkstra.
5. Définir l'intersection avec le temps temporaire minimal comme active. Marquer son temps de parcours comme permanent.
6. Répétition des étapes 3 à 5 jusqu'à ce qu'il ne reste plus d'intersections avec un temps permanent, dont les voisins ont encore un temps de parcours temporaire.

Algorithme 2: Dijkstra pour trouver le chemin le plus rapide

```
Input : source , destination , graphe des chemin possible ;
Output : path [ ] chemin le plus rapide;
for sommet v in Graph do
    temps[v]:= infinité;
    précédent[v]: = undefined;
end
path[]: = undefined;
temps[source]: = 0 ;
Q: = l'ensemble de toutes les intersections dans le Graph;
while Q n'est pas vide do
    u:= intersection dans Q avec le plus petit temps[];
    supprimer u de Q;
    for voisine v de u do
        alt:= temps[u] + temps entre (u, v);
        if alt est plus grand que temps[v] then
            temps[v]:= alt;
            précédent[v]:= u;
        end
    end
end
i=0;
while destination ≠ précédent[i] do
    path[i]:= précédent[i];
    i++;
end
path[i]:=destination;
return path [ ];
```

Pour mettre en œuvre cette application, nous avons opté pour le langage de programmation orienté agent SARL (System Multi-agent Programming Language), le SGBD MongoDB, et le langage JavaScript pour construire le service de navigation d'urgence pour les ambulanciers en se basant sur OpenStreetMap. Le chemin le plus rapide est dessiné en bleu (voir figure 6, carte 1) et est mis à jour jusqu'à (cartes 2 et 3) l'arrivée à destination à chaque intersection. La ligne dessinée en vert représente le chemin de recommandation de départ (carte 4).

Fig. 6 – Mis à jour du chemin recommandé jusqu'à l'arrivée à destination.

Sur la base des résultats obtenus, nous travaillons actuellement à l'amélioration de l'équation proposée (6) pour inclure des paramètres plus détaillés tels que l'état de la route (travaux publics, présence d'accidents, …), le type de foule de véhicules, la vitesse de la foule de véhicules, ...

5. Conclusion and Perspectives

L'objectif de cette contribution est de proposer une solution de recommandation de trafic en état d'urgence afin d'éviter que les ambulances ne soient coincées dans les embouteillages au moment de déplacements, interventions. Pour cela, nous avons proposé une architecture basée sur les techniques de machine learning de classification et la collecte des informations requises pour prédire le chemin le plus rapide aux ambulanciers. Vu l'absence de données empiriques, nous avons proposé un algorithme de génération d'un trafic aléatoire pour expérimenter notre solution de recommandation. Basé sur le dataset généré, nous avons proposé une approche hybride composée d'une approche locale pour visualiser l'état de congestion sur la carte de la ville de Marrakech sous OpenStreetMap et d'une approche globale permettant de suggérer le chemin le plus rapide au fur et à mesure du déplacement des ambulances. L'approche locale repose la technique KNN pour prédire la congestion des tronçons en trois classes : tronçons congestionnés, peu congestionnés et non congestionnés. Alors que l'approche globale implémente l'algorithme de Dijkstra en utilisant le langage de modélisation des systèmes multi-agents SARL.

La solution de recommandation proposée permet de prédire le chemin le plus rapide aux ambulanciers d'une origine à une destination. Vu que ces recommandations peuvent ou pas être prises en considération par les ambulanciers, nous sommes actuellement en train d'intégrer un module d'apprentissage de la mobilité en état d'urgence qui peut être soit recommandée par le système ou décidée par les ambulanciers.

6. Références

Athavan K., Balasubramanian G., Jagadeeshwaran S., Dinesh N . (2012). Automatic ambulance rescue system. IEEE. Publié dans second international conference on advanced computing & communication technologies.

Azimi S., Delavar R., Rajabifard A. (2017). Multi-agent simulation of allocating and routing ambulances under condition of street blockage after natural disaster. The International Archives of the Photogrammetry, Remote sensing and spatial Information Sciences, Volume XLII-4/W4.

Buisson C. (2010)., Lesort J Comprendre le trafic routier. Méthodes et calculs. CERTU, 111p, 2010, Coll. Références, 978-2-11-098892-8.

Ibri S., Nourelfath M., Drias H. (2012). A multi-agent approach for integrated emergency vehicle dispatching and covering problem, engineering applications of artificial intelligence,Volume 25, Issue 3.

Poulton M., Noulas A., Weston D., Roussos G. (2018). Modelling metropolitan-area ambulance mobility unde blue light conditions. IEEE.

Vasuki S., Ruthvik G.(2015). Automated traffic signal for hassle free movement of ambulance. Publié dans IEEE international conference on electrical, computer and communication technologies (icecct).

Weisfeldt M., Becker L. (2002). Resuscitation after cardiac arrest: a 3-phase time-sensitive model, JAMA, vol. 288, no. 23, pp.3035–3038.

Summary

With the increasing evolution of advanced technologies and techniques such as the Internet of Things, AI and Big Data, the traffic management systems industry has acquired new methodologies for creating advanced and intelligent services and applications for traffic management and safety. Our current contribution focuses on the implementation of a path recommendation service for paramedics in emergency situations, which we characterize as one of the most critical and complex issues in traffic management for the survival of individuals involved in emergency incidents. We mainly focused this study on the response time to life-threatening incidents, which is an indicator for emergency ambulance services and for recommending a shorter ambulance route. To this end, we focused on machine learning techniques to predict the fastest route to reach a destination. More specifically, we first propose a local approach based on the KNN technique to predict the congestion of different sections of a map from an origin to a destination in OpenStreetMap. Secondly, we suggest a global approach to show the fastest path to ambulance drivers in real time as they move.

Keywords: Recommendation systems, emergency urban traffic, ambulance mobility, emergency services, navigation services.

Prévision de la consommation d'électricité à l'échelle individuelle dans les secteurs résidentiel et tertiaire

Fatima Fahs*, Frédéric Bertrand*,**, Myriam Maumy*,**.

*IRMA - UMR 7501 - LabEx IRMIA
Université de Strasbourg et CNRS
7 rue René-Descartes, 67084 Strasbourg Cedex
fatima.fahs@unistra.fr, mmaumy@unistra.fr, fbertran@unistra.fr
**UR LIST3N
Université de technologie de Troyes
12 rue Marie Curie, CS 42060, 10004 Troyes Cedex
myriam.maumy@utt.fr, frederic.bertrand@utt.fr

Résumé. La prévision de la demande d'électricité au niveau individuel attire de plus en plus l'attention en raison de son importance dans de nombreuses applications industrielles. Nous aborderons ce sujet ici dans un contexte de services personnalisés de la consommation d'électricité pour des clients des secteurs résidentiel et tertiaire. Dans cet article, nous proposons trois modèles de prévision à court terme de la consommation à l'échelle individuelle au pas demi horaire. Les modèles sont testés sur dix courbes de charge très disparates et une approche de prévision pour les données les plus volatiles sera proposée.

Mots-clés : Modèles de prévision de la consommation d'électricité, échelle individuelle, secteurs résidentiel et tertiaire.

1 Introduction

Pour se positionner sur le marché "Énergie et Utilités" qui est devenu de plus en plus concurrentiel, les entreprises se servent de la digitalisation pour accroître leur compétitivité et valoriser leur potentiel disruptif. La digitalisation a modifié le paysage traditionnel du secteur de l'électricité et a mis au défi les entreprises de toutes tailles avec tout son potentiel de disruption. Les entreprises d'électricité ont augmenté leurs investissements dans les technologies numériques comme les objets connectés (*IoT*), les *smart grids* et les compteurs intelligents (*Linky*) au cours des dernières années. Par exemple, l'investissement mondial dans les infrastructures et les logiciels électriques numériques a augmenté de plus de 20% par an depuis 2014 (IEA (2017)). Les fournisseurs d'énergie investissent des milliards de dollars dans le remplacement des compteurs analogiques par des compteurs intelligents. Ces derniers peuvent mesurer, stocker et transférer des données de consommation d'électricité à haute fréquence. Ils peuvent être installés sur des circuits électriques appartenant à des ménages ou à des bâtiments, et sont également utilisés pour mesurer la consommation d'énergie au niveau des appareils

électriques. Enfin, les compteurs intelligents ont des capacités de communication bidirectionnelle : envoyer et recevoir des informations vers et depuis d'autres appareils. En Europe, le parlement européen a promulgué le déploiement des compteurs intelligents, avec un objectif de l'ordre de 80% des sites en 2020 dans la plupart des pays (De Somer (2018)). Ils aideront les fournisseurs à optimiser la commercialisation d'électricité et informeront les clients en temps quasi réel sur leur consommation. L'essor du digital a transformé les habitudes du consommateur. Il souhaiterait un service entièrement adapté à ses besoins et au pilotage de sa consommation en temps réel ainsi que recevoir des conseils personnalisés et des alertes en cas de dépassement. Les entreprises innovantes qui aborderont la numérisation de manière stratégique et qui investiront davantage dans la personnalisation de la relation clientèle gagneront un avantage concurrentiel significatif, améliorant leur capacité à s'adapter. Aujourd'hui, certains fournisseurs d'électricité proposent à leur clientèle une application, accessible sur différents supports, permettant de suivre leurs consommations et leurs dépenses énergétiques en temps réel. Par exemple, Enedis et EDF disposent notamment de leurs propres applications mobiles. Les fonctionnalités les plus courantes proposées dans ces applications sont celles qui permettent d'identifier les pics de consommation d'électricité, de visualiser en euros et en kWh les consommations avec la possibilité de les comparer par périodes et de fixer des objectifs à atteindre pour pouvoir effectuer des économies sur la facture d'énergie (MaPetiteEnergie, 2020). L'objectif de nos travaux de recherche est d'étudier la possibilité d'intégrer des fonctionnalités prédictives pour chaque client dans ce type d'application de pilotage de la consommation.

2 État de l'art

La prévision de la charge au niveau national est un domaine de recherche très mature. Par conséquent, il existe une vaste littérature décrivant et testant une variété de techniques, y compris les modèles de séries temporelles ((*S*)*ARIMA*), les modèles de lissage exponentiels (Hong, 2010; Taylor, 2010), les modèles d'apprentissage automatique (réseaux de neurones artificiels, *ANN*, et séparateurs à vastes marges, *SVM*, (Kaytez et al., 2015) qui ont la capacité également de modéliser une relation non linéaire entre la charge et les variables météorologiques). Les modèles d'espace d'état (*SSM*) (Dordonnat et al., 2012) ont été également utilisés. Parmi les méthodes non et semi paramétriques, Antoniadis et al. (2012) ont utilisé un modèle de régression basé sur l'estimateur à noyau pour prédire la courbe de charge considérée comme un processus stochastique à valeurs fonctionnelles. Fan et al. (2020) et d'autres chercheurs ont proposé des modèles hybrides pour la prévision de la consommation d'électricité à cette échelle. Certains de ces modèles permettent la quantification de l'influence de la température sur la prévision de la charge (Toros et Aydın, 2019).

Récemment, les données récoltées par les compteurs intelligents, l'émergence d'applications dans le contexte des réseaux intelligents et des maisons intelligentes ainsi que la gestion et la réponse à la demande, ont encouragé le développement d'approches de prévision à l'échelle individuelle allant des ménages jusqu'aux distributeurs. Pourtant, la haute volatilité de la charge d'électricité des ménages rend difficile les prévisions à court terme qui n'ont pas encore été réalisées de manière satisfaisante. Cette charge dépend de plusieurs facteurs comme le nombre et le mode de vie des occupants, leurs appareils électriques, la classe énergétique du logement, les caractéristiques du bâtiment et la météo (Tascikaraoglu et al., 2014). Les fluctuations de

comportement des occupants complique particulièrement la prévision de la charge d'électricité à cette échelle.

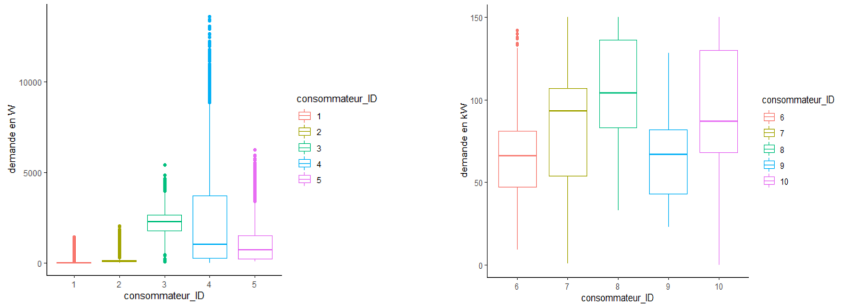
Un article recensant les méthodes les plus récentes d'utilisation des données de compteurs intelligents telles que la prévision, la classification et l'optimisation a été publiée par Yildiz et al. (2017). Il a été observé dans cet article qu'il existe des différences remarquables entre les études publiées à ce sujet non seulement en raison de la différence des applications envisagées mais encore en raison de la diversité des caractéristiques des données utilisées. Dans cet article, la performance de la prévision de la demande est connue pour être assez médiocre au niveau des ménages, les erreurs de pointe varient entre 5% et 60%. Des techniques comme la modélisation de l'activité des ménages associée à des modèles de prévision classiques ont été proposées également afin d'améliorer la performance de ces derniers (Gajowniczek et Ząbkowski, 2017). Bien que ces méthodes se révèlent performantes, elles nécessitent des informations sur les appareils électriques de chaque ménage qui ne sont pas toujours accessibles.

Les approches mises en œuvre pour la prévision de la charge à l'échelle des ménages ne sont pas toutes commercialement viables en raison de leurs complexités (difficile à adapter sur de nouvelles données, non prise en compte des contraintes de temps de calcul et des ressources en mémoire). Parmi le peu d'études sur ce sujet qui se concentrent sur les aspects techniques et qui cherchent à trouver des réponses aux défis du monde réel, nous citons l'article Gerossier et al. (2018). Ce dernier propose un cadre de prévision hiérarchique basé sur un total de cinq modèles probabilistes de complexité variable et évalué sur les données de 20 ménages.

3 Entre les besoins et les contraintes

Le but de ce travail de recherche est d'obtenir un modèle de prévision à $J+1$ fiable et interprétable permettant de prédire la consommation d'électricité du lendemain pour chaque client à partir de son historique de consommation et des données météorologiques. Aucune information sur le client comme le nombre d'appareils électriques et la consommation de ces derniers, le nombre d'occupants dans le logement et la localisation géographique n'est renseignée. Le modèle que nous développerons sera intégré dans une application mobile. Ensuite, la solution envisagée doit également être compatible avec les contraintes industrielles telles que :

- La précision. Le modèle doit donner une analyse prédictive fiable et interprétable pour chaque client.
- La rapidité. Le modèle doit calculer rapidement des prévisions quotidiennes pour de nombreux ménages, en se limitant à l'infrastructure existante et les ressources en mémoire.
- L'adaptabilité et la robustesse. Le modèle doit être facilement adaptable pour de nombreuses typologies de ménages et pour différents profils de courbes de charge.
- La simplicité. Une préférence incontestable pour les modèles explicites qui peuvent être facilement utilisés par des non spécialistes. Ceci exclut ainsi les méthodes d'apprentissage profond pour leur défaut d'interprétabilité.



(a) Boîtes à moustaches pour les courbes de charge résidentielles.

(b) Boîtes à moustaches pour les courbes de charge tertiaires.

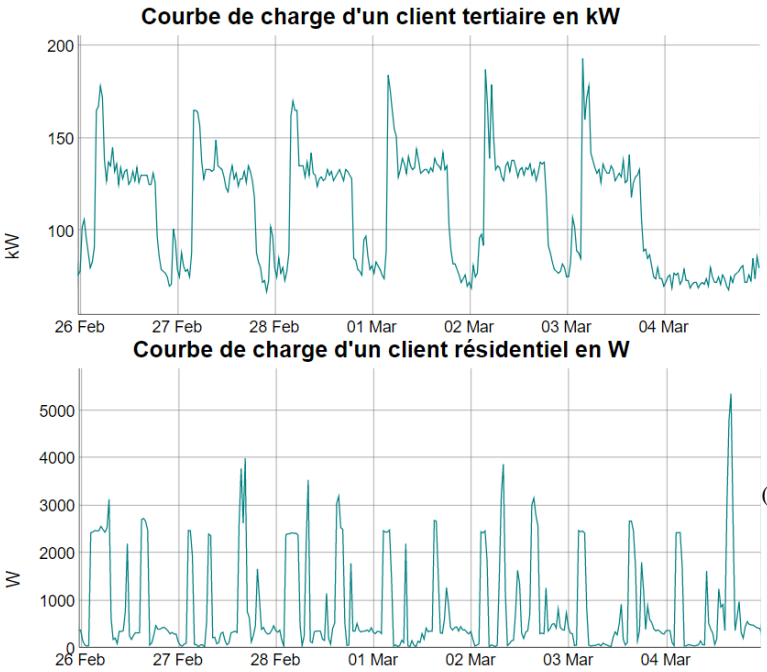
Figure 1: Boîtes à moustaches pour les courbes de charge.

4 Description des données

Nous travaillons sur des données anonymes fournies par une entreprise d'électricité pour différents profils de clients équipés de compteurs intelligents. Cet ensemble de données est confidentiel et ne peut pas être partagé. Il comporte des courbes de charge de dix consommateurs, relevées avec un pas de 30 minutes sur une période de deux ans (2017 à 2019). Cinq courbes (ID : 1, 2, 3, 4 et 5 voir figure 1a) appartiennent à la catégorie des clients résidentiels et les cinq autres (ID : 6, 7, 8, 9 et 10 voir figure 1b) à la catégorie des entreprises tertiaires. 80% des données sont utilisées pour entraîner les modèles de prévision et les 20% restantes pour tester. Une représentation graphique de dix courbes de charge individuelles est présentée dans la figure 1 à l'aide des boîtes à moustaches. Il est important de noter que les profils de courbes de charge électrique résidentielles (voir figure 1a) présentent un nombre important de valeurs aberrantes, représentées par les points en dehors des boîtes de moustaches. En pratique, ils correspondent aux pics de charge et surviennent lors de la mise en marche d'appareils électriques les plus énergivores. De plus, les courbes de charge tertiaires (voir figure 1b) sont moins volatiles en raison de la régularité de l'activité économique dans ce secteur.

Les courbes de charge d'électricité comportent trois cycles saisonniers : journalier, hebdomadaire et annuel. Le cycle journalier est représenté par les pics et les creux de la consommation au cours de la journée. Le cycle hebdomadaire se caractérise par la variation de la consommation entre les jours ouvrés et les week-ends. Le cycle annuel s'identifie par les cycles des saisons et les périodes des vacances (Hahn et al. (2009)). La courbe de charge d'un client tertiaire a un profil journalier marqué par les heures d'ouverture et de fermeture, avec une chute de la demande pendant les week-ends liée à une baisse d'activité (voir la figure 2a). La courbe de charge d'un client résidentiel est beaucoup plus volatile, avec une augmentation de la demande pendant les week-ends (voir la figure 2b).

La variation saisonnière de la consommation d'électricité est fortement liée à l'utilisation du chauffage électrique dans les secteurs résidentiel et tertiaire. Cette thermosensibilité mène à une forte augmentation de la consommation en hiver. Nous distinguons dans notre étude deux profils de clients, les clients thermosensibles (voir la figure 3a) et les clients non thermosensibles (voir la figure 3b).



(a) Client tertiaire.

(b) Client résidentiel.

Figure 2: Exemples de courbes de charge du lundi 26 février au dimanche 4 mars 2018.

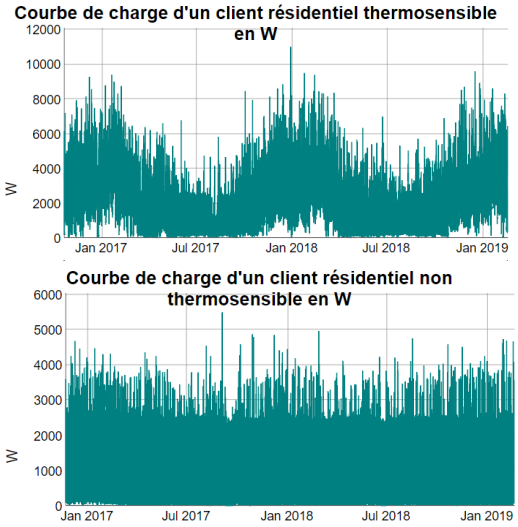
La relation entre la consommation d'électricité et la température est très complexe et non-linéaire. En effet, l'augmentation de la différence entre la température extérieure et intérieure déclenche le démarrage du chauffage ou de la climatisation et par suite augmente la demande d'électricité. De plus, il existe un effet dynamique en raison de l'inertie thermique des logements, la charge au temps t ne dépend pas de la température au temps t uniquement, mais aussi des températures des jours précédents (Le Comte et Warren (1981)).

5 Les modèles de prévision

Nous avons testé plusieurs modèles de régression non paramétrique. Deux versions sont conçues : la première est dédiée à la prévision de la charge des profils thermosensibles en ajoutant la température extérieure comme variable exogène et la deuxième à la prévision de la charge des profils non thermosensibles à partir de l'historique de la consommation uniquement.

5.1 Le modèle *KWF* (*Functional Wavelet Kernel*)

Cette méthode fonctionnelle non paramétrique est basée sur la régression à noyau (Antoniadis et al. (2014)). Elle a été appliquée par EDF au niveau national et implémentée dans le `package enercast` du logiciel R (Cugliari et al., 2017). Le principe général de cette approche consiste à prédire le futur à partir des contextes similaires du passé. La courbe de charge est donc traitée comme un processus à valeurs fonctionnelles, chaque courbe de charge



(a) Client résidentiel thermosensible.

(b) Client résidentiel non thermosensible.

Figure 3: Exemples de courbes de charge de 2017 à 2019.

journalière $(Z_n)_n$ est décomposée dans des bases d'ondelettes. Des poids de similarité $(w_{m,n})$ sont calculés entre les courbes de charge journalières (Z_m) et (Z_n) à partir de leurs coefficients d'ondelettes. La courbe de charge journalière à prédire, Z_{n+1}, est donc estimée par une méthode noyau à partir des courbes de charge journalières similaires parmi Z_1, Z_2,..., Z_n .

$$\widehat{Z_{n+1}} = \sum_{m=1}^{n} w_{m,n} Z_m.$$

La méthode exige la vérification des hypothèses de stationnarité mais ces dernières sont loin d'être vérifiées sur des courbes de charge journalières : premièrement à cause de la variation du niveau moyen de ces courbes avec le temps et deuxièmement l'existence de groupes de jours distincts comme les jours ouvrés et les week-ends. Plusieurs stratégies ont été mises en œuvre pour prendre en compte les diverses sources de non-stationnarité par Cugliari (2011), notamment l'intégration des groupes calendaires (les jours de la semaine et les jours fériés). De plus, une méthode de *clustering* a été proposée par Cugliari (2011) pour intégrer de l'information exogène comme la température extérieure. Cette dernière utilise la distribution d'énergie à travers les échelles pour identifier des *clusters* homogènes des courbes de charge journalières. Enfin elle prédit le futur à partir des jours similaires du même *cluster*. Nous avons utilisé cette dernière méthode pour prédire la charge des clients thermosensibles et le modèle *KWF* avec des groupes calendaires pour les profils non thermosensibles.

5.2 Le modèle *GAM* (*Generalized Additive Model*)

Le modèle additif généralisé, abrégé par *GAM*, a été élaboré par Hastie et Tibshirani (1990). Ce dernier se base sur une somme de P fonctions lisses $f_1, \ldots, f_P$ des variables explicatives

$(X^1, \ldots, X^P)$. L'équation du modèle s'écrit, pour tout temps t :

$$g(\mathbb{E}(Y_t)) = \beta_0 + f_1(X_t^1) + \cdots + f_P(X_t^P) + \epsilon_t$$

où g désigne une fonction de lien et ϵ_t l'erreur du modèle. La fonction lisse f_k, pour tout k variant de 1 à P est la somme des fonctions de la base B de dimension q :

$$f_k(x) = \sum_{i=1}^{q} \beta_i b_i(x) \quad \text{pour tout } \beta_i \text{ réel.}$$

Ce modèle est capable de modéliser à la fois des relations linéaires et des interactions de non-linéarités entre les variables indépendantes ce qui rend le modèle *GAM* semi paramétrique. Le choix des fonctions f_k, qui déterminent le caractère non linéaire du modèle, constitue une véritable problématique. Ces fonctions peuvent être sélectionnées parmi un ensemble de fonctions classiques : les fonctions polynomiales, les fonctions en escaliers ou les *splines*. Ces modèles sont très utilisés dans la littérature car ils modélisent bien la thermosensibilité de la courbe de charge et sont faciles à faire évoluer (Pierrot et Goude, 2011). Afin de détecter des relations complexes entre la consommation et la température extérieure et des interactions non linéaires entre les différentes variables d'influence, nous avons implémenté un modèle *GAM*, pour la prévision des courbes de charge à l'échelle individuelle. Il existe plusieurs méthodes pour estimer le modèle *GAM*. Nous avons choisi d'utiliser la méthode des moindres carrés itératifs repondérés pénalisés (Wood et Augustin, 2002) implémentée dans le `package mgcv` (Wood, 2015) de R. La table 1 présente les différents facteurs impactant la charge électrique au niveau des ménages et les variables les modélisant.

Facteur	Variable
Saisonnalité journalière	Le décalage d'un jour de la courbe de charge Variable catégorielle (indiquant le pas de temps)
Saisonnalité hebdomadaire	Variable catégorielle (type du jour de la semaine)
Saisonnalité annuelle	Coefficients de Fourier
Calendrier	Variable catégorielle (les jours fériés et les jours de pont)
Thermosensibilité	Les températures extérieures lissées (modélisent l'inertie thermique des logements) Les températures actuelles et décalées d'un jour

Table 1: Facteurs impactant la charge électrique au niveau ménage et variables les modélisant.

5.3 Le modèle *MARS* (*Multivariate Adaptive Regression Splines*)

MARS est une méthode de régression non paramétrique (Friedman, 1991). Pour estimer la relation non linéaire entre un vecteur de variables explicatives noté $X = (X^1, ..., X^P)$ et

Y la variable de réponse, un modèle d'estimation flexible est construit à l'aide de fonctions élémentaires linéaires par morceaux (les *splines*) de la forme :

$$(X^j - t)_+ = \begin{cases} X^j - t, & \text{si } x < t \\ 0 & \text{sinon} \end{cases} \quad \text{ou} \quad (t - X^j)_+ = \begin{cases} t - X^j, & \text{si } x < t \\ 0 & \text{sinon.} \end{cases}$$

L'ensemble de ces *splines* forme une base de dimension q notée *B*. Un modèle *MARS* s'écrit :

$$Y_t = f_1(X_t^1) + ... + f_P(X_t^P) + \epsilon_t \quad \text{où} \quad f_k(x_t^k) = \sum_{i=1}^{q} \beta_i B_i(x_t^k) \quad \forall \beta_i \in \mathbb{R}$$

où $B_i(x)$ représente une fonction de l'ensemble *B* ou le produit de deux ou plusieurs de ces fonctions. Ce modèle introduit automatiquement des termes d'interaction entre les variables, *a contrario* des *GAM*. C'est cette différence qui permet aux modèles *MARS* de se démarquer sensiblement des modèles traditionnels. Nous avons testé cette méthode dans le but de détecter des relations complexes entre la consommation et la température extérieure en profitant du processus de sélection automatique des variables indépendantes. Il faut noter que nous avons utilisé pour ce modèle le `package earth` (Milborrow, 2020) de R ainsi que les mêmes variables d'entrée que pour le modèle *GAM* (voir la table 1).

6 Évaluation de la performance des modèles de prévision

Deux métriques sont utilisées pour évaluer la performance des modèles. Nous comparerons les valeurs prédites par chaque modèle avec les observations réelles d'une période de test.
La première métrique est l'erreur d'échelle absolue moyenne notée *MASE* qui est définie par :

$$MASE = \frac{\frac{1}{N}\sum_{t=1}^{N} |Y_t - F_t|}{\frac{1}{T-m}\sum_{t=m+1}^{T} |Y_t - Y_{t-m}|}$$

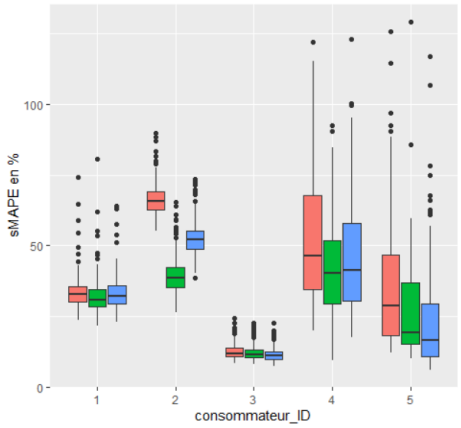
où N représente le nombre de prévisions, Y_t désigne la valeur réelle et F_t la valeur de prévision. Le dénominateur de *MASE* est l'erreur absolue moyenne de la "méthode de prévision naïve saisonnière" qui utilise la valeur réelle de la saison précédente comme prévision. Dans notre cas nous avons choisi la saisonnalité hebdomadaire pour la méthode de prévision naïve.
La deuxième métrique est l'erreur en pourcentage absolue moyenne symétrique notée *sMAPE*. Elle est généralement définie par :

$$sMAPE = \frac{\sum_{t=1}^{N} |Y_t - F_t|}{\sum_{t=1}^{N} (Y_t + F_t)}.$$

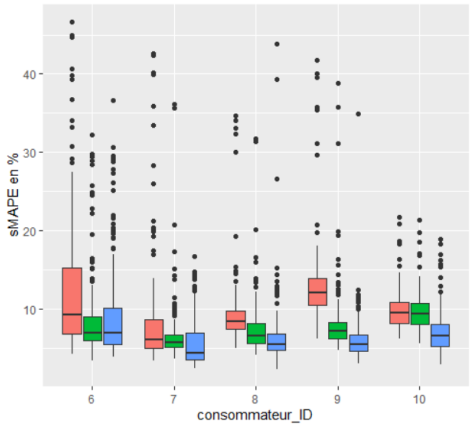
Ces métriques d'erreur sans échelle peuvent être utilisées pour comparer la précision des prévisions entre les séries temporelles de différentes échelles. Lors de la comparaison des méthodes de prévision, la méthode avec la *MASE* la plus basse est la méthode la plus performante.

7 Résultats et discussions

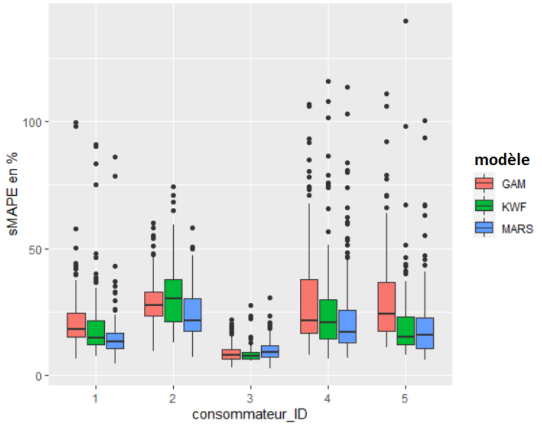
Nous avons testé les trois modèles de prévision à court terme du jour $J+1$ de chacune des dix courbes de charge. Nous avons calculé les erreurs *sMAPE* journalières pour chaque modèle. Les figures 4a et 4b présentent les boîtes à moustaches de l'erreur *sMAPE* journalière pour chaque consommateur et pour la période test allant de septembre 2018 à février 2019.



(a) Courbes de charge résidentielles.



(b) Courbes de charge tertiaires.

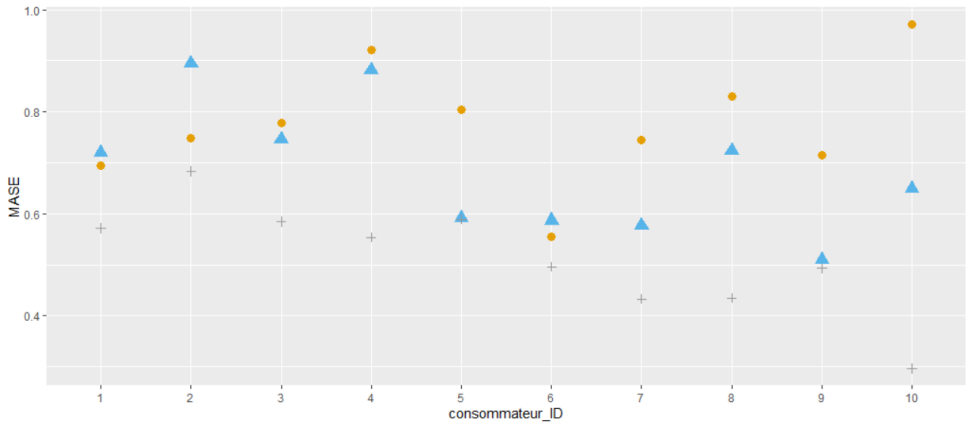


(c) Courbes de charge **cumulée** résidentielles.

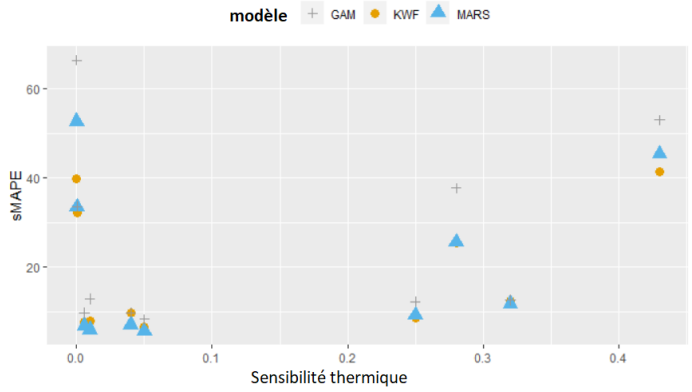
Figure 4: Distribution de l'erreur journalière *sMAPE* de la prévision à $J+1$ par les trois modèles *KWF*, *GAM* et *MARS*.

Ces graphiques révèlent que l'augmentation de la volatilité dans la courbe de charge (ID: 1, 2 et 3 voir la figure 4a) entraîne une augmentation du *sMAPE* de tous les modèles testés et par conséquence les performances de la prévision des courbes de charge tertiaires sont meilleures que celles des courbes de charge résidentielles puisqu'elles sont moins volatiles. La figure 5a montre que les trois modèles testés ont de meilleures performances qu'un modèle de prévision naïf saisonnier qui utilise la valeur réelle de la semaine précédente comme prévision. En

plus, le modèle *KWF* fonctionne mieux pour la prévision des courbes de charge résidentielles non thermosensibles (ID: 1 et 2 voir la figure 4a) alors que le modèle *MARS* a de meilleures performances que les deux autres modèles pour la prévision des courbes de charge tertiaires thermosensibles (ID: 8, 9 et 10 voir la figure 4b).



(a) *MASE* (axe des ordonnées) pour les trois modèles *KWF*, *GAM* et *MARS*.



(b) *sMAPE* (axe des ordonnées) en fonction leur sensibilité thermique (axe des abscisses, sans unité)

Figure 5: Variation des performances de la prévision à $J + 1$.

De plus, la figure 5b montre que les courbes de charge ayant une plus grande sensibilité thermique sont plus difficiles à prévoir. La sensibilité thermique est définie comme étant le carré de la corrélation entre la consommation d'électricité et la température extérieure. Nous constatons aussi, d'après la figure 5b, que les performances des modèles de prévision varient fortement entre les différents profils de courbes de charge ayant une sensibilité thermique similaire. Cela est dû aux comportements aléatoires des consommateurs. Par exemple, le décalage du déclenchement d'un appareil électrique énergivore comme le lave-vaisselle, le lave-linge, dans la journée engendre du bruit dans la courbe de charge. Cela a pour conséquence d'affecter

la performance du modèle de prévision alors que la quantité d'électricité consommée pendant cette journée reste égale à celle consommée pendant une journée similaire dans le passé.

Les résultats de la prévision de ce type de courbe de charge notamment résidentielle sont ainsi en retard par rapport aux résultats obtenus pour la prévision des charges dans le secteur tertiaire où l'activité économique est plus régulière et la charge est moins volatile. Dans le but d'améliorer la performance de la prévision des courbes de charge résidentielles les plus volatiles en restant dans le cadre de proposition de services de prévision fiables aux consommateurs, nous avons testé les modèles de prévision sur la consommation journalière cumulée au pas 30 minutes en kWh à la place de la puissance exprimée en W. Nous cherchons, par cette approche, à lisser le bruit engendré par le décalage des habitudes de consommation d'électricité dans la journée afin d'améliorer la performance de la prévision des courbes de charge résidentielles. Enfin nous fournirons une approche interprétable au consommateur qui s'intéresse plus à la quantité d'énergie consommée à un temps t de la journée au lieu de la puissance. Nous avons testé les trois modèles de prévision sur les données de consommation journalière cumulée en kWh des profils résidentiels. Les résultats de la figure 4c montrent une amélioration importante dans la performance de tous les modèles testés.

8 Conclusion

Dans cet article, nous avons présenté trois modèles pour la prévision à court terme des courbes de charge individuelles au pas demi horaire. Les trois modèles ont été évalués sur cinq courbes de charge résidentielles et cinq courbes de charge issues du tertiaire. Concernant le choix du modèle le plus performant parmi les trois proposés, nous avons conclu que les modèles *KWF* et *MARS* peuvent être utilisés pour la prévision de différentes courbes de charge individuelles en fonction de leurs caractéristiques. Compte tenu de l'hétérogénéité des courbes de charge, il sera intéressant d'explorer la possibilité d'attribuer un modèle par classe de clients à déterminer selon leur profil.

Notre analyse a montré que la performance de la prévision dépend fortement de deux caractéristiques : la volatilité de la courbe de charge et sa sensibilité thermique. La prévision des courbes de charge les moins volatiles est plus performante et les résultats obtenus sont plus satisfaisants notamment pour le secteur tertiaire.

Afin d'améliorer la qualité de la prévision des données les plus volatiles, nous avons proposé une approche de prévision de l'énergie cumulée dans la journée en Kilowatt-heure au lieu de la prévision de la puissance en Watt. Une comparaison de ces deux approches faite sur les données de cinq ménages de profils disparates montre que l'approche d'énergie est plus performante que celle de la puissance, elle permettra de réduire le pourcentage d'erreur dans les prévisions. Par ailleurs, dans la pratique elle est plus facilement interprétable par le consommateur résidentiel. Ceci prouve que les données de la consommation individuelles hautement volatiles peuvent être exploitées afin de fournir des analyses prédictives fiables et interprétables aux consommateurs de leur consommation.

References

Antoniadis, A., X. Brossat, J. Cugliari, et J.-M. Poggi (2012). Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. application à la consommation d'électricité. *Journal de la Société Française de Statistique 153*(2), 52–78.

Antoniadis, A., X. Brossat, J. Cugliari, et J.-M. Poggi (2014). Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité. *Journal de la Société Française de Statistique 155*(2), 202–219.

Cugliari, J. (2011). *Prévision non paramétrique de processus à valeurs fonctionnelles : application à la consommation d'électricité*. Theses, Université Paris Sud - Paris XI.

Cugliari, J., A. Castrillejo, et I. Ramirez (2017). Electrical energy forecast tools. `https://github.com/cugliari/enercast/`.

De Somer, S. (2018). The powers of national regulatory authorities as agents of eu law. In *ERA Forum*, Volume 18(4), pp. 581–595. Springer.

Dordonnat, V., S. J. Koopman, et M. Ooms (2012). Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling. *Computational Statistics & Data Analysis 56*(11), 3134–3152.

Fan, G.-F., X. Wei, Y.-T. Li, et W.-C. Hong (2020). Forecasting electricity consumption using a novel hybrid model. *Sustainable Cities and Society 61*, 102320.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics 19*(1), 1–67.

Gajowniczek, K. et T. Ząbkowski (2017). Electricity forecasting on the individual household level enhanced based on activity patterns. *PloS one 12*(4), e0174098.

Gerossier, A., R. Girard, A. Bocquet, et G. Kariniotakis (2018). Robust day-ahead forecasting of household electricity demand and operational challenges. *Energies 11*(12), 3503.

Hahn, H., S. Meyer-Nieberg, et S. Pickl (2009). Electric load forecasting methods: Tools for decision making. *European journal of operational research 199*(3), 902–907.

Hastie, T. J. et R. J. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC press.

Hong, T. (2010). *Short Term Electric Load Forecasting.* Theses, North Carolina State University.

IEA (2017). Digitalisation and energy. Technology report.

Kaytez, F., M. C. Taplamacioglu, E. Cam, et F. Hardalac (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power & Energy Systems 67*, 431–438.

Le Comte, D. M. et H. E. Warren (1981). Modeling the impact of summer temperatures on national electricity consumption. *Journal of Applied Meteorology and Climatology 20*(12), 1415–1419.

MaPetiteEnergie (2020). Suivi de consommation d'électricité et de gaz : quelles applications ? `https://www.monpetitforfait.com/energie/aides/applications-suivi-conso-energie`. Accessed: 2020-08-18.

Milborrow, S. (2020). earth: Multivariate adaptive regression splines. R package version 5.3.0.

Pierrot, A. et Y. Goude (2011). Short-term electricity load forecasting with generalized additive models. In *16th Intelligent System Applications to Power Systems Conference, ISAP 2011*, pp. 410–415. IEEE.

Tascikaraoglu, A., A. Boynuegri, et M. Uzunoglu (2014). A demand side management strategy based on forecasting of residential renewable sources: A smart home system in turkey. *Energy and Buildings 80*, 309–320.

Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research 204*(1), 139–152.

Toros, H. et D. Aydın (2019). Prediction of short-term electricity consumption by artificial neural networks using temperature variables. *Avrupa Bilim ve Teknoloji Dergisi 14*, 393–398.

Wood, S. N. (2015). Package 'mgcv'. R package version 1.29.

Wood, S. N. et N. H. Augustin (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological modelling 157*(2-3), 157–177.

Yildiz, B., J. I. Bilbao, J. Dore, et A. B. Sproul (2017). Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy 208*, 402–427.

Summary

Forecasting electricity consumption at the individual level has been gaining an increasing attention for its importance in many industrial applications. In this article, we tackle this subject in the context of personalized electricity consumption management services for consumers in residential and tertiary sectors. In this study, we propose three day-ahead forecasting models to predict half-hourly electricity consumption at the individual level. The performance of models is evaluated on real data for 10 disparate load curves from residential and tertiary sectors. We also provide a forecasting approach for the most volatile load curves.
Keywords: Forecasting electricity consumption, individual level, pilot services, residential and tertiary sectors.

APPRENTISSAGE SUPERVISÉ RAPIDE POUR DES DONNÉES TENSORIELLES

Ouafae Karmouda*, Jérémie Boulanger*
Rémy Boyer*

*Université de Lille, CNRS, CRIStAL, 59655 Lille, France,
prenom.nom@univ-lille.fr

Résumé. L'apprentissage supervisé est une tâche majeure dans la classification. Dans notre contexte, nous nous intéressons à la classification à partir de jeux de données de tenseurs d'ordre élevé. La «malédiction de la dimensionnalité» stipule que les complexités en terme de stockage et de calcul augmentent de façon exponentielle avec l'ordre des tenseurs. En conséquence, la méthode de l'état de l'art basée sur la SVD d'ordre supérieur (HOSVD) fonctionne bien mais souffre de limitations sévères en termes de complexité. Dans ce travail, nous proposons une méthode rapide à noyau grassmannien rapide pour l'apprentissage des tenseurs d'ordre élevé basée sur l'équivalence entre les décompositions de Tucker et en Train de Tenseurs. Notre approche est liée au réseaux de tenseurs, dont le but est de décomposer le tenseur initial d'ordre élevé en une collection de tenseurs d'ordre faible (au plus d'ordre 3). Nous montrons sur des données réels que la méthode proposée atteint un taux de classification similaire à la méthode de l'état de l'art mais pour une complexité beaucoup plus faible.

1 Introduction

De nos jours, les données ont besoin de plus en plus de dimensions pour être décrites Cichocki et al. (2015). Une manière naturelle de représenter de telles données consiste à utiliser des tableaux multidimensionnels appelés tenseurs Sidiropoulos et al. (2017). Les tenseurs d'ordre Q Kolda et Bader (2009) sont des tableaux multi-dimensionnelles de dimension Q. Ils généralisent les notions de vecteurs (tenseurs du premier ordre) et de matrices (tenseurs du second ordre). Une décomposition tensorielle importante est la SVD d'ordre élevé (HOSVD) Lathauwer et De Moor (2000) mais cette décomposition souffre de la fameuse " malédiction de la dimensionalité " signifiant que les coûts de stockage et de calcul croissent exponentiellement avec l'ordre du tenseur. Ceci est une limitation sévère pour les tenseurs avec $Q > 3$ Cichocki et al. (2016); Zniyed et al. (2020a).
Dans le cadre de la classification supervisée, les machines à vecteurs support (SVM) Burges (1998); Smola et Schölkopf (2004) ont été largement utilisées en raison de leurs solides fondements théoriques solides, de leurs performances et de leur facilité de mise en œuvre. Bien qu'ils ne traitent que la classification linéaire, ils peuvent être modifiés pour traiter des problèmes non linéaires via les méthodes à noyaux. L'idée principale est de projeter les données

initialement non linéairement séparables dans un espace de dimension supérieure où elles deviennent linéairement séparables en utilisant une application ϕ. En pratique (grâce à l'astuce du noyau), le calcul explicite de Φ n'est pas nécessaire tant qu'une expression pour le noyau : $k(.,.) =< \Phi(.), \Phi(.) >$ existe.

Différents travaux ont opté pour la généralisation des SVMs pour les données tensorielles. Citons Kotsia et Patras (2011) où les paramètres de poids sont supposés suivre le modèle de Tucker. Le problème est formulé afin qu'il puisse être résolu de manière itérative, où à chaque itération les paramètres cherchés correspondent à la projection le long d'un seul mode tensoriel et sont estimés en résolvant un problème de type SVMs. Cependant, STuMs est conçu pour des problèmes de classification linéairement séparables ce qui n'est pas toujours le cas pour la majorité des datasets réels.

Dans Kotsia et al. (2012), le tenseur poids est supposé de rang un pour capturer la structure des données. Cependant, le pouvoir expressif d'un tenseur de rang 1 est restrictif pour de nombreuses données du monde réel. Pour surpasser cette limitation, dans Chen et al. (2019) ils introduisent STTMs pour machines à support de train de tenseurs en remplaçant le tenseur poids dans STMs avec un train de tenseurs. La méthode STTMs est également utilisée dans le cas où les données sont linéairement séparables.

L'utilisation de la HOSVD et SVMs sur les données tensorielles a été introduite dans Signoretto et al. (2011). La méthode proposée montre de bonnes performances de classification avec une complexité de calcul élevée due à l'utilisation du HOSVD. L'idée principale de la méthode proposée, notée FAKSETT (Fast Kernel Subspace Estimation based on Tensor Train Decomposition) est de diminuer cette complexité en utilisant un résultat théorique récent donnant un lien algébrique entre le format Tucker et Tensor-Train Zniyed et al. (2020a).

Dans ce travail, les scalaires seront désignés par des lettres minuscules (*e.g.* : a), les matrices seront désignées par des lettres majuscules (*e.g.* : A) tandis que les tenseurs seront désignés par des lettres calligraphiques (*e.g.* : $\mathcal{A}$). L'ordre d'un tenseur sera généralement noté Q et nous considérerons le cas où $Q > 3$. Le q-ième dépliment d'un tenseur $\mathcal{A}$ est une matrice et sera noté $\mathcal{A}_{<q>}$ dont les éléments sont donnés par :

$$\mathcal{A}_{<q>}(i_q\ ,\ i_1...i_{q-1}i_{q+1}...i_Q) = \mathcal{A}_{i_1\,,\,...\,,\,i_Q}.$$

Le mode-n entre $\mathcal{A}$ et B sera noté $\times_n$ selon l'expression suivante :

$$(\mathcal{A} \times_n B)_{i_1\,,\,...\,,\,i_Q} = \sum_i \mathcal{A}_{...\,,\,i_{n-1}\,,\,i\,,\,...}\ B_{i\,,\,i_n}.$$

Le produit de contraction $\times_n^m$ entre $\mathcal{A}$ et $\mathcal{B}$ est un tenseur construit de manière similaire en faisant la somme sur le n-ième indice sur $\mathcal{A}$ et sur le m-ème indice de $\mathcal{B}$.

2 Noyau pour les données tensorielles

Nous nous concentrons sur les problèmes de classification supervisée où les données sont des tenseurs d'ordre élevé. Les méthodes de classification à noyau Newman et al. (2017); Peeters et al. (2008) nécessitent une mesure de similarité. Il est standard Signoretto et al. (2011) de considérer un noyau entre deux tenseurs $\mathcal{X}$ et $\mathcal{Y}$ comme par exemple le noyau gaussien

RBF :

$$k(\mathcal{X}, \mathcal{Y}) = \exp(-\gamma ||\mathcal{X} - \mathcal{Y}||_F^2) \tag{1}$$

où $\gamma > 0$ est un hyper-paramètre et $||.||_F$ est la norme de Frobenius. Cependant, ce noyau ne prend pas en compte la structure multidimensionnelle des tenseurs de données.
Définition : Décomposition de Tucker. Un tenseur $\mathcal{X}$ suit une décomposition de Tucker (TD) s'il peut être écrit comme suit Lathauwer et De Moor (2000) :

$$\mathcal{X} = \mathcal{G} \times_1 U_1 \times_2 ... \times_Q U_Q \tag{2}$$

où U_q sont de taille $I \times R_q$, $1 \leq q \leq Q$ et $\mathcal{G}$ est le tenseur-coeur de taille $R_1 \times \cdots \times R_Q$. Les rangs multilinéaires (rangs-m) de $\mathcal{X}$ sont le Q -uplet $\{R_1, ..., R_Q\}$.
Définition : HOSVD. Une forme contrainte importante du TD est la HOSVD. Dans cette dernière, les facteurs U_q sont orthonormés et le tenseur-noyau $\mathcal{G}$ est all-orthogonal. Afin de calculer le HOSVD présentée dans l'équation (2), Signoretto et al. (2011) considère les R_q vecteurs singuliers dominants à gauche du q -ième dépliement $\mathcal{X}_{<q>}$. La complexité du HOSVD pour un tenseur d'ordre cubique Q de taille $I_1 \times \cdots \times I_Q$ est $O(QRI^Q)$ où $I = \ max_q\{I_q\}$ et $R = \ max_q\{R_q\}$. La complexité de la HOSVD croît linéairement et exponentiellement par rapport à l'ordre Q. Pour des tenseurs d'ordre assez petit Papastergiou et al. (2018); Makantasis et al. (2018), cette complexité reste acceptable mais cette limitation devient rapidement sévère pour les tenseurs d'ordre élevé ($Q > 3$).

2.1 Similarité basée sur les facteurs HOSVD

Afin de tenir compte de la structure multidimensionnelle des tenseurs, une idée présentée dans Signoretto et al. (2011) consiste à décomposer chaque tenseur en sa HOSVD :

$$\mathcal{X} = \mathcal{G} \times_1 U_1 \times_2 ... \times_Q U_Q, \tag{3}$$

$$\mathcal{Y} = \mathcal{H} \times_1 V_1 \times_2 ... \times_Q V_Q. \tag{4}$$

La fonction noyau de la méthode proposée est donnée par :

$$k(\mathcal{X}, \mathcal{Y}) = \prod_q^Q k_q\left(U_q, V_q\right) \tag{5}$$

où $k_q(.,.)$ est un noyau défini positif sur $\mathbb{R}^{I \times R_q} \times \mathbb{R}^{I \times R_q}$.

2.2 Noyau sur une variété Grassmannienne

Notons que les décompositions de l'équation (3) ne sont pas uniques Kolda et Bader (2009). La classification sera affectée par cette non-unicité. Pour éviter ce problème dans le contexte de l'apprentissage, on peut considérer les sous-espaces engendrés par les facteurs $\{U_q, ..., U_q\}$ et $\{V_1, ..., V_Q\}$. En effet, le sous-espace engendré par le facteur $U_q, q \leq 1 \leq N$ (respectivement V_q) est invariant à toute multiplication à droite par une matrice non singulière.
Par conséquent, considérons les sous-noyaux k_q ayant la forme suivante :

$$k_q\left(U_q, V_q\right) = \tilde{k}_q\left(\text{span}(U_q), \text{span}(V_q)\right) \tag{6}$$

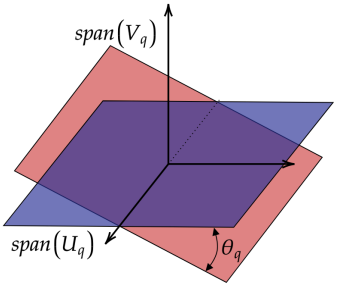


FIG. 1 – *Illustration de l'angle θ_q figurant dans l'équation* (7)

où $\tilde{k}_q$ est un noyau défini sur la variété Grassmannienne $\mathcal{G}(R_q, I)$, *i.e.* les sous-espaces de $\mathbb{R}^I$ de dimension R_q.
Un choix populaire pour $\tilde{k}_q$ qui donne lieu à un noyau défini positif et utilisé dans Signoretto et al. (2011) est donné par :

$$k_q\left(U_q, V_q\right) = \exp\left(-\gamma \sin^2\left(\theta_q\right)\right) \tag{7}$$

où θ_q est l'angle principal entre span(U_q) et span(V_q) (cf.figure. 1). Il convient de bien noter que θ_q est la distance géodésique entre les deux sous-espaces. Le lecteur intéressé peut se référer à Jayasumana et al. (2015) pour des méthodes explicites de calcule des angles principaux. Dans notre cas, il est possible d'utiliser directement les projecteurs :

$$\sin^2\left(\theta_q\right) = 2\left|\left|U_q U_q^T - V_q V_q^T\right|\right|_F^2. \tag{8}$$

3 La méthode FAKSETT : une alternative rapide à la méthode de [9]

Afin de réduire la complexité de la HOSVD discutée dans la section 2.1, nous proposons d'utiliser une méthode de projection multilinéaire rapide proposée dans Zniyed et al. (2020a). Les fondements théoriques de cette méthode sont basés sur la théorie des réseaux de tenseurs Cichocki et al. (2016) et en particulier sur l'équivalence entre les formats de Tucker et du Train de Tenseurs introduits dans Zniyed et al. (2020a) et décrits ci-après. Nous commencerons par une définition de la décomposition en Train de Tenseurs (TTD).

Définition : La décomposition en Trains de Tenseurs (TTD). Un Tenseur $\mathcal{X}$ admet une TTD avec des rangs TT $(R'_1, \ldots, R'_{Q-1})$ s'il peut être exprimé comme suit :

$$\mathcal{X} = G_1 \times_2^1 \mathcal{G}_2 \cdots \times_{Q-1}^1 \mathcal{G}_{Q-1} \times_Q^1 G_Q \tag{9}$$

où :
— $G_1 \in \mathbb{R}^{I \times R'_1}$
— $\mathcal{G}_q \in \mathbb{R}^{R'_{q-1} \times I \times R'_q},\ \forall q : 1 < q < Q$

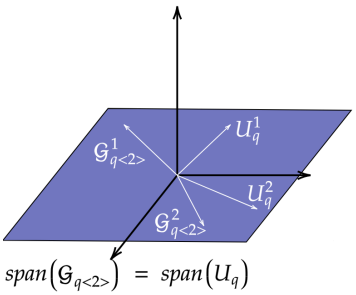


FIG. 2 – *Dans le cas où $R_q = 2$ avec $U_q = [U_q^1, U_q^2]$: Malgré des facteurs différents, HOSVD et FAKSETT donnent des facteurs qui engendrent le même sous-espace.*

— $G_Q \in \mathbb{R}^{R'_{Q-1} \times I}$

Pour estimer les tenseurs-cœurs, nous pouvons utiliser l'algorithme TT-SVD Oseledets (2011) ou sa généralisation TT-HSVDZniyed et al. (2020b).
Supposons qu'un tenseur $\mathcal{X}$ suit une TD de m-rangs $\{R_1, \cdots R_Q\}$ avec des facteurs orthonormaux U_q. La complexité de la TT-SVD est évaluée à $O(R'^Q)$ where $R' = \max_q R'_q$. Une équivalence entre la TD et la TTD est présentée dans Zniyed et al. (2020a). Chaque noyau extrait de la TD donné par l'équation (9) suit un modèle de Tucker d'ordre 3 avec deux matrices latentes dans ses première et troisième dimensions. Dans la deuxième dimension, nous avons la propriété intéressante que les R_q vecteurs singuliers dominants du deuxième dépliement engendrent le même sous-espace que U_q. En conséquence, nous avons pour $2 \leq q \leq Q-1$:

$$\text{span}(U_q) = \text{span}\left(\mathcal{G}_{q<2>}\right). \tag{10}$$

et $\text{span}(U_1) = \text{span}(G_1)$, $\text{span}(U_Q) = \text{span}(G_Q^T)$.
Cette propriété est illustrée sur la figure. 2 Par conséquent, l'expression de l'équation (7) peut être obtenue à partir du calcul des vecteurs singuliers dominants gauches grâce à la SVD associé aux $(Q-2)$ tenseurs-coeurs avec une complexité moindre d'un facteur Q qui devient rapidement très significative quand le l'ordre des tenseurs devient élevé.
Les rangs multi-linéaires et les rangs TT vérifient la relation suivante :

$$R'_q = \min\left(\prod_{p=1}^{q} R_p, \prod_{p=q+1}^{Q} R_p\right).$$

La dernière étape de la méthode FAKSETT est de calculer le noyau défini dans les équations (5) and (7).

4 Experiences numériques

Pour les bases de données suivantes, la tâche de classification est réalisée via SVMs Burges (1998). Ceci repose sur la matrice de similarité obtenue à l'aide du noyau défini précédemment.

FIG. 3 – *Trois classes de la base de données Extended Yale B.*

FIG. 4 – *Deux classes de la base de données UCF11.*

Le noyau est calculé avec FAKSETT et est comparé à la méthode de l'état de l'art Signoretto et al. (2011).

4.1 Bases de données

— **UCF11** : Cette base de données Liu et al. (2009) contient 1600 clips vidéo contenant 11 actions humaines telles que : *plonger*, *sauter*, *marcher*. Deux actions humaines sont choisies : *sauter* et *marcher* et sont présentées dans la figure. 4. . Pour chaque vidéo, nous considérons une séquence de 240 images couleur où la résolution de chaque image est de 320 $\times 240 \times 3$. Ces vidéos peuvent être interprétés comme des tenseurs d'ordre 4 de dimensions 240 $\times 240 \times 320 \times 3$. Un total de 109 tenseurs sont présents dans chaque classe. On sélectionne aléatoirement 60 $\%$ de la taille de la base de donnée pour constituer l'ensemble d'apprentissage. Le reste est laissé pour le test.

— **Extended Yale B** : Cette base de données Georghiades et al. (2001) contient 28 images faciales. Pour chaque sujet, il y a 576 images de taille 480×640 prises sous 9 poses. Chaque pose est prise sous 64 illuminations différentes. Dans ce cas, 3 sujets sont arbitrairement choisis et sont représentés dans la figure.3. Ceci représente 3 classes pour un problème de classification. Afin de construire l'ensemble d'entraînement et l'ensemble de test, nous décomposons le tenseur de chaque sujet en 16 tenseurs en considérant chaque 4 illuminations dans un tenseur de taille $9 \times 480 \times 640 \times 4$.

4.2 Résultats de classification

Dans cette section, nous présentons des expériences numériques dans lesquelles nous utilisons comme mesure de performance le nombre de données correctement classifiées sur le nombre total de données de test.
Le paramètre de régularisation du SVM et de la bande passante du noyau γ de l'équation (7) sont sélectionnés dans la grille de valeurs $\{2^{-9}, 2^{-8}, ..., 2^{8}, 2^{9}\}$ par une validation croisée à 5 blocs. Toutes les expériences sont menées sur un ordinateur équipé d'un processeur Intel Core

i7 de 9e génération 2,6 GHZ et d'une mémoire RAM de 32 Go exécutant Windows 10. Les calculs des SVDs sont réalisés à l'aide de la bibliothèque optimisée TensorLy (Tensor Learning in Python).

— Les Tables 1 et 2 montrent des scores de classification très proches entre FAKSETT et de la méthode de Signoretto et al. (2011) pour les tâches de classification considérées sur les deux bases de données réelles. Cela indique que la méthode FAKSETT fonctionne aussi efficacement que la méthode de l'état de l'art Signoretto et al. (2011). La réduction de la taille de l'ensemble de données d'entraînement (*i.e.* entraînement avec moins de données) n'a pas d'impact significatif sur les performances.

— Cependant, on observe dans la table 3 que FAKSETT réduit significativement le temps d'exécution pour le calcul des facteurs, bien que l'ordre des tenseurs considéré est $Q = 4$. Les tenseurs d'ordre plus élevées conduiraient à un gain de temps encore plus conséquent entre les deux méthodes.

s%	m-ranks	FAKSETT	Méthode de Signoretto et al. (2011)
%50	[2,2,2,2]	$0.72(10^{-2})$	$\mathbf{0.73(10^{-2})}$
%60	[3,3,3,3]	$\mathbf{0.7(10^{-2})}$	$\mathbf{0.7(10^{-2})}$
%80	[3,3,3,3]	$0.76(10^{-2})$	$\mathbf{0.77(10^{-2})}$

TAB. 1 – *Moyenne (écart type) sur les données de test pour la base de données UCF11.*

s%	m-ranks	FAKSETT	Méthode de Signoretto et al. (2011)
%50	[1,3,2,1]	$0.98(10^{-2})$	$\mathbf{0.99(10^{-2})}$
%60	[1,2,2,1]	$\mathbf{0.99(10^{-2})}$	$\mathbf{0.99(10^{-2})}$

TAB. 2 – *Moyenne (écart type) sur les données de test pour la base de données Extended Yale B*

Database	m-ranks	FAKSETT	Méthode de Signoretto et al. (2011)
UCF11	[2,2,2,2]	$\mathbf{14(0.42)}$	$69(3)$
	[3,3,3,3]	$\mathbf{15(0.63)}$	$104(5)$
Extended Yale	[1,2,2,1]	$\mathbf{2.56(0.09)}$	$9.47(0.1)$

TAB. 3 – *Temps moyen (écart type) en secondes pour calculer la HOSVD sur différentes bases de données pour à différentes valeurs de rangs multi-linéaires.*

5 Conclusion

Dans cet article, les SVMs à noyau pour les données tensorielles basées sur une métrique grassmannienne entre des sous-espaces extraits de HOSVD a été proposé. Malgré de bons taux de classification obtenus sur des bases de données réelles, cette méthode souffre d'une complexité élevée, en particulier pour les jeux de données associés à des tenseurs d'ordre Q lorsque $Q > 3$. Dans ce travail, nous avons exploité un lien algébrique récent entre la

HOSVD et le Trains de Tenseurs pour accélérer la méthode de l'état de l'art. Nous appelons ce nouveau schéma d'apprentissage supervisé FAKSETT (pour Fast Kernel Subspace Estimation based on Tensor-Train decomposition). Sur des bases de données réelles, nous avons montré que le schéma FAKSETT atteint des résultats de classification très similaire à la méthode de l'état de l'art mais pour un temps d'exécution considérablement réduit sur des ensembles de données réels utilisés. D'autres décompositions tensorielles seront plus convenables pour plusieurs d'autres applications dans le monde réel et seront exploitées dans de futurs travaux.

Références

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery 2*(2), 121–167.

Chen, C., K. Batselier, C.-Y. Ko, et N. Wong (2019). A support tensor train machine. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Cichocki, A., N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, et D. P. Mandic (2016). Tensor networks for dimensionality reduction and large-scale optimization : Part 1 low-rank tensor decompositions. *Foundations and Trends® in Machine Learning 9*(4-5), 249–429.

Cichocki, A., D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, et H. A. Phan (2015). Tensor decompositions for signal processing applications : From two-way to multiway component analysis. *IEEE Signal Processing Magazine 32*(2), 145–163.

Georghiades, A. S., P. N. Belhumeur, et D. J. Kriegman (2001). From few to many : Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence 23*(6), 643–660.

Jayasumana, S., R. Hartley, M. Salzmann, H. Li, et M. Harandi (2015). Kernel methods on riemannian manifolds with gaussian rbf kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(12), 2464–2477.

Kolda, T. G. et B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*(3), 455–500.

Kotsia, I., W. Guo, et I. Patras (2012). Higher rank support tensor machines for visual recognition. *Pattern Recognition 45*(12), 4192–4203.

Kotsia, I. et I. Patras (2011). Support tucker machines. In *CVPR 2011*, pp. 633–640.

Lathauwer, L. et B. De Moor (2000). A multi-linear singular value decomposition. *Society for Industrial and Applied Mathematics 21*, 1253–1278.

Liu, J., J. Luo, et M. Shah (2009). Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1996–2003.

Makantasis, K., A. D. Doulamis, N. D. Doulamis, et A. Nikitakis (2018). Tensor-based classification models for hyperspectral data analysis. *IEEE Transactions on Geoscience and Remote Sensing 56*(12), 6884–6898.

Newman, E., M. Kilmer, et L. Horesh (2017). Image classification using local tensor singular value decompositions. *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*.

Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing 33*(5), 2295–2317.

Papastergiou, T., E. Zacharaki, et V. Megalooikonomou (2018). Tensor decomposition for multiple-instance classification of high-order medical data. *Complexity 2018*, 1–13.

Peeters, T., A. Vilanova, et B. ter Haar Romeny (2008). *Analysis of Distance/Similarity Measures for Diffusion Tensor Imaging*, pp. 113–136. Springer.

Sidiropoulos, N. D., L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, et C. Faloutsos (2017). Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing 65*(13), 3551–3582.

Signoretto, M., L. De Lathauwer, et J. A. K. Suykens (2011). A kernel-based framework to tensorial data analysis. *Neural networks : the official journal of the International Neural Network Society 24 8*, 861–74.

Smola, A. J. et B. Schölkopf (2004). A tutorial on support vector regression. *Statistics and computing 14*(3), 199–222.

Zniyed, Y., R. Boyer, A. L. De Almeida, et G. Favier (2020a). High-order tensor estimation via trains of coupled third-order cp and tucker decompositions. *Linear Algebra and its Applications 588*, 304 – 337.

Zniyed, Y., R. Boyer, A. L. de Almeida, et G. Favier (2020b). A tt-based hierarchical framework for decomposing high-order tensors. *SIAM Journal on Scientific Computing 42*(2), A822–A848.

Amélioration de l'entreposage des données spatio-temporelles massives

Hanen Balti*,** Nedra Mellouli** Ali Ben Abbes*, Imed Riadh Farah* Myriam Lamolle**
Yangfang Sang***

*Laboratoire RIADI, Université de la Manouba, La Manouba, Tunisie
hanen.balti@ensi-uma.tn,
ali.benabbes@yahoo.fr,
imedriadh.farah@isamm.uma.tn,
**Laboratoire LIASD, Université Paris 8, Paris, France
n.mellouli@iut.univ-paris8.fr,
m.lamolle@iut.univ-paris8.fr
***Key Laboratory of Water Cycle and Related Land Surface Processes,
Institute of Geographic Sciences and Natural Resources Research,
Chinese Academy of Sciences, Pékin, Chine
sangyf@igsnrr.ac.cn

Résumé. Ici, nous proposons un entrepôt de données volumineuses et des analyses avancées pour la surveillance de la sécheresse. Afin de soutenir le processus de prise de décision, un cadre d'entreposage de données volumineuses a été conçu et mis en œuvre. Le cadre proposé se compose de 4 étapes principales : la collecte des données, le pré-traitement des données, le stockage des données et l'étape de prise de décision. Les données utilisées sont multi-sources et hétérogènes. Cela permet donc d'offrir une vue des différentes dimensions telles que la distribution spatiale, la distribution temporelle et la détection de la gravité de ce phénomène. Nous évaluons également l'efficacité de notre cadre proposé dans une étude de cas réelle de surveillance de la sécheresse en Chine de 2000 à 2020. En outre, nous présentons le retour de connaissance de certaines requêtes et de différentes visualisation.

Mots-clés : Analyse des données massives, Entrepôt de données, Stockage, Sécheresse, Hive

1 Introduction

Les données massives d'observation de la Terre (DOT) sont des données complexes vérifiant les propriétés des données massives. Elles s'articulent autour des sciences de la Terre et comprennent principalement les données d'observation de la Terre, par exemple les données de télédétection, les données météorologiques, les données biophysiques, les données atmosphériques, l'activité dérivée de l'homme, etc. (Merritt et al., 2018). Les DOT sont caractérisées comme étant massives, multi-sources, hétérogènes, multi-temporelles, multi-scalaires, hautement dimensionnelles, hautement complexes, non-stationnaires et non structurées. Il s'agit

donc de toutes les données relatives à la Terre, y compris les composantes de la Terre, la surface, l'atmosphère et l'environnement proche de l'espace. Les big data d'observation de la Terre sont caractérisés par 4 V : le volume se réfère à la quantité de données collectées ; plus de 50 000 TeraBytes (TB) de données ont été stockés en 2015, et le volume atteindra 350 000 TB d'ici 2030 (Balti et al., 2020), la vitesse se réfère aux taux auxquels les nouvelles données sont générées ou aux taux auxquels les données sont traitées ; la variété se réfère au fait que les big data de la Terre sont multi-sources, hétérogènes, multi-temporelles, multi-scalaires, non-stationnaires, et non-structurées ; et la véracité se réfère à la qualité globale des données disponibles. La qualité des données peut être affectée par du bruit ou des anomalies dans le processus original de collecte des données (Balti et al., 2020). Il est donc important de concevoir et de mettre en œuvre un entrepôt de données (DW) efficace pour la gestion des catastrophes. Un DW définit une représentation uniforme des données par le biais de son modèle de schéma et stocke les ensembles de données dérivés afin qu'ils puissent être analysés pour en extraire des connaissances utiles. Il existe essentiellement trois modèles possibles pour organiser les données stockées dans un entrepôt de données : les modèles en étoile, en flocon de neige et en constellation. Ces modèles sont principalement composés de faits et de dimensions. Le tableau des faits aide l'utilisateur à analyser les dimensions du problème, pour prendre des décisions afin d'améliorer ses résultats. En outre, les tableaux multidimensionnels aident à rassembler les dimensions avec lesquelles les mesures doivent être prises (Liu et al., 2012). La principale différence entre le tableau de faits et le tableau de dimensions est que le tableau multidimensionnel contient des attributs avec lesquels les mesures sont prises dans le tableau de faits. Par conséquent, la modélisation en étoile est le modèle le plus simple. Il se compose d'une table de faits et de plusieurs tables multidimensionnelles. Le schéma flocon de neige est un type de schéma en étoile qui inclut la forme hiérarchique des tables de dimension. Dans ce modèle, il existe une table de faits composée de différentes tables de dimensions et de sous-dimensions liées par des clés primaires et étrangères à la table de faits. Le fractionnement permet de réduire les redondances et d'éviter les pertes de mémoire. Un schéma en flocon de neige est plus facile à gérer mais complexe à concevoir et à comprendre. Le schéma en constellation de faits comporte plusieurs tables de faits partageant des tables multidimensionnelles. Ce modèle est plus complexe que les modèles en étoile et en flocon de neige. Malgré les avantages qu'offre le DW, il y a un manque de littérature qui relève de la conception du DW dans le but de permettre la gestion des catastrophes, considérée incontestablement comme de l'analyse et l'exploitation massives des données de crise. La conception de DW à grande échelle est très difficile. En effet, les DOT sont spatiales, temporelles, complexes, hétérogènes, hautement dimensionnelles et collectées à partir de sources multiples. De toute évidence, elles sont constituées des dimensions du Big Data : volume, variété, vélocité et véracité. Par conséquent, les sources de données d'observation de la terre sont très diversifiées et ont différents niveaux de qualité.

Cet article aborde certaines questions relatives aux systèmes d'entreposage de données volumineuses pour la gestion des catastrophes en utilisant des données d'observation de la terre. Les principales contributions sont : (i) l'intégration de nouveaux DW adaptés aux données massives ; (ii) proposition d'une manière adéquate le stockage et le traitement des données ; (iii) interrogation et fourniture d'une aide à la décision interprétable pour la surveillance de la sécheresse. Ce document est organisé comme suit : Dans la section 2, nous présentons l'état de l'art de l'utilisation des big data et des entrepôts de données dans plusieurs domaines. Dans

la section 3, nous présentons la méthodologie proposée. Dans la section 4, nous présentons l'expérimentation, les résultats et la discussion. Nous concluons ce travail dans la section 5 et nous présentons les perspectives imminentes et plus à moyen et court termes.

2 Etat de l'art

Dans la littérature, plusieurs travaux ont utilisé les données massives et les entrepôts de données dans différents domaines. (Sebaa et al., 2018) a proposé une architecture basée sur Hadoop et un modèle conceptuel de données pour élaborer un entrepôt de données massives médicales et a fourni un détail de mise en œuvre de l'entrepôt avec l'écosystème Hadoop afin d'assurer une allocation optimale des ressources de santé. (Ngo et al., 2020) ont proposé un entrepôt de données agricoles au niveau continental, ils évaluent sa performance et présentent quelques requêtes pour extraire des connaissances sur la gestion des cultures. (Jenhani et al., 2019) a proposé une architecture hybride basée sur Storm et Hadoop pour l'extraction d'événements structurés à partir de données de médias sociaux et leur intégration dans l'entrepôt de données. Ils tirent parti de l'analyse en temps réel des données en continue de Storm et du traitement par lots des quantités massives de données de Hadoop, permettant de relever le défi de l'analyse des données de médias sociaux en flux continu. (Ngo et Kechadi, 2020) a proposé une méthode d'intégration des données agricoles utilisant un schéma de constellation conçu pour être suffisamment flexible incorporant d'autres ensembles de données et modèles. Ils ont extrait des connaissances en vue d'améliorer le rendement des cultures. Il s'agit notamment de trouver des quantités appropriées de propriétés du sol, d'herbicides et d'insecticides pour à la fois augmenter le rendement des cultures et protéger l'environnement. (Liu et al., 2012) a proposé un modèle de DW en couches, donnant un processus ETL efficace pour l'intégration des données et concevant deux types d'interfaces. Cette étude en particulier a prouvé qu'un entrepôt de données est une solution efficace pour les DOT.

Dans la littérature, plusieurs outils et architectures ont été proposés pour les entrepôts de données volumineuses dans plusieurs domaines comme Hive, et Oracle. Pour supporter le traitement distribué et parallèle, Apache Hadoop a largement déployé des DW dans de nombreux domaines tels que l'éducation, la médecine, et l'agriculture. Aussi, Hadoop est capable de traiter une grande quantité de données hétérogènes à une vitesse très importante. Ensuite, plusieurs types de schémas DW ont été utilisés dans ces travaux. Le choix du schéma dépend principalement de l'objectif de l'application et de la quantité de données. A notre connaissance, les entrepôts de Big data pour le suivi de la sécheresse n'ont pas été proposés dans la littérature. Ainsi, dans ce travail, nous proposons un entrepôt de données pour la surveillance de la sécheresse.

3 Approche proposée

L'objectif de cette méthodologie est de proposer un entrepôt de données pour la surveillance de la sécheresse afin de stocker l'énorme quantité de données multi-sources (par exemple, les données climatiques, les données de télédétection, les données hydrologiques). La méthodologie proposée consiste en 4 étapes principales : La première étape est la collecte

des données, la deuxième étape est le pré-traitement des données, puis vient le stockage des données et enfin nous avons l'étape de visualisation et de prise de décision. La figure 1 représente le flux de travail de l'architecture proposée.

3.1 Collection des données

La collecte de données consiste à générer et à collecter des données à partir de différentes ressources (Balti et al., 2020). La sécheresse est l'interaction de plusieurs types de facteurs tels que les données de télédétection, les données climatiques et les données géographiques du sol (Mishra et Singh, 2011). Ces données sont massives et hétérogènes. Les données de télédétection sont : indice normalisé de végétation du sol (NDVI), la température de la surface (LST)) ; les données climatiques sont : indice normalisé de l'évapotranspiration des précipitations(SPEI), l'évapotranspiration(ETP), humidité, précipitations, vitesse du vent , pression ; les données géographiques du sol en particulier humidité. Ces données sont un mélange de données structurées, semi- ou non-structurées. Elles sont également multidimensionnelles dont les dimensions peuvent être par exemple, multi-spectrales, multi-résolution, multi-temporelles. Cette couche englobe différentes sources de données pertinentes pour les observations de la Terre et traite différents types de formats de données (c'est-à-dire : NetCDF, CSV, hdr). Chaque mois, le volume de données augmente considérablement. Ainsi, chaque jour, des gigaoctets de données sont générés à partir de différentes sources. Par exemple, les données de télédétection (.i.e. NDVI, LST) sont formatées en différents formats tels que .hdr, alors que les données TRMM et les données géographiques du sol sont formatées en .NetCDF et les données climatiques sont formatées en .csv. La variété des sources de données présentées dans cette couche témoigne de l'hétérogénéité liée à leurs sources d'acquisition et les gestionnaires associés aux systèmes de stockage déployés. Le Tableau 1 décrit la quantité et la temporalité des données.

Données	Résolution Spatiale	Résolution temporelle	Volume	Variété	Véracité	Vélocité
LST	1km	8 jours	x	-	x	x
NDVI	1km	16 jours	x	-	x	x
Précipitation	0.25°x0.25°	Journalier	-	-	-	x
Données climatiques	-	Journalier	-	x	x	x
Humidité du sol	0.25°x0.25°	-	-	-		x

TAB. 1 – *Description des données.*

3.2 Pré-traitement des données

Les données collectées proviennent de différentes sources. Elles contiennent donc différents types d'imperfections. Par exemple, les images satellites peuvent présenter des distorsions géométriques et des bruits atmosphériques. Les données climatiques peuvent inclure des valeurs erronées. Ainsi, il est important d'effectuer différentes opérations sur ces données pour améliorer leur qualité en appliquant plusieurs opérations telles que le mosaïque, la correction des données et la récupération des données brutes (Balti et al., 2020). L'opération de mosaïque est appliquée aux données de télédétection. Les images collectées sont une quantité massive de tuiles couvrant la zone d'étude. Le but de cette opération est de reconstruire une image

satellite composée de nombreuses bandes. L'opération de correction des données consiste en trois types de correction différents. La correction géométrique consiste à éviter les distorsions géométriques d'une image déformée, en établissant la relation entre le système de coordonnées de l'image et le système de coordonnées géographiques en utilisant les données de calibration du capteur. La correction atmosphérique consiste à récupérer la réflexion de surface qui caractérise les propriétés de la surface à partir des données de télédétection en supprimant les effets atmosphériques et la correction de valeur qui consiste à corriger les valeurs erronées ou à identifier les valeurs manquantes. Enfin, l'extraction des données brutes consiste à extraire des informations précieuses de différentes sources d'information telles que le calcul du NDVI, du LST, du SPEI et de l'ETP.

3.3 Stockage des données

L'entrepôt de données est basé sur une conception simple et efficace pour l'analyse des grandes données d'observation de la terre sous la forme d'un modèle multidimensionnel. Dans cet article, un schéma en flocon de neige est proposé. Le schéma en flocon de neige présente plus de détails sur les données que le schéma en étoile. Il offre la possibilité d'utiliser des requêtes plus complexes, ce qui signifie qu'il prend en charge des analyses puissantes et des relations de type plusieurs à plusieurs. La fig.1 est composée d'une table de faits et de 13 tables de dimensions D=(Product_Dimension, Sensor_Dimension, Image_Dimension, SatelliteFeature_Dimension, Drought_Index_Dimension, ClimateStation_Dimension, Date_Dimension, ClimateFeature_Dimension, BiophysicalFeature_Dimension, BiophysicalStation_Dimension, Location_Dimension, Country, Province).
Le DW est présenté par : (F, D, HD_i) où :

- F : la table des faits
- D : {D_1, ..., D_n} : fait référence aux dimensions définies ci-dessous, n étant le nombre de tableaux multidimensionnels
- HD_i{} : fait référence aux hiérarchies pour chaque dimension Di définie par HD_i={h_1, ..., h_k} avec k le nombre de hiérarchies pour chaque dimension.

La table des faits : Chaque fait est défini par F : (NameF, M{}) où :

- NameF : est le nom du fait
- M{} = (m_1, ..., m_n) fait référence aux mesures

Le tableau des dimensions :
Une dimension est définie par D : (NomD, A{}, H{}, TypeD) où :

- NameD : nom de la dimension
- A{}=(a_1, ..., a_l) est un ensemble d'attributs
- H{}=(h_1, ..., h_z) est un ensemble de hiérarchies
- TypeD [T, S] : une dimension peut être temporelle ou spatiale.

La mesure :
Une mesure M est définie par M : (NameM, TypeM, FuncM) où

- NameM : nom de la mesure
- TypeM : type de la mesure
- FuncM : ensemble de fonctions d'agrégation compatibles avec la propriété de compression de la mesure où $FuncM \subset SUM, AVG, MAX, MIN$..... Dans notre cas, la table de faits s'appelle OperationFact, Sensor_Dimension est un exemple des tables de dimensions et la mesure est par exemple AVG_TEMP().

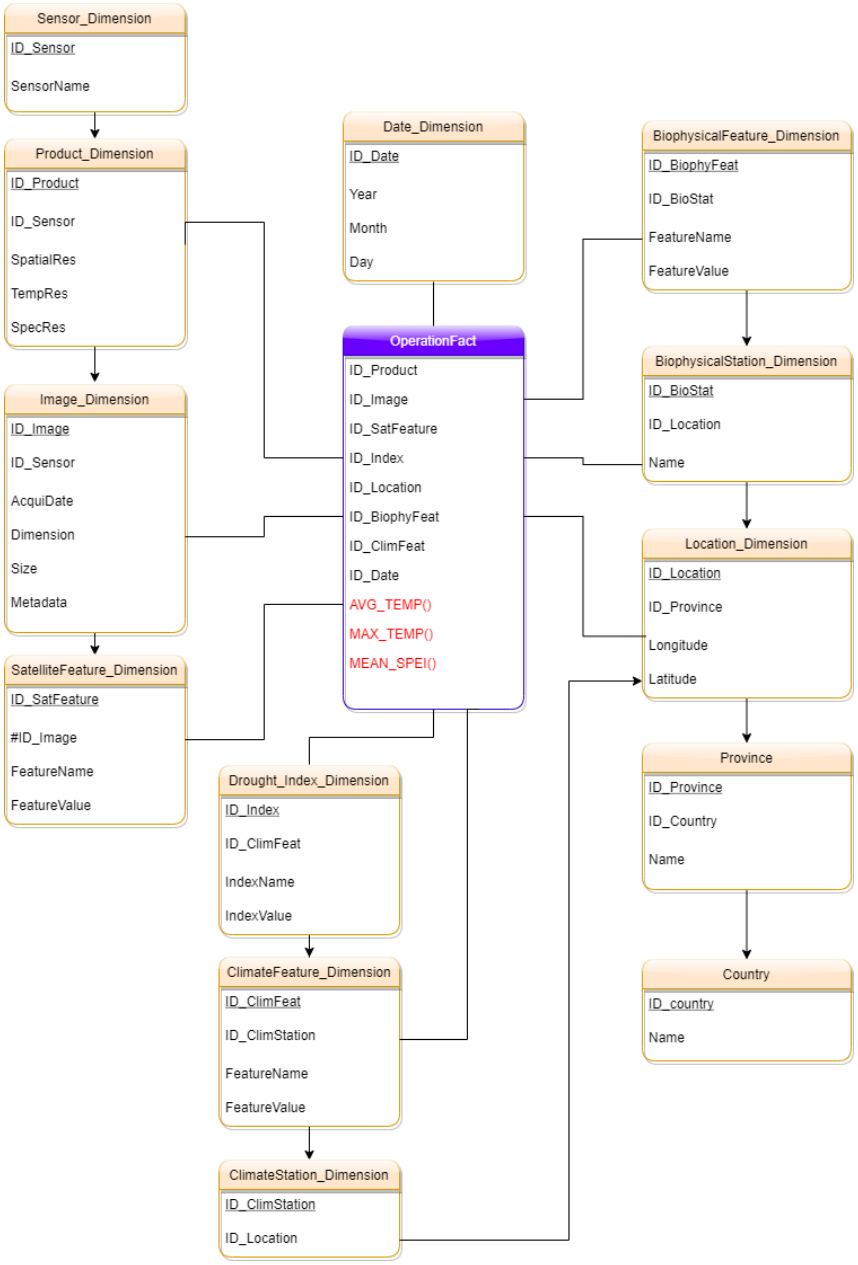


FIG. 1 – *Schéma conceptuel multidimensionnel pour l'entrepôt de données big data de la sécheresse.*

Pour exploiter les données stockées dans le DWH proposé, HiveQL est utilisé. HQL est de type SQL. Il fournit un environnement de type SQL dans Hive pour traiter les tables, les bases

de données et les requêtes.

3.4 Interprétation et visualisation

Les décideurs et les scientifiques ont besoin de comprendre les phénomènes de sécheresse. Ainsi, ils ont besoin d'utiliser les big data pour développer le processus traditionnel de prise de décision. Par conséquent, cette étape vise à présenter les résultats finaux sous forme de représentations qui les aident à comprendre et à déduire les idées potentielles. Pour dialoguer, les données stockées à l'étape précédente, les utilisateurs et les décideurs doivent interroger le DWH afin d'extraire des informations précieuses pour la prise de décision. L'utilisation d'Apache Hive permet d'effectuer diverses requêtes telles que la modélisation des données par la création de dimensions et de faits, les fonctionnalités ETL comme l'extraction, la transformation et le chargement des données, et un outil de requête plus rapide utilisant Hadoop. Plusieurs représentations sont utilisées pour la visualisation, comme des graphiques, des cartes et des rapports textuels. Le but de ces représentations est de montrer et de discuter les tendances, la variabilité spatio-temporelle, et l'intensité de la sécheresse dans une région donnée.

4 Expérimentation et validation

Pour valider notre méthodologie, le domaine d'étude est présenté dans cette section, puis la mise en œuvre de l'architecture de l'entrepôt de données volumineuses est décrite et enfin, certains résultats sont interprétés et discutés.

4.1 Zone d'étude

La Chine est située à l'est du continent asiatique. Les nombreuses formes de terrain et les positions géographiques de la Chine produisent des disparités climatiques majeures entre les différentes régions du monde et génèrent des distributions de sécheresse différentes à travers le pays. La Fig. 2 représente la distribution spatiale des stations climatiques sur la Chine.

4.2 Description des données

Pour valider notre méthodologie, différents types de données ont été utilisés.
Données de télédétection : L'ensemble de données LST a été extrait de MOD11A2, 8 jours avec une résolution spatiale de 1 km et a été collecté à partir de NASA-EOSDIS entre 2000 et 2020 les données NDVI ont été collectées à partir de MYD13A2 16 jours avec Résolution spatiale de 1 km de 2000 à 2020 et les données de précipitations TRMM sont collectées à partir du produit quotidien 3B42 de l'ensemble de données TRMM avec une résolution spatiale de 0,25°x 0,25°.
Données géographiques du sol : Les données d'humidité du sol sont basées sur un ensemble de données GLDAS-NOAH025-M.2.1 mensuel de 0,25 ° x 0,25 ° de 2000 à 2020
Données climatiques : SPEI est basée sur des données climatiques. Le SPEI est utilisé pour déterminer la durée de la sécheresse et permet de comparer la gravité de la sécheresse dans le temps et dans l'espace. Les données SPEI collectées ont quatre résolutions spatiales différentes

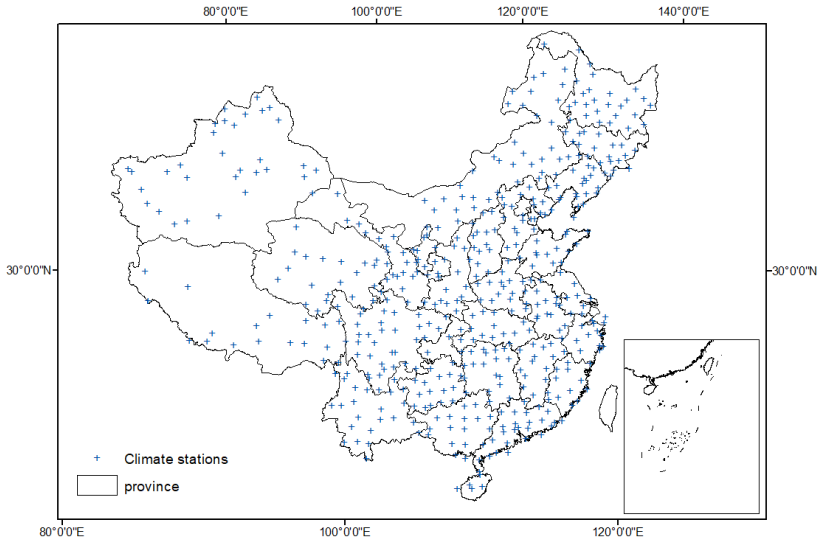


FIG. 2 – *Distribution spatiale des stations climatiques sur la China.*

(par exemple 1, 3, 6, 12 mois), les données ETP sont collectées quotidiennement. Cette caractéristique peut être calculée à l'aide de nombreuses formules, dans notre cas, nous avons utilisé la formule de Penman-Monteith, et enfin, les données climatiques des stations (par exemple, température, humidité, pression, durée d'ensoleillement, vitesse du vent) ont été collectées à partir de 511 stations en Chine.

4.3 Implémentions de l'entrepôt de données

Dans ce travail, Apache Hive a été mis en œuvre pour la construction de DW. Ce choix est rapporté à plusieurs points. Tout d'abord, Hive est construit avec Apache Hadoop qui est la plateforme de cloud computing la plus puissante pour le Big Data. En outre, Hive fonctionne avec un HQL qui est similaire à SQL.
Pour tester l'architecture proposée, quelques requêtes ont été illustrées. L'objectif principal de ces requêtes est de fournir des informations sur la durée et la sévérité de la sécheresse.

Premier exemple de requête : La moyenne de l'indice de sécheresse SPEI-1 en 2000 :
SELECT AVG (ID.IndexValue)
FROM OperationFact OF, Date_Dimension D, Index_Dimension ID
WHERE D.ID_Date=OF.ID_Date and OF.ID_Index= ID.ID_Index
and D.Year="2000" and ID.IndexName="SPEI-1"

Deuxième exemple de requête : La moyenne annuelle des précipitations entre 2000 et 2019. Pour l'humidité et la température, le CFD.FeatureName doit être remplacé par le bon nom de caractéristique.

```
SELECT AVG (CFD.FeatureValue)
FROM OperationFact OF, Date_Dimension D, Climate_Feature_Dimension CFD
WHERE D.ID_Date=OF.ID_Date and OF.ID_Index= CFD.ID_Index
and CFD.FeatureName="Precipitation"
GROUP BY D.Year
```

Troisième exemple de requête : La province ayant la valeur SPEI-3 minimale en 2019

```
SELECT PR.name
FROM Province PR, Location_Dimension LD, Drought_Index_Dimension DID, Operation-
Fact OF, Date D
WHERE PR.ID_Province=LD.ID_Province and LD.ID_Location=OF.ID_Location
and DID.ID_Index=OF.ID_Index and OF.ID_Data=D.ID_Date and D.Year=2019
and DID.IndexValue= (SELECT MIN (IndexValue) FROM Drought_Index_Dimension WHERE
IndexName="SPEI-3")
```

4.4 Résultats et interprétation

Le tableau 1 illustre les différentes requêtes utilisées pour tester notre approche proposée. Dans chaque requête, plusieurs commandes ont été utilisées, telles que WHERE, GROUP BY, HAVING, LEFT (RIGHT) JOIN, ORDER BY et UNION. La combinaison de ces commandes pouvait affecter le temps d'exécution des requêtes. Le tableau 2 représente le nombre de paramètres utilisés dans chaque requête. La figure 3 illustre le temps d'exécution en secondes de chaque requête.

Les résultats ont révélé que lorsque le nombre de commandes utilisé augmente, le temps

Requête	Nombre de commandes	Nombre de tables
Q1	3	3
Q2	3	3
Q3	2	5
Q4	5	4

TAB. 2 – *Nombre de paramètres utilisé pour chaque requête.*

d'exécution augmente. De plus, le nombre de tables utilisées dans une requête affecte la vitesse d'exécution. En fait, dans Q3 et Q4,respectivement 5 et 4 tables sont utilisées avec respectivement 2 et 5 commandes. Les résultats montrent que le temps d'exécution de Q3 était plus élevé que celui de Q4.

Pour la visualisation, plusieurs formes sont utilisées. Des cartes sont utilisées pour montrer la variation spatio-temporelle de l'intensité de la sécheresse en Chine. La Figure 4, la Figure 5 illustre la cartographie de la sécheresse en Chine en 2000 et 2019.

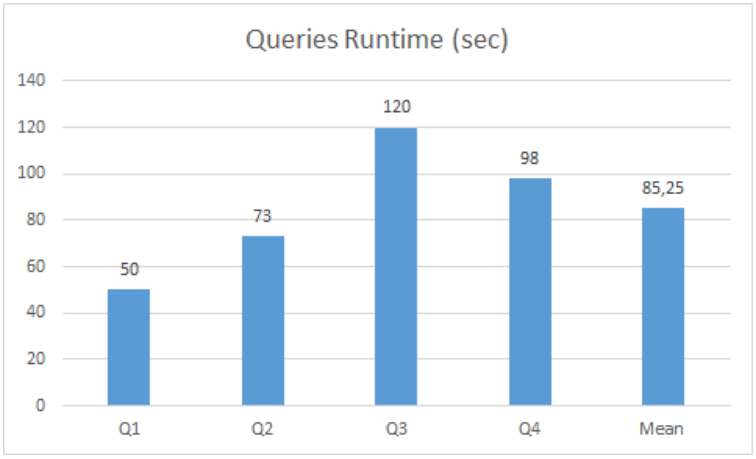


FIG. 3 – *comparaison entre le temps d'exécution des requêtes.*

Les résultats de la première application en 2000 et 2019 montrent que la moyenne de SPEI-1

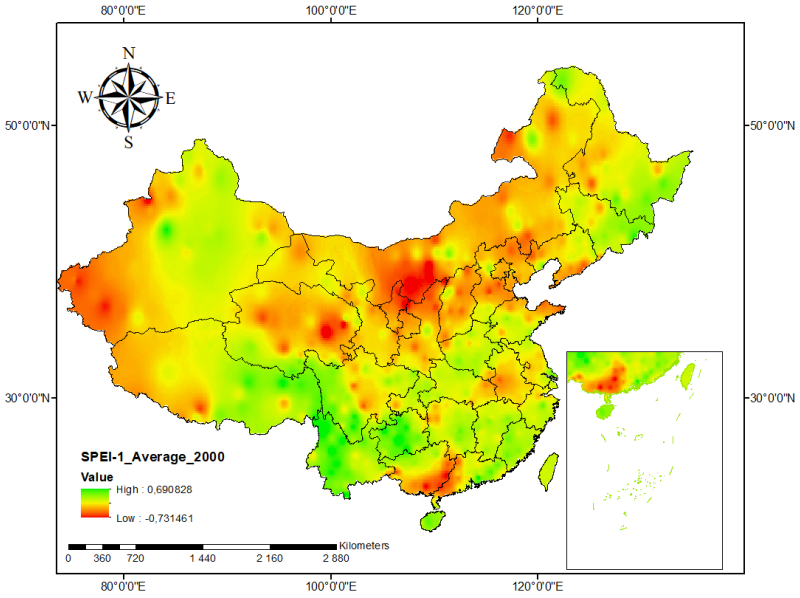


FIG. 4 – *Visualisation de la moyenne de SPEI-1 en 2000.*

varie entre 0,69 et -0,7 en 2000 et 1,5 et -1,5 en 2019. Les figures 4, 5 montrent que certaines provinces sont passées de la catégorie WET (régions en bleu foncé) à la catégorie sèche (régions en rouge) sur la base des valeurs SPEI. Par exemple, le sud-ouest de la Chine est dans un

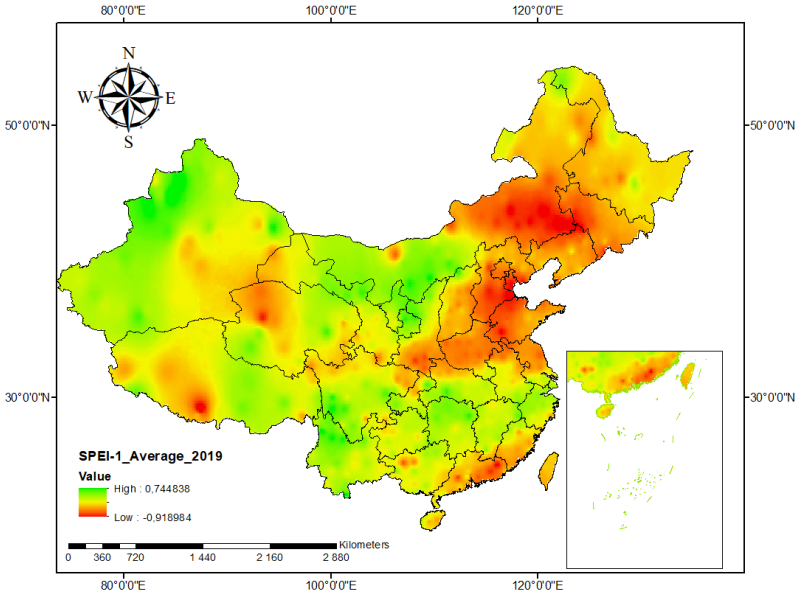


FIG. 5 – *Visualisation de la moyenne de SPEI-1 en 2019.*

état plus humide en 2000 mais en 2019 et le nord-est de la Chine est dans un état plus sec en 2000 et plus humide en 2019.

5 Conclusuion

Dans cet article, nous avons proposé un schéma flocon de neige comme architecture globale pour intégrer divers ensembles de données d'observation de la terre pour la surveillance de la sécheresse. Ce schéma est élastique et compatible avec la modélisation Big Data pour la prévention d'autres risques naturels. Sur la base de ce schéma, nous avons extrait, migré et chargé les informations de différents ensembles de données dans une représentation unique de l'ensemble de données sur la sécheresse. Malheureusement, Hive n'a pas la capacité de traiter des données en temps réel, par conséquent, une combinaison avec un outil en temps réel sera proposée dans un travail futur. En outre, nous souhaitons développer un nouvel outil basé sur un service web pour fournir des informations en temps réel à l'échelle spatio-temporelle.

Références

Balti, H., A. B. Abbes, N. Mellouli, I. R. Farah, Y. Sang, et M. Lamolle (2020). A review of drought monitoring with big data : Issues, methods, challenges and research directions. *Ecological Informatics 60*, 101136.

Jenhani, F., M. S. Gouider, et L. B. Said (2019). Streaming social media data analysis for events extraction and warehousing using hadoop and storm : Drug abuse case study. *Procedia Computer Science 159*, 1459–1467.

Liu, S., C. Han, S. Wang, et Q. Luo (2012). Data warehouse design for earth observation satellites. *Procedia Engineering 29*, 3876–3882.

Merritt, P., H. Bi, B. Davis, C. Windmill, et Y. Xue (2018). Big earth data : a comprehensive analysis of visualization analytics issues. *Big Earth Data 2*(4), 321–350.

Mishra, A. K. et V. P. Singh (2011). Drought modeling–a review. *Journal of Hydrology 403*(1-2), 157–175.

Ngo, V. M. et M.-T. Kechadi (2020). Crop knowledge discovery based on agricultural big data integration. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, pp. 46–50.

Ngo, V. M., N.-A. Le-Khac, et M.-T. Kechadi (2020). Data warehouse and decision support on integrated crop big data. *arXiv preprint arXiv :2003.04470*.

Sebaa, A., F. Chikh, A. Nouicer, et A. Tari (2018). Medical big data warehouse : Architecture and system design, a case study : Improving healthcare resources distribution. *Journal of medical systems 42*(4), 1–16.

Summary

Here, we propose a big data warehousing and advanced analytics for drought monitoring. In order to support the decision-making process, a big data warehousing framework was designed and implemented. The proposed framework consists of 4 main steps: data collection, data preprocessing, data storage and decision-making step. The used data are multi-source and heterogeneous. Hence it leads to offers a view of the different dimensions such as the spatial distribution, the temporal distribution, and the severity detection of this phenomenon. We also evaluate the efficiencies of our proposed framework in a real case study of drought monitoring in China from 2000 to 2020. In addition, we present return knowledge of some queries and different mappings.

Keywords: Big data analytics, data warehouse, storage, disaster management, drought, Hive

Alignement non supervisé d'embeddings de mots dans le domaine biomédical

Félix Gaschi*,** Parisa Rastin** Yannick Toussaint**

*SAS Posos, 53 rue de la Boétie, 75008 Paris
prenom@posos.fr,
www.posos.co
**LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy
nom.prenom@loria.fr
www.loria.fr

Résumé. Notre objectif est de créer un alignement non supervisé et multilingue d'embeddings de mots (ou plongements lexicaux) basés sur des corpora de domaines différents. Plus précisément, nous cherchons à aligner un embedding cible anglais du domaine biomédical avec un embedding source du domaine général d'une autre langue, puisque les textes à traiter sont dans diverses langues (français, espagnol...) et que le vocabulaire du domaine biomédical est essentiellement disponible en anglais. Notre méthode pour aligner deux embeddings de domaines et langages différents repose sur un autre embedding pivot de même domaine que la source et de même langage que la cible. Notre méthode aligne d'abord les embeddings de même domaine pour créer un dictionaire qui sert ensuite à aligner les embeddings de domaines et languages distincts. Elle est évaluée sur une tâche de traduction du domaine biomédical dans plusieurs langues. Bien que notre algorithme ne dépasse pas les méthodes d'alignement entre embeddings de même domaine, elle dépasse ces mêmes méthodes appliquées à des embeddings de domaines différents. Ce travail préliminaire montre qu'aligner des embeddings de domaines différents est possible de manière non supervisé.

Mots-clés : embeddings de mots, traitement automatique du langage, multilingue, apprentissage non supervisé

1 Introduction

Les embeddings de mots (ou plongements lexicaux) fournissent des représentations utiles pour de nombreuses tâches de machine learning (Mikolov et al., 2013b; Pennington et al., 2014). Les embeddings ont été généralisés dans un context multilingue en un concept d'embedding multilingue non supervisé (EMN) (Mikolov et al., 2013a; Artetxe et al., 2017; Zhang et al., 2017a; Lample et al., 2018b; Artetxe et al., 2018). Les données spécifiques à un domaine précis, comme les publications scientifiques, peuvent être rares pour des langues autres que l'anglais. En particulier dans le domaine biomédical, les quantités de ressources disponibles en anglais (plus de 21 millions d'articles sur PubMed) dépassent largement celles disponibles

dans d'autres langues. C'est pourquoi nous avons pour but de créer des représentations multilingues qui alignent un embedding cible spécifique au domaine biomédical en anglais avec un embedding source issu du domaine général dans une autre langue. Notre objectif n'est pas tant de surpasser l'état de l'art en termes d'alignement non supervisé d'embedding, mais plutôt d'explorer la possibilité d'aligner de manières non supervisée des embeddings de domaines distincts.

Les EMN visent à apprendre des représentations multilingues uniquement à partir de données monolingues. Les méthodes existantes pour construire de tels embeddings multilingues reposent souvent sur une transformation linéaire orthogonale et montrent de bons résultats tant que les données sont issues du même domaine (Mikolov et al., 2013a; Artetxe et al., 2017; Zhang et al., 2017a; Lample et al., 2018b; Artetxe et al., 2018). Cependant, elles ont tendance à produire de moins bons résultats avec des données issues de domaines différents (Søgaard et al., 2018). Ces méthodes s'appuient sur l'hypothèse que les espaces métriques des embeddings de mots sont approximativement isométriques (Mikolov et al., 2013a; Zhang et al., 2017a; Lample et al., 2018a).

Bien que Søgaard et al. (2018) aient montré que cette similarité isométrique approximative n'est pas conservée pour des embeddings entraînés sur des domaines différents, l'hypothèse initiale peut tout de même rester valable pour des sous-ensembles bien choisis. Nous apportons en effet une amélioration sur les alignements d'embeddings de domaines différents en s'appuyant sur un embedding pivot de la même langue que l'embedding cible et du même domaine que l'embedding source. Nous mettons en pratique notre méthode pour aligner un embedding source entraîné sur Wikipedia en français sur un embedding cible entraîné sur PubMed en anglais. Notre pivot est donc un embedding entraîné sur le domaine général en anglais dans nos expériences.

2 Travaux connexes

Embedding cross-lingues et non supervisés. Peu après avoir introduit les modèles Skip-gram et de Continuous Bag-of-Words (CBOW) (Mikolov et al., 2013c) pour apprendre des embeddings de mots, Mikolov et al. (2013a) proposent d'aligner des embeddings de différentes langues dans un espace partagé avec l'aide d'un dictionaire bilingue. Avec l'apparition de l'auto-apprentissage itératif (Artetxe et al., 2017), qui alterne entre l'apprentissage d'un alignement et celui d'un dictionaire, les méthodes d'alignement d'embeddings progressent et requièrent alors de moins en moins de paires de mot dans le dictionnaire initial d'entraînement. Finalement, des méthodes entièrement non supervisées reposent sur l'apprentissage adversaire (Zhang et al., 2017a; Lample et al., 2018b) et des heuristiques d'initialisation (Artetxe et al., 2018). Ces dernières s'appuient en grande partie sur une transformation linéaire orthogonale appliquée à des embeddings normalisés (Xing et al., 2015; Smith et al., 2017) assurant ainsi une invariance des distances entre les mots au sein d'une même langue.

Limites de la contrainte orthogonale. Søgaard et al. (2018) démontrent que les EMNs qui s'appuient sur une transformation orthogonale ont besoin de trois conditions pour être efficaces : (1) les langues à aligner doivent être morphologiquement similaires, (2) les corpora d'entraînement monolingues doivent être issus du même domaine, et (3) le même modèle doit être utilisé (un embedding CBOW en espagnol ne pourra pas être aligné avec un embedding

Skip-gram en anglais). En effet, utiliser une transformation orthogonale implique que les embeddings soient à-peu-près isométriques (Mikolov et al., 2013a; Zhang et al., 2017a; Lample et al., 2018b). Søgaard et al. (2018) montrent, grâce à une comparaison de valeurs propres, que les graphes de voisinage des mots ne sont pas isomorphiques. Zhang et al. (2017b) montrent que la distance de Wasserstein entre des embeddings est corrélée à la similarité typologique entre les langues. Patra et al. (2019) utilisent une autre métrique basée sur des homologies persistentes pour évaluer la similarité entre des embeddings de mots de langues différentes. Ces observations mènent finalement à la création de plusieurs méthodes utilisant une contrainte d'orthogonalité faible (Zhang et al., 2017a; Patra et al., 2019). Pour tenir compte de variations locales de la densité des embeddings, un critère local de mise à l'échelle (Cross-domain Similarity Local Scaling ou CSLS) (Lample et al., 2018a) est souvent utilisé dans les modèles basés sur des transformations orthogonales (Lample et al., 2018a; Artetxe et al., 2018; Joulin et al., 2018).

Dépasser les limites des embeddings cross-lingues. Søgaard et al. (2018) montrent qu'un des modèles basés sur des tranformations orthogonales (Lample et al., 2018b) obtient une précision proche de zéro lorsqu'on aligne des embeddings de domaines différents et pour des langues éloignées. De plus, ils montrent qu'il est possible d'augmenter la précision en utilisant les mots écrits de manières identiques dans deux langues comme signal de supervision faible. Cependant, pour autant que nous le sachions, avec l'exception de cette supervision faible, il n'y a pas eu de travaux proposant une méthode pour aligner des embeddings de domaines différents de manière non supervisée. Des méthodes semi-supervisées avec une contrainte orthogonale faible ont été proposées (Patra et al., 2019). Shakurova et al. (2019) ont appliqué avec succès des méthodes d'alignement d'embeddings sur des domaines spécifiques, mais toujours pas entre domaines distincts. Si l'alignement entre des embeddings de domaines différents a, semble-t-il, suscité peu d'intérêt, le cas des langues distantes ou faibles en ressources textuelles a été plus largement étudié. À titre d'exemple, Nakashole (2018) développe un modèle basé sur les voisinages pour améliorer l'alignement entre des embeddings de langues distantes et Nakashole et Flauger (2017) utilisent un langage riche en donnée monolingues disponibles comme pivot pour aligner des langues pour lesquelles moins de ressources sont disponibles, ce qui a en partie inspiré notre idée d'utiliser un troisième embedding intermédiaire pour l'alignement entre domaines différents.

La plupart des approches pour aligner des langues distantes s'appuient sur une contrainte orthogonale faible voire s'en débarrasse complètement, ce qui est justifié par l'apparente absence d'isométrie entre les embeddings de langues distantes. Nous pensons que cette absence profonde d'isométrie ne s'applique pas dans les cas où les embeddings proviennent simplement de domaines différents. Par la suite, nous détaillons brièvement le lien entre transformation orthogonale et isométrie et nous faisons l'hypothèse que, bien que cette condition d'isométrie ne soit pas valable sur l'intégralité des embeddings de domaines différents elle pourrait l'être pour des sous-ensembles bien choisis de ces embeddings.

3 Considérations sur l'isométrie entre des sous-ensembles d'embeddings

Les méthodes supervisées pour l'apprentissage d'embeddings multilingues construisent généralement une transformation linéaire entre les représentations des entrées d'un dictionaire bilingue (Mikolov et al., 2013a). En suivant le formalisme de Lample et al. (2018a), nous avons :

$$W^* = \arg \min_{W \in \mathcal{O}_d} ||AW - B|| \quad (1)$$

Avec $A \in \mathbb{R}^{N \times d}$ et $B \in \mathbb{R}^{N \times d}$ les représentations de dimension d des entrées du dictionnaire bilingue dans les embeddings source et cible. W^* est la transformation linéaire apprise, dans $\mathcal{O}_d$, l'ensemble des matrices orthogonales. La transformation est choisie orthogonale pour conserver les distances au sein des embeddings monolingues. Comme une transformation orthogonale préserve les distances, cela signifie également que la fonction $f : \mathbb{R}^d \to \mathbb{R}^d$ telle que $f(A_i) = B_i$ doit être une isométrie pour que la méthode fonctionne, c'est-à-dire que pour toute paire de lignes issus de A, $A_i, A_j \in \mathbb{R}^d$, on doit avoir $l(A_i, A_j) \approx l(B_i, B_j)$ avec B_i, B_j les lignes de B correspondantes et l une distance.

Pour les méthodes entièrement non supervisées, le dictionnaire doit être appris en même temps que la transformation orthogonale. Pour deux ensembles de vecteurs donnés, représentés par les matrices X pour la source et Z pour la cible, on doit trouver l'application $f : X \to Z$ qui est un dictionnaire bilingue et doit être une quasi-isométrie. Cela signifie que X et Z définissent des espaces métriques qui sont eux-même quasi-isométriques.

Alors qu'une telle fonction peut exister dans le cas mono-domaine, ça n'est pas le cas dans un cas multi-domaine (Søgaard et al., 2018) ; les mots spécifiques à un domaine dans un embedding peuvent ne pas trouver de traduction dans l'embedding d'un autre domaine.

Nous faisons l'hypothèse qu'il est possible d'améliorer les méthodes non supervisées pour la création d'embeddings multilingues et multi-domaines en essayant d'aligner des sous-ensembles bien choisis de chaque vocabulaire. Comme schématisé en figure 1, cet "alignement partiel" des embeddings pourrait être utile pour certaines tâches spécifiques à un domaine, comme le domaine biomédical.

Pour valider cette hypothèse, nous réalisons deux expériences. En suivant les travaux de Patra et al. (2019), nous utilisons la distance de bottleneck pour mesurer à quel point des sous-ensembles d'embeddings de différents domaines sont proche de l'isométrie. Les détails du calcul de la distance de bottleneck sont détaillés dans la section suivante. Ensuite, nous proposons une méthode simple d'alignement non supervisée basé sur une transformation orthogonale qui s'applique à des embeddings de domaines différents. Nous la détaillons dans la section 5.

4 Une métrique pour la quasi-isométrie

Le calcul de la distance de bottleneck nous permet d'évaluer à quel point deux embeddings s'éloignent de l'isométrie. L'utilisation de la distance de bottleneck pour un tel usage est proposée par Patra et al. (2019).

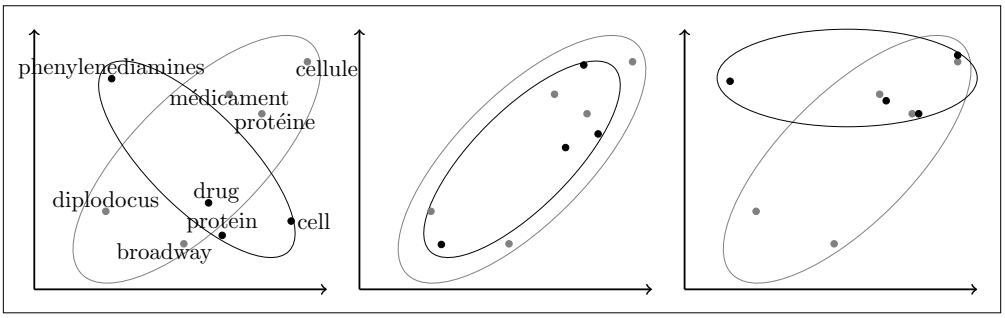


FIG. 1: Exemple jouet. Quand on aligne un domaine (gris) avec un autre (noir), les embeddings initialement non-alignés (gauche) s'alignent mal lorsqu'on essaye d'aligner tous les mots (centre). Notre objectif est de les aligner partiellement (droite).

Les embeddings sont normalisés comme le font Artetxe et al. (2018) : l1-normalisation, centrés à la moyenne, puis l1-normalisation à nouveau. La l1-normalisation permet notamment d'avoir une équivalence entre la distance cosinus et la distance l2.

Une mesure de la distance entre deux espaces métriques $(\mathcal{X}, d)$ et $(\mathcal{Y}, d)$ peut-être donnée par la distance de Hausdorff qui est la distance maximale entre les paires de plus proches voisins :

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = \max\left(\sup_{x\in\mathcal{X}} \inf_{y\in\mathcal{Y}} d(x,y), \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} d(x,y)\right) \tag{2}$$

Cette distance de Hausdorff mesure à quel point deux espaces métriques "coïncident". Pour mesurer à quel point ils sont isométriques, il faut connaître la distance de Hausdorff minimale entre toutes les transformations isométriques possibles de $\mathcal{X}$ et $\mathcal{Y}$. C'est la distance de Gromov-Hausdorff :

$$\mathcal{GH}(\mathcal{XY}) = \min_{f,g} \mathcal{H}(f(\mathcal{X}), g(\mathcal{Y})) \tag{3}$$

Avec f et g des isométries. Calculer cette distance n'est pas réalisable dans notre cas, mais on peut l'approximer en utilisant la distance bottleneck (Chazal et al., 2009). Pour la calculer, on construit d'abord les diagrammes de persistance de premier ordre sur les complexes de Vietoris-Rips des deux espaces métriques. Concrètement, pour une distance t qui varie de 0 à $+\infty$, on on construit des simplexes pour chaque ensemble de points qui se trouvent à une distance inférieure à t. En faisant croître t, on commence avec autant de composante connexes que de points dans l'espace métrique, et on diminue leur nombre en les fusionnant petit à petit. Le diagramme de persistance est constitué des points $(t_{\text{birth}}, t_{\text{death}})$ pour chaque composante connexe qui est apparue à $t = t_{\text{birth}}$ et a été fusionnée avec une autre à $t = t_{\text{death}}$.

Soit nos deux diagrammes de persistance f et g, la distance de bottleneck est le minimum pour toute bijection γ entre f et g de la distance de Manhattan maximale entre les éléments de f et leur image par γ :

$$\mathcal{B}(f,g) = \inf_{\gamma}(\sup_{u\in f} ||u - \gamma(u)||_\infty) \tag{4}$$

5 Méthode d'alignement avec pivot

Søgaard et al. (2018) utilisent les mots identiques comme signal de supervision faible. De manière similaire, nous voulons construire un dictionnaire de manière non supervisée qui pourrait être utilisé pour apprendre une transformation satisfaisant l'equation 1.

Les embeddings source et cible sont de domaines et langages différents. Dans nos expériences, il s'agira par exemple d'un embedding du domaine général entraîné sur Wikipedia en français et d'un embedding du domaine biomédical entraîné sur PubMed en anglais. En plus de ces deux embeddings, on dispose d'un embedding supplémentaire intermédiaire, de même domaine que la source et de même langage que la cible (Wikipedia en anglais dans notre exemple).

La méthode que nous proposons prend comme entrée ces trois embeddings entraînés avec FastText (Bojanowski et al., 2017) : X l'embedding source, Z l'embedding cible et Y le pivot, de même domaine que X et de même langage que Z. Tous les embeddings sont normalisés de la même manière que Artetxe et al. (2018) : normalisés par la norme, centrés à la moyenne, puis normalisé par la norme à nouveau.

D'abord, nous alignons X et Y (la source et le pivot) de manière non supervisée à l'aide de l'algorithme VecMap (Artetxe et al., 2018). X et Y sont des embeddings de même domaines sur lesquels des méthodes comme VecMap ont fait leurs preuves. VecMap apprends d'abord un dictionnaire initial en représentant chaque mot par sa similarité à ses plus proches voisins et construit les paires d'un dictionnaire bilingue par simple recherche de plus proche voisin dans cette nouvelle représentation. Ensuite, une transformation orthogonale et un dictionnaire sont affinés alternativement dans une boucle d'auto-apprentissage.

À partir des embeddings source et pivot de même domaines maintenant alignés grâce à VecMep, $\tilde{X}$ et $\tilde{Y}$, on peut maintenant inférer un dictionnaire. Mais nous pouvons restreindre ce dictionnaire à l'intersection du vocabulaire entre le pivot $\tilde{Y}$ et la cible Z puisque ce sont deux embeddings de même langage. Plus précisément, pour chaque mot de la source $\tilde{X}$ nous constituons une paire avec son plus proche voisin dans l'embedding pivot aligné $\tilde{Y}$ si le mot correspondant se trouve aussi dans le vocabulaire de la cible. Et pour chaque mot qui est à la fois dans Z et $\tilde{Y}$ nous constituons une paire avec le mot de $\tilde{X}$ le plus proche. On peut ainsi construire les matrices A et B de notre dictionnaire bilingue et l'injecter dans l'équation 1 pour aligner la source et la cible. La solution est donnée par la décomposition en valeur singulière (SVD) de $A^\top B = USV^\top$ en écrivant $W^* = UV^\top$.

6 Expériences et résultats

Dans un premier temps, nous voulons mesurer à quel point différents sous-ensembles de paires d'embeddings sont proche de l'isométrie. Dans un second temps, nous évaluons la méthode proposée sur une tâche de traduction spécifique au domaine biomédical.

Dans ce qui suit, l'embedding source est systématiquement un embedding entraîné sur Wikipédia dans une langue autre que l'anglais (français, allemand, espagnol ou portugais). L'embedding cible est entraîné quant à lui sur PubMed [1] pour l'alignement *multi-domaine*,

1. un ensemble d'environ 21 million d'articles scientifiques du domaine biomédical, écrits en anglais, fournis par la U.S. National Library of Medicine `https://www.nlm.nih.gov/databases/download/pubmed_medline.html`

	source-cible	source-pivot
20 000 mots		
fréquents	0.1532	0.0626
MeSH	0.0638	0.0806
Dictionnaire en sortie de VecMap		
dico.	0.1413	0.0763

TAB. 1: Distance de bottleneck entre différents sous-ensembles d'embeddings en français et en anglais.

et sur Wikipedia en anglais pour l'alignement *mono-domaine*. L'embedding intermédiaire de notre méthode est entraîné sur Wikipedia en anglais.

Mesure de la distance bottleneck. Nous avons fait l'hypothèse que certains sous-ensembles d'embeddings de domaines différents peuvent être malgré tout à-peu-près isométriques. Pour s'assurer de la validité de cette hypothèse, nous utilisons la distance de bottleneck comme définie plus haut (section 4). Plus la distance de bottleneck est proche de zéro, plus la paire d'espaces métriques évaluée est proche d'être isométrique.

Nous reportons nos résultats dans le tableau 1. Les premiers 20 000 mots les plus fréquents (*fréquents* dans le tableau) de deux embeddings de même domaine, la source et la cible (*src-cib*), sont plus proches en terme de distance de bottleneck que les 20 000 mots les plus fréquents de deux embeddings de même domaine, la source et l'intermédiaire (*src-pivot*). Cela expliquerait pourquoi des méthodes comme VecMap basées sur une transformation orthogonale fonctionnent mieux dans le cas mono-domaine que dans le cas multi-domaine, puisqu'il s'agit de construire un dictionnaire sur ces 20 000 mots les plus fréquents.

À l'inverse, quand on évalue cette même distance entre les 20 000 mots les plus fréquents du MeSH [2], une ontologie biomédicale en anglais avec sa traduction en français [3], on obtient le résultat inverse. Les ensembles des représentation vectorielles des mots les plus fréquents dans le vocabulaire biomédical semblent plus proches en terme de distance de bottleneck quand on utilise l'embedding spécifique au domaine en question.

Il semblerait donc que l'hypothèse d'isométrie puisse toujours être valable en choisissant soigneusement les sous-ensembles du vocabulaire que l'on cherche à aligner. Avec notre méthode, qui s'appuie sur un alignement mono-domaine et l'intersection des vocabulaires pour générer un dictionnaire spécifique au domaine recherché, nous fournissons une première heuristique simple pour montrer qu'une sélection du vocabulaire peut améliorer les résultats sur des alignements multi-domaines. Cependant, nous montrons également dans le tableau 1 que la distance de bottleneck évaluée sur le vocabulaire du dictionaire bilingue inféré grâce à VecMap (*dico.*) est plus grande dans le cas multi-domaine. Le sous-ensemble de vocabulaire que nous avons sélectionné n'est donc pas optimal et pourrait donc être amélioré.

2. fournie par le NLM `https://www.nlm.nih.gov/mesh/meshhome.html`
3. fournie par l'INSERM `http://mesh.inserm.fr/FrenchMesh/`

	fr-en	es-en	de-en	pt-en
Wikipedia xx avec PubMed en non supervisé				
VecMap	0.093	0.062	0.068	0.065
MUSE	0.053	0.064	0.055	0.069
WP	0.081	0.081	0.052	0.053
notre méthode	0.382	0.503	0.313	0.460
Wikipedia xx avec PubMed en faiblement supervisé				
VecMap	0.299	0.365	0.254	0.289
Wikipedia xx avec Wikipedia en non supervisé				
VecMap	**0.455**	**0.582**	0.373	**0.555**
MUSE	0.434	0.579	**0.398**	0.532
WP	0.447	0.571	0.363	0.513
soumission UCAM run 3				
UCAM	-	*0.708*	*0.612*	-

TAB. 2: score BLEU-1 sur la tâche de traduction Biomedical WM19.

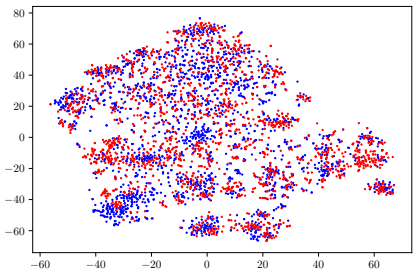
Évaluation de notre méthode. Nous évaluons notre méthode sur l'ensemble de test du dataset Biomedical WMT19[4]. Cette tâche fournie des résumés d'articles PubMed dans divers langages et leur traduction en anglais. Comme notre méthode est une méthode d'alignement de vecteurs de mots et non un modèle de traduction, nous l'utilisons pour traduire les phrases mot-à-mot en la comparant avec d'autres méthodes d'alignement d'embedding, mono-domaines et multi-domaines. On évalue donc les performances à l'aide du score BLEU-1 (Papineni et al., 2002) sur les résumés d'articles entiers.

Nous avons choisi une tâche de traduction plutôt que d'induction de dictionnaire car celle-ci ne tient notamment pas compte des variations morphologiques des mots (Czarnowska et al., 2019) et donne trop d'importance à certain mot comme les noms propres (Kementchedjhieva et al., 2019). De plus, nous pouvons envisager d'utiliser des méthodes d'alignement multi-domaine comme heuristique d'initialisation pour des modèles de traduction non supervisée spécifiques au domaine biomédical, à la manière des travaux de Artetxe et al. (2017); Lample et al. (2018a); Artetxe et al. (2019).

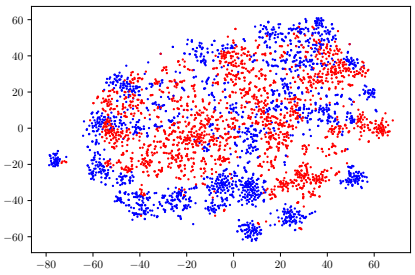
Les résultats sur la tâche de traduction de quatre langues différentes (français, espagnol, allemand, portugais) vers l'anglais sont montrés dans le Tableau 2. Notre méthode est comparée avec les méthodes non supervisées VecMap (Artetxe et al., 2018), MUSE (Lample et al., 2018b) et Wasserstein-Procrustes (WP) (Grave et al., 2018) appliquées à des paires d'embeddings de domaines différents aussi bien qu'à des paires d'embeddings de même domaines. Les résultats de la soumission UCAM (Saunders et al., 2019) au challenge Biomedical WM19 sont aussi donnés à titre de borne supérieure de référence, puisqu'il s'agit d'un modèle de deep learning a priori inégalable avec une simple méthode d'alignement d'embeddings. On présente également les résultats de la méthode faiblement supervisée proposée par Søgaard et al. (2018) basés sur les mots identiques au sein d'une paire de langues.

Dans toutes les langues, notre méthode dépasse les autres méthodes multi-domaines, y compris la méthode faiblement supervisée. En revanche les méthodes mono-domaines appli-

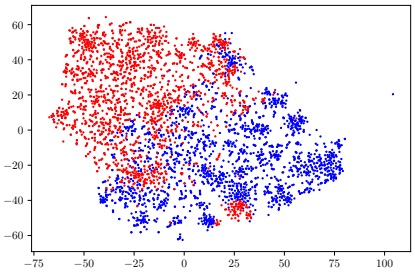
4. http://www.statmt.org/wmt19/biomedical-translation-task.html



(a) VecMap, même domaine.



(b) VecMap, domaines différents.



(c) Notre méthode, domaines différents.

FIG. 2: t-SNE sur des embeddings alignés français (bleu) et anglais (rouge).

quées au domaine général semblent généraliser assez bien au domaine biomédical et obtiennent de meilleures résultats que notre méthode.

Visualisation des alignements. Pour conclure nos expériences, nous avons utilisé une méthode de réduction de dimension, la t-SNE (van der Maaten et Hinton, 2008), pour vérifier qu'un "alignement partiel" comme prévu par notre hypothèse avait bien lieu (cf. figure 1). Nous reportons en figure 2 les résultats d'une telle t-SNE sur les embeddings français (en bleu) et anglais (en rouge) alignés selon trois techniques. L'alignement mono-domaine à l'aide de VecMap des embeddings basés sur Wikipedia en français et en anglais (2a) montre que les nuages de points des deux langues sont globalement superposés et on observe la superposition de certaines grappes. Cette superposition locale est moins, voire pas du tout visible, lorsqu'on utilise la même méthode, VecMap, pour aligner des embeddings de domaines différents (2b). Les nuages de points des deux langues sont toujours globalement superposés, mais peu de grappes semblent se superposer localement. Ce résultat est corrélé au fait que VecMap sur des embeddings de domaines distincts obtient des scores proches de zéro dans notre tâche de traduction. Enfin, nous observons que pour la notre méthode appliquée aux embeddings de domaines différents (2c), les nuages de points ne sont plus globalement alignés. On retrouve ce qui ressemble à l'idée "d'alignement partiel" schématisée dans notre exemple jouet (figure

1). Toutefois, sur la partie où les deux langues se superposent, nous n'observons pas d'alignements significatifs de grappes de points. Cette visualisation tend donc à confirmer l'idée qu'un alignement partiel est possible, mais elle souligne également que notre méthode est perfectible.

7 Conclusion

Nous avons montré que l'hypothèse de quasi-isométrie est encore valable pour des sous-ensembles bien choisis d'embeddings de domaines différents. Nous avons fait la démonstration que les alignements non supervisés basés sur des transformations orthogonales ne sont pas voués à l'échec dans le cas multi-domaine. Cependant, notre méthode reste assez naïve et la distance de bottleneck élevée entre les représentations de notre dictionnaire intermédiaire montrent qu'il y a encore des possibilités d'amélioration.

L'amélioration qu'apportent des méthodes comme la nôtre ou celle faiblement supervisée proposée par Søgaard et al. (2018) par rapport à d'autres sur des embeddings de domaines différents suggèrent que les performances de méthodes d'alignement d'embeddings sont très sensibles à l'initialisation. De futures recherches sont à mener sur les méthodes d'initialisation et leur adaptation à des embeddings de domaines différents.

Références

Artetxe, M., G. Labaka, et E. Agirre (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Vancouver, Canada, pp. 451–462. Association for Computational Linguistics.

Artetxe, M., G. Labaka, et E. Agirre (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia, pp. 789–798. Association for Computational Linguistics.

Artetxe, M., G. Labaka, et E. Agirre (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 194–203. Association for Computational Linguistics.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics 5*, 135–146.

Chazal, F., D. Cohen-Steiner, L. J. Guibas, F. Mémoli, et S. Y. Oudot (2009). Gromov-hausdorff stable signatures for shapes using persistence. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, Goslar, DEU, pp. 1393–1403. Eurographics Association.

Czarnowska, P., S. Ruder, E. Grave, R. Cotterell, et A. Copestake (2019). Don't forget the long tail ! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 974–983. Association for Computational Linguistics.

Grave, E., A. Joulin, et Q. Berthet (2018). Unsupervised alignment of embeddings with wasserstein procrustes.

Joulin, A., P. Bojanowski, T. Mikolov, H. Jégou, et E. Grave (2018). Loss in translation : Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 2979–2984. Association for Computational Linguistics.

Kementchedjhieva, Y., M. Hartmann, et A. Søgaard (2019). Lost in evaluation : Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 3336–3341. Association for Computational Linguistics.

Lample, G., A. Conneau, L. Denoyer, et M. Ranzato (2018a). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lample, G., A. Conneau, M. Ranzato, L. Denoyer, et H. Jégou (2018b). Word translation without parallel data. In *International Conference on Learning Representations*.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Q. V. Le, et I. Sutskever (2013b). Exploiting similarities among languages for machine translation. *CoRR abs/1309.4168*.

Mikolov, T., Q. V. Le, et I. Sutskever (2013c). Exploiting similarities among languages for machine translation.

Nakashole, N. (2018). NORMA : Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 512–522. Association for Computational Linguistics.

Nakashole, N. et R. Flauger (2017). Knowledge distillation for bilingual dictionary induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2497–2506. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, et W.-J. Zhu (2002). Bleu : A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, USA, pp. 311–318. Association for Computational Linguistics.

Patra, B., J. R. A. Moniz, S. Garg, M. R. Gormley, et G. Neubig (2019). Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 184–193. Association for Computational Linguistics.

Pennington, J., R. Socher, et C. D. Manning (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Saunders, D., F. Stahlberg, et B. Byrne (2019). UCAM biomedical translation at WMT19 : Transfer learning multi-domain ensembles. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3 : Shared Task Papers, Day 2)*, Florence, Italy, pp. 169–174. Association for Computational Linguistics.

Shakurova, L., B. Nyari, C. Li, et M. Rotaru (2019). Best practices for learning domain-specific cross-lingual embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Florence, Italy, pp. 230–234. Association for Computational Linguistics.

Smith, S. L., D. H. P. Turban, S. Hamblin, et N. Y. Hammerla (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

Søgaard, A., S. Ruder, et I. Vulić (2018). On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia, pp. 778–788. Association for Computational Linguistics.

van der Maaten, L. et G. Hinton (2008). Visualizing data using t-sne. *Journal of Machine Learning Research 9*(86), 2579–2605.

Xing, C., D. Wang, C. Liu, et Y. Lin (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Denver, Colorado, pp. 1006–1011. Association for Computational Linguistics.

Zhang, M., Y. Liu, H. Luan, et M. Sun (2017a). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Vancouver, Canada, pp. 1959–1970. Association for Computational Linguistics.

Zhang, M., Y. Liu, H. Luan, et M. Sun (2017b). Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 1934–1945. Association for Computational Linguistics.

Summary

We aim to create an unsupervised cross-lingual embedding based on corpora from different domains. More precisely we align a biomedical English embedding with a general-domain non-English embedding under the hypothesis that monolingual data in the biomedical domain is mainly available in English. Our method for aligning two embeddings from different domains and languages relies on a proxy embedding of same domain as one embedding and same language as the other. The same-domain embeddings are aligned together in order to generate a dictionary for aligning the cross-domain embeddings. We evaluate our proposed algorithm on a biomedical translation task in several languages. While our method gives results below same-domain alignment approaches, it outperforms other cross-domain alignment techniques. This preliminary work ultimately intends to show that aligning different domains in an unsupervised manner is possible.

Keywords: word embedding, natural language processing, multilingual, unsupervised learning

Problème d'apprentissage supervisé en tant que problème inverse basé sur une fonction de perte L^1

Soufiane LYAQINI*, Mourad Nachaoui*, Mohamed QUAFAFOU**

*Université sultan moulay sliman, Béni Mellal, Maroc
** Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Résumé. Dans cet article, nous nous intéressons à l'étude et la résolution de certains problèmes d'apprentissage supervisé via leur formulation en problèmes inverses. La contribution principale réside dans l'utilisation des fonctions de pertes non forcément quadratiques (L^1 par exemple). Dans ce cadre, nous proposons une nouvelle technique d'approximation basée sur la formulation du problème inverse associé à un problème de minimisation d'une fonction de coût régulière, qui est résolu en utilisant l'algorithme de Newton et la régularisation de Tikhonov. Afin de valider notre approche, nous présentons deux types d'expériences numériques. Le premier consiste à approcher un problème d'apprentissage avec des données synthétiques. Tandis que le deuxième considère deux modèles d'apprentissage concrets. Notamment, le modèle de prédiction des signaux ECG. Les résultats numériques ainsi obtenus ont montré les bonnes performances des approches proposées.

Mots-clés : Apprentissage supervisé, régularisation de Tikhonov, ECG, algorithme de Newton.

1 Introduction

Dans ce travail nous nous intéressons au problème d'apprentissage supervisé. Dans ce contexte, nous notons qu'il existe une large gamme d'algorithmes d'apprentissage supervisé qui proposent en général un modèle de calcul qui est formellement décrit comme un problème d'optimisation. Ce dernier repose sur la minimisation des erreurs fonctionnelles sur des ensembles paramétrés de données d'entrée-sortie [1, 2, 3, 4, 5].

Récemment, certaines études se sont concentrées sur la définition d'une relation entre l'apprentissage supervisé et les problèmes inverses [6]. Tout d'abord, Kurkova [7] et Mukherjee et al. [8] ont montré que l'apprentissage supervisé, modélisé comme un problème d'optimisation qui consiste à minimiser une fonction d'erreur, peut être reformulé comme un problème inverse défini par des opérateurs d'évaluation et d'inclusion. Ensuite, De Vito et al. [9] ont établi un lien clair entre l'approche de consistance dans la théorie de l'apprentissage, la stabilité et la propriété de convergence dans des problèmes inverses mal posés, basé sur la théorie de la régularisation [10]. Cependant, cette connexion entre le problème d'apprentissage et les problèmes inverses a été faite dans le cas particulier où le problème inverse est reformulé comme

un problème de minimisation avec une fonction de coût quadratique (c'est-à-dire une fonction de coût L^2). Or, il est bien connu que la fonction de coût peut être L^1, L^2 ou toute fonction positive qui mesure l'écart entre les données prédites et celles observées. En effet, pour un problème d'apprentissage supervisé, l'utilisation de la fonction de perte L^1 donne des résultats plus performants (voir [11]). Cela consolide l'idée de reformuler le problème inverse en un problème de minimisation utilisant la fonction de coût L^1.

Dans ce travail nous étudions certains problèmes d'apprentissage supervisé via leur formulation en problèmes inverses. L'originalité réside dans l'utilisation des fonctions de pertes non forcément quadratiques (L^1 par exemple). Ceci a un impact considérable sur la performance et la qualité des solutions. En effet, l'utilisation de la fonction de perte L^1 dans des problèmes d'apprentissage supervisé donne des résultats plus pertinents par rapport à la fonction de perte L^2 généralement utilisée dans ce cadre. Cependant, la fonction de perte L^1 n'est pas différentiable, ce qui empêche l'utilisation d'outils d'optimisation standards. Pour surmonter cette difficulté, nous proposons une nouvelle technique d'approximation basée sur la reformulation du problème inverse associé en un problème de minimisation d'une fonction de coût oblique [12], qui est résolu en utilisant l'algorithme de Newton et la régularisation de Tikhonov [10]. Cette approche conduit à développer certains schémas numériques très efficaces permettant de résoudre des problèmes d'apprentissage supervisé dans le cadre le plus général.

2 Position du problème

Étant donné un ensemble d'apprentissage $Z = \{(x_i, y_i)\}_{i=1}^{\ell}$, où $x_i \in \mathcal{X} \subset \mathbb{R}^d$ appelé l'espace des entrées et $y_i \in \mathcal{Y}$ un sous ensemble de $\mathbb{R}$ appelé l'espace des sorties. Le problème de l'apprentissage supervisé est alors, à partir de l'ensemble d'apprentissage, de trouver un modèle $f : \mathcal{X} \longrightarrow \mathcal{Y}$, telle que $f(x)$ soit une bonne estimation de la sortie y quand une nouvelle entrée x est donnée dans $\mathcal{X}$. Par conséquent, le modèle optimal peut être obtenue en résolvant le problème de minimisation régularisé suivant

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x), y) + \lambda \|f\|_{\mathcal{H}}^2, \tag{1}$$

Où $\mathcal{H}$ désigne l'espace de Hilbert à noyau reproduisant [13], $\|.\|_{\mathcal{H}}$ la norme correspondante, λ est un paramètre de régularisation positif et $\mathcal{L}$ une fonction de perte pénalisant l'écart entre $f(x_i)$ et y_i.

Tout au long de cet article, nous utilisons la fonction de perte valeur absolue donnée par

$$\mathcal{L} = |f(x) - y|,$$

Dans ce contexte, le problème de minimisation (1) devient

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} |f(x) - y| + \lambda \|f\|_{\mathcal{H}}^2, \tag{2}$$

Pour surmonter les difficultés liées au fait que le terme de fidélité de problème de minimisation (2) n'est pas différentiable, nous optons par la suite pour l'approximation du terme de fidélité de problème (2) par une fonction deux fois différentiable et convexe.

3 Valeur absolue régularisée

Une manière classique pour résoudre un problème d'optimisation consiste à utiliser des méthodes du gradient de descente. Cependant, la fonction de perte valeur absolue dans le problème d'optimisation (2) n'est pas différentiable, on peut ainsi utiliser des méthodes d'optimisation globales, telles que la méthode du sous-gradient, qui ne nécessite aucune hypothèse de différentiabilité sur la fonction de coût. Cependant, ces méthodes sont très coûteuses au niveau temps de calcul [14, 15]. C'est pourquoi, nous optons pour une approximation de la fonction valeur absolue par une fonction régulière qui est deux fois différentiable et convexe. Cela nous permet en fait d'utiliser des méthodes de type gradient pour résoudre le problème d'optimisation résultant.

La fonction de perte valeur absolue peut être approchée par une fonction qui est deux fois différentiable et convexe, comme suit :

Considérons une régularisation de la fonction $\mathcal{L}(\omega) = |\omega|$ par la fonction régulière.

$$\mathcal{L}_\mu(\omega) = \sqrt{\omega^2 + \mu^2},$$

où $\mu > 0$ est un paramètre de régularisation.

On a $\mathcal{L}_\mu$ converge vers $\mathcal{L}$, lorsque μ tend vers zéro. Ceci est illustré en fait dans la figure suivante et sera démontré analytiquement par la suite.

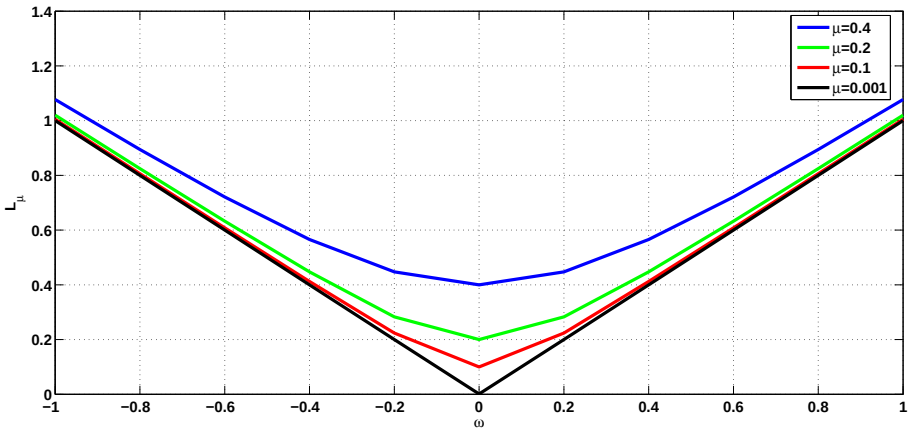


FIG. 1: Fonction de perte valeur absolue régularisée pour différents paramètres $\mu = 0.4$, $\mu = 0.2$, $\mu = 0.1$ et $\mu = 0.001$.

Nous allons maintenant montrer quelques propriétés mathématiques que $\mathcal{L}_\mu$ doit satisfaire pour être naturellement une fonction de perte.

Lemme 1. *i) $\mathcal{L}_\mu$ est une fonction localement lipschitzienne , c'est-à-dire pour tout $\omega_1, \omega_2 \in [-M, M]$, nous avons*

$$|\mathcal{L}_\mu(\omega_1) - \mathcal{L}_\mu(\omega_2)| \leq |\omega_1 - \omega_2|.$$

ii) Pour $\mathcal{Y} = [a, b]$, avec $a, b \in \mathbb{R}$, tel que $a < b$, nous avons

$$\mathcal{L}_\mu(y) \leq \beta + \mu, \quad \forall y \in \mathcal{Y}$$

où $\beta = \max\{|a|, |b|\}$.

Démonstration. Montrons $i)$, puisque $\mathcal{L}_\mu$ est une fonction différentiable sur $[-M, M]$, alors pour tous $\omega_1, \omega_2 \in [-M, M]$, nous avons

$$|\mathcal{L}_\mu(\omega_1) - \mathcal{L}_\mu(\omega_2)| \leq \sup_{\omega \in [-M,M]} |\mathcal{L}'_\mu(\omega)| \ |\omega_1 - \omega_1|,$$

avec

$$\mathcal{L}'_\mu(\omega) = \frac{\omega}{\sqrt{(\omega)^2 + \mu^2}}.$$

Donc il est claire que

$$\sup_{\omega \in [-M,M]} |\mathcal{L}'_\mu(\omega)| \leq 1.$$

Par conséquence

$$|\mathcal{L}_\mu(\omega_1) - \mathcal{L}_\mu(\omega_2)| \leq |\omega_1 - \omega_1|.$$

Montrons maintenant $ii)$, soit $y \in Y = [a, b]$, nous avons

$$\begin{aligned} \mathcal{L}_\mu(y) &= \sqrt{y^2 + \mu^2} \\ &\leq |y| + \mu \\ &\leq \beta + \mu, \end{aligned}$$

où $\beta = \max\{|a|, |b|\}$. □

Ainsi, notre approche vise à remplacer le problème de minimisation (2) par le problème de minimisation quadratique suivant

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\mu(f(x), y) + \lambda \|f\|^2_{\mathcal{H}}. \tag{3}$$

Dans la suite, nous allons montrer que le terme de fertilité du problème (2) $\mathcal{F}^\mu := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\mu(f(x), y)$ converge vers le terme de fertilité du problème (3) $\mathcal{F} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x), y)$, lorsque μ tend vers zero. Pour cela, nous introduisons le lemme suivant.

Lemme 2. *Pour tout $\omega \in \mathbb{R}$, on a*

$$\mathcal{L}(\omega) \leq \mathcal{L}_\mu(\omega) \leq \mathcal{L}(\omega) + \mu. \tag{4}$$

Démonstration. Ceci découle du fait que

$$\omega^2 \leq \omega^2 + \mu^2 \leq (|\omega| + \mu)^2.$$

Donc, en appliquant la fonction racine carrée, nous obtenons

$$|\omega| \leq \sqrt{\omega^2 + \mu^2} \leq |\omega| + \mu.$$

D'où

$$\mathcal{L}(\omega) \leq \mathcal{L}_\mu(\omega) \leq \mathcal{L}(\omega) + \mu.$$

□

En utilisant ce lemme, nous obtenons le résultat suivant

Proposition 1. *Pour tout $\mu > 0$, nous avons*

$$|\mathcal{F}^\mu[f] - \mathcal{F}[f]| \leq \mu. \tag{5}$$

Démonstration. A partir de l'équation (4), nous avons

$$\mathcal{L}(f(x_i) - y_i) \leq \mathcal{L}_\mu(f(x_i) - y_i) \leq \mathcal{L}(f(x_i) - y_i) + \mu, \quad \forall i = 1, ..., \ell.$$

Donc

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \mathcal{L}(f(x_i) - y_i) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} \mathcal{L}_\mu(f(x_i) - y_i) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} \mathcal{L}(f(x_i) - y_i) + \mu.$$

D'où

$$\mathcal{F}[f] \leq \mathcal{F}^\mu[f] \leq \mathcal{F}[f] + \mu.$$

Par conséquence

$$|\mathcal{F}^\mu[f] - \mathcal{F}[f]| \leq \mu.$$

□

Ce qui prouve que le risque empirique quadratique $\mathcal{F}^\mu$ converge vers $\mathcal{F}$ quand $\mu \to 0$.

Ensuite, nous allons utiliser les propriété des espaces de Hilbert à noyau reproduisant afin de pouvoir transformer le problème de minimisation (3) en un problème de minimisation non paramétrique.

4 Régression non paramétrique

Dans cette section, nous allons minimiser le problème (3) sur un espace de Hilbert à noyau reproduisant. L'avantage de l'optimisation dans un EHNR est que les solutions optimales peuvent être déterminées comme combinaison linéaire d'un nombre fini de fonctions de base, dont les coefficients sont inconnus, quelle que soit la dimension de l'espace $\mathcal{H}$ dans lequel l'optimisation est effectuée. Le résultat suivant connu sous le nom du théorème de représentation formalise cette notion (voir [16]).

Théorème 1. *Soint $\mathcal{H}$ un espace de Hilbert à noyau reproduisant K. Toute fonction $f \in \mathcal{H}$ minimisant le problème* (3) *admet une représentation de la forme :*

$$f(x) = \sum_{i=1}^{\ell} \alpha_i K(x_i, x), \quad \forall x \in \mathcal{X}, \tag{6}$$

où $\alpha_i \in \mathbb{R}$, pour tout $i = 1, ..., \ell$.

Ceci signifie que la solution du problème de minimisation (3) se représente dans le sous-espace vectoriel engendré par les fonctions noyaux dépendantes des données d'apprentissage : $K_{x_i}(.)$, pour $i = 1, ..., \ell$. Il s'ensuit que la solution peut être déterminée dans un espace dimension fini ℓ, même si l'espace $\mathcal{H}$ est lui-même de dimension infinie.

Ainsi en substituant (6) dans (3), on peut calculer la fonction optimale par optimisation numérique dans un espace de dimension finie. Ainsi, le problème d'optimisation (3) peut être réduit à trouver $\alpha = (\alpha_1, ..., \alpha_\ell)^T$ solution de

$$\min_{\alpha \in \mathbb{R}^\ell} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\left(\sum_{j=1}^{\ell} \alpha_j K(x_i, x_j) - y_i \right)^2 + \mu^2 \right)^{\frac{1}{2}} + \lambda \| \sum_{j=1}^{\ell} \alpha_j K_{x_j}(.) \|^2.$$

En posant alors

$$K = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_\ell) \\ \vdots & \ddots & \vdots \\ K(x_\ell, x_1) & \cdots & K(x_n, x_\ell) \end{pmatrix}, K_i = \begin{pmatrix} K_{x_i}(x_1) \\ \vdots \\ K_{x_i}(x_\ell) \end{pmatrix} \text{ et } y = \begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix},$$

la fonction de coût s'écrit

$$\mathcal{J}_\lambda^\mu(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\left(\alpha^\top K_i - y_i \right)^2 + \mu^2 \right)^{\frac{1}{2}} + \lambda \alpha^\top K \alpha.$$

Donc il suffit de résoudre le problème

$$\min_{\alpha \in \mathbb{R}^\ell} \mathcal{J}_\lambda^\mu [\alpha]. \tag{7}$$

Par la suite, nous allons proposer la validation numérique de la méthodologie de calcul proposée et aux preuves numériques pour étayer les résultats théorique.

5 Expérimentations

Pour évaluer un algorithme nouvellement conçu, il est souvent souhaitable d'avoir des ensembles de données de référence normalisés. Dans notre cas, nous choisissons d'évaluer la méthode proposée à l'aide d'exemples académiques et d'ensembles de données réels, qui modélisent certains problèmes concrets. L'objectif est de démontrer que notre approche couvre un

large spectre de modèle. Plus précisément, nous montrerons que contrairement aux approches classiques, qui ont besoin de connaissances préalables sur les modèles prévus, notre approche peut fidèlement prédire tous les modèles basés uniquement sur un ensemble de formation. La méthodologie de validation proposée consiste tout d'abord à considérer les modèles académiques antérieurs connus, de sorte que l'objectif est de les reconstruire. Ensuite, des modèles réels avec des données réelles sont considérés afin de montrer la capacité et l'efficacité de notre approche lorsqu'il s'agit de traiter des problèmes concrets.

5.1 Validation sur un problème synthétique

Nous considérons un problème dans lesquels le modèle est connu analytiquement. Nous donnons de brèves informations sur le problème de test dans cet exemple, et nous commençons à trouver le modèle basé sur un ensemble d'apprentissage. Ensuite, nous testons le modèle prévu sur un ensemble de test. En particulier, nous cherchons un modèle f défini de $\Omega \subset \mathbb{R}^\ell$ ($\ell > 1$) à valeurs dans $\mathbb{R}$. Dans le cinquième exemple numérique, on s'intéresse à l'approximation d'une fonction non régulière (avec un seul point singulier), définie comme suit :

$$f(x) = |x|\,, \quad x \in [-1, 1]. \tag{8}$$

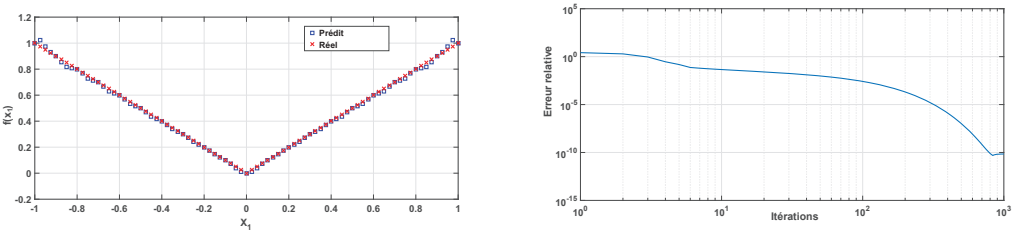
Tout d'abord, nous construisons la base de données en utilisant une discrétisation de $[-1, 1]$ en N éléments par

$$x_j^i = j\frac{2}{N}, \quad \text{et} \quad y_j = f(x_j), \quad x_j \in [-1, 1], \text{ pour } j = 0, \dots N, \tag{9}$$

où $N = 100$ est la taille de la base de données. L'ensemble d'apprentissage est construit comme suit

$$x_0 = -1,$$
$$x_{i+1} = x_i + \frac{1}{2}, \quad i = 1, ..., \ell - 1,$$

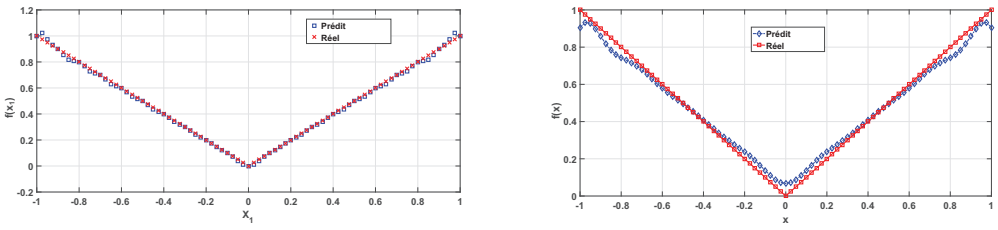
où $\ell = 21$ est la taille de l'ensemble d'apprentissage. Sur la base de l'ensemble d'apprentissage, nous avons prédit le modèle en résolvant le problème (7). La comparaison des données de sortie générées par le modèle prédit et celles tirées de la base de données construite (9) est présentée sur la figure 2. Comme on peut l'observer, le modèle prédit et l'exact donnent des courbes superposées



(a) (b)

FIG. 2: (a) Modèle réel et appris, pour $\sigma = 1$, $\mu = 10^{-3}$ et $\lambda = 10^{-3}$. (b) le risque empirique.

La figure 3 montre la comparaison entre notre approche et celle quadratique dans le cas où le modèle exact est décrit par une fonction qui a un seul point singulier. Comme on peut le voir, notre méthode donne une bonne approximation des fonctions non régulières que celle utilisant la fonction de perte quadratique.



(a) (b)

FIG. 3: Modèle réel et prévisionnel, pour $\sigma = 1$, $\mu = 10^{-3}$ et $\lambda = 10^{-3}$. (a) en utilisant la fonction de perte valeur absolue régularisée et (b) en utilisant la fonction de perte quadratique

5.2 Validation sur des données réelles

Dans cette section, nous allons étudier l'efficacité de l'algorithme proposé ci-dessus sur la prédiction des signaux ECG en se basant sur la base de données d'arythmie MIT-BIH.

L'organisation mondiale de la santé place les maladies cardiovasculaires (MCV) au premier rang des décès dans le monde. Ces MCV surviennent en raison de l'effet à long terme des arythmies cardiaques. Généralement, les arythmies cardiaques ne mettent pas souvent la vie en danger, mais peuvent entraîner une mort cardiaque ou une insuffisance cardiaque à long terme et doivent être détectées à temps. Un électrocardiogramme (ECG) mesure l'activité électrique du c ?ur et a été largement utilisé pour détecter les maladies cardiaques en raison de sa simplicité et de sa nature non invasive. En analysant le signal électrique de chaque rythme cardiaque, c'est-à-dire la combinaison de formes d'onde d'impulsion d'action produites par différents tissus cardiaques spécialisés trouvés dans le c ?ur, il est possible de détecter certaines de ses anomalies. De nos jours, il existe de nombreuses approches pour mesurer/enregistrer l'ECG. Da Silva et al. [17] ont fourni une taxonomie des méthodes de mesure ECG de pointe : en personne, sur personne et hors personne. Dans cette section, nous essayons de prédire les signaux ECG en se basant sur la base de données d'arythmie MITBIH. Cette base de données contient 48 extraits d'une demi-heure d'enregistrements ECG ambulatoires à deux canaux, obtenus à partir de 47 sujets étudiés par le laboratoire d'arythmie BIH entre 1975 et 1979.

Pour l'implémentation, nous avons sélectionné $2 \times 360\times$ éléments de temps, avec temps = 10. Puis, nous construisons notre base de données d'apprentissage en prenant 20 observations équidistribuées de toute base. Ensuite, en utilisant l'approche proposée, nous générons le modèle de signaux ECG.

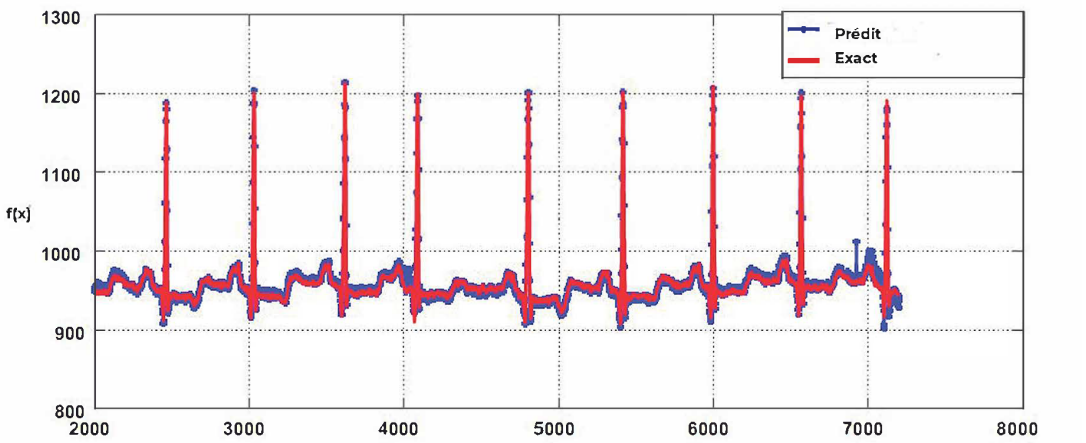


FIG. 4: Modèle réel et prévisionnel, pour $\sigma = 1.5$, $\epsilon = 10^{-3}$ et $\lambda = 10^{-3}$.

La figure 4 montre que le modèle prédit formé par cet ensemble d'apprentissage imite le vrai modèle avec tous ses pics. Comme pour le premier exemple, ces résultats nous garantissent la performance de la méthode proposée.

Le tableau ci-dessous répertorie l'erreur relative en fonction du nombre d'itérations pour ces deux exemples.

Ensemble de données set	Erreur relative
Exemple académique	0.0151
ECG	0.023

TAB. 1: L'erreur relative par rapport au nombre d'itérations pour ces deux exemples.

Les résultats quantitatifs présentés dans le tableau 1 confirment que la méthode proposée résout les problèmes avec plus de succès.

En conclusion, nous pouvons dire que la méthode proposée est un moyen plus efficace pour atteindre une solution optimale précise.

6 Conclusion

Dans ce travail, nous avons modélisé et résolu le problème d'apprentissage supervisé dans un cadre général. Cela se fait en le reformulant comme un problème inverse. Contrairement aux résultats existants, cette formulation est basée sur une fonction de perte L^1, qui est plus précise et donne un bon taux de convergence. Le problème inverse ainsi obtenu est défini par une fonction de coût non régulière avec un espace d'hypothèses générales. Cela nous a permis de développer des méthodes numériques rapides et efficaces, basées sur la technique d'approximation régulière, la régularisation de Tikhonov et la méthode de Newton. Dans les futures travaux nous allons adapter l'approche proposée aux problèmes d'apprentissage profond.

Références

[1] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995.

[2] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39 :1–49, 2002.

[3] Ali Emrouznejad. *Big Data Optimization : Recent Developments and Challenges*, volume 18. 2016.

[4] Soufiane Lyaqini, Mohamed Quafafou, Mourad Nachaoui, and Abdelkrim Chakib. Supervised learning as an inverse problem based on non-smooth loss function. *Knowledge and Information Systems*, pages 1–20, 2020.

[5] S Lyaqini, M Nachaoui, and M Quafafou. Non-smooth classification model based on new smoothing technique. In *Journal of Physics : Conference Series*, volume 1743, page 012025. IOP Publishing, 2021.

[6] Andreas Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Springer-Verlag, Berlin, Heidelberg, 1996.

[7] Vera Kurkova. Supervised learning as an inverse problem. page 1377â1384, 2004.

[8] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory : stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25 :161–193, 2006.

[9] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6, 2005.

[10] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed problems*. W.H. Winston, 1977.

[11] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same ? *Neural Comput.*, 16 :1063–1076, 2004.

[12] Hsinchun Chen, Roger H L Chiang, and Veda C. Storey. Business intelligence and analytics : From big data to big impact. *MIS Quarterly : Management Information Systems*, 36 :1165–1188, 2012.

[13] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68 :337–404, 1950.

[14] Claude Lemarechal. Nondifferentiable optimisation subgradient and εâsubgradient methods. In *Optimization and Operations Research*, pages 191–199. Springer, 1976.

[15] Wim van Ackooij and René Henrion. (sub-) gradient formulae for probability functions of random inequality systems under gaussian distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1) :63–87, 2017.

[16] K-R Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2) :181–201, 2001.

[17] AndreLourenco Ana Fred Rui Cesar das Neves Rui Ferreira Hugo Plecido da Silva, Carlos Carreiras. Off-the-person electrocardiography : performance assessment and clinical

correlation. *Health and Technolog*, 4 :pages 309â318, 2015.

Summary

In this paper, we deal with a supervised learning problem reformulated as an inverse problem. This inverse problem has been in fact reformulated into an unconstrained minimization one using an accurate smooth approximation of L1 loss function and solved by means of Newton's method. However, despite the quality of the obtained solution, this method still time consuming. In this work, based on this smooth approximation, we propose a theoretical validation of this proposed approach and then suggest a fast and efficient algorithm for solving this supervised learning problem. We also present some numerical results from academic and real-life datasets showing the efficiency of the proposed approach and confirming that it is better in terms of the stability and the convergence speed.

Keywords : Supervised learning, Tikhonov regularization, ECG, Newton Algorithm

Analyse statistique robuste et apprentissage profond à partir de séquences spectrales d'EEG pour la détection de somnolence

Antonio Quintero-Rincón*, Hadj Batatia**

*Departament of Electronic, Catholic University of Argentina (UCA), Buenos Aires, Argentina
antonioquintero@uca.edu.ar,
http://uca.edu.ar
**Heriot-Watt University, MACS School, Knowledge park, Dubai-Campus
h.batatia@hw.ac.uk
https://researchportal.hw.ac.uk/en/persons/hadj-batatia

Résumé. La somnolence des conducteurs est une cause majeure d'accidents de la route. L'électroencéphalogramme (EEG) est considéré comme le prédicteur le plus robuste de somnolence. Cet article propose une méthode nouvelle, simple et rapide pour détecter la somnolence de conducteurs, qui peut être implémentée en temps réel en utilisant une seule électrode. L'étude vise deux objectifs. Le premier consiste à déterminer le canal EEG unique le plus pertinent pour surveiller la somnolence. Cela est fait en procédant par analyse de covariance maximale. Le second objectif consiste à développer une méthode d'apprentissage profond à partir de ce canal. Pour cela, des caractéristiques spectrales du signal sont d'abord extraites. Un modèle de réseau récurrent à mémoire court et long terme (LSTM) est alors utilisé pour détecter les états de somnolence. Des expériences ont été conduites avec 12 signaux EEG pour discriminer les états de somnolence et d'alerte. Notre résultat principal est que le canal le plus significatif est TP7 situé dans la région temporo-pariétale gauche. Cela correspond à une zone partagée entre la conscience spatiale et la navigation spatiale visuelle. Ce canal est aussi relié à la faculté de prudence. En plus, malgré le petit nombre de données, la méthode proposée permet de prédire la somnolence avec une précision de 75% et un délai de 1.4 secondes. Ces résultats prometteurs mettent en lumière des données intéressantes à considérer pour la surveillance de la somnolence.

Mots-clés : analyse robuste, biLSTM, somnolence, EEG.

1 Introduction

L'électroencéphalogramme (EEG) est l'enregistrement de l'activité électrique du cerveau, détectée par des électrodes placées sur le cuir chevelu. Les variations des ondes cérébrales captées par l'EEG sont corrélées aux conditions neurologiques, aux états physiologiques et au

niveau de conscience. La somnolence des conducteurs est une cause majeure des accidents de la route, liée à 20% des cas (BalasuBramanian et Bhardwaj, 2018). Ces accidents causent plus de mortalité et morbidité que d'autres types d'accidents à cause de la grande vitesse au moment de l'impact (Horne et Reyner, 1995). Pour prévenir ces accidents, la détection de somnolence en temps réel chez les conducteurs est un besoin crucial. Cela permettrait le développement d'appareils d'alerte efficaces. Selon Lal et Craig (Lal et Craig, 2002), l'EEG peut être l'un des prédicteurs les plus efficaces de la somnolence. Par conséquent, un grand nombre de travaux de recherche ont étudié des méthodes de détection par analyse de signaux EEG.
Dans (Lal et Craig, 2002), les signaux EEG ont été analysés dans différents rythmes cérébraux et les résultats rapportés indiquent que les ondes delta (δ) et theta (θ) augmentent respectivement de 22% et 26% durant la transition vers l'état de somnolence. Ils indiquent aussi que l'activité des ondes alpha (α) et beta (β) n'augmente que de 9% et 5% respectivement. Partant de ces résultats, Lal et Craig (Lal et al., 2003) ont développé un algorithme robuste de détection de somnolence qui classe les bandes de fréquences en 4 phases : alerte, transition vers la somnolence, transition-posttransition, et posttransition.
Jap et al. (Jap et al., 2009) ont évalué ses quatre phases pendant la conduite monotone. Ils ont montré une augmentation du rapport onde-lente sur onde-rapide dans les activités EEG, avec $(\theta + \alpha)/\beta$ donnant l'augmentation la plus large. Wei et al. (Wei et al., 2012) ont transformé les données EEG en bandes (θ, α et β) et ils ont évalué 11 paramètres d'énergie pour déterminer l'indicateur optimal de somnolence. Ils ont sélectionné FP1 et O1 comme les électrodes significatives en utilisant une analyse en composantes principales à noyau (kACP). Ils rapportent une précision de 92%. Simon et al. (Simon et al., 2011) ont étudié les alpha spindles. Il s'agit de rafales courtes ($0.5 - 2.0$ s) d'activité alpha de haute fréquence (Lawhern et al., 2013). Les résultats rapportés indiquent que les paramètres de rafales alpha aboutissent à une augmentation de la sensibilité et la spécificité de détection de la somnolence.
Les méthodes d'entropie ont été étudiées par plusieurs auteurs pour détecter la somnolence à partir de signaux EEG. Hu (Hu, 2017) ont extrait des paramètres de quatre mesures d'entropie (entropie spectrale, entropie approximative, entropie d'échantillon, entropie floue) à partir d'un seul canal EEG. Dix classifieurs ont été utilisés pour comparer ces paramètres. Le classifieur forêts d'arbres décisionnels a donné la performance optimale en utilisant l'entropie floue du canal FP4, avec une précision de 96.6%. Mu et al. (Mu et al., 2017) ont combiné les mêmes 4 entropies calculées de signaux EEG enregistrés dans des états d'alerte et de somnolence. Les résultats (précision de 98.57%) montrent que la combinaison d'entropies donne une meilleure performance que l'utilisation d'une seule entropie. L'évaluation des différentes électrodes a montré que T5, TP7, TP8 et FP1 donnent de meilleurs résultats. Min et al. (Min et al., 2017) ont proposé la fusion de plusieurs entropies d'EEG et ont rapporté une précision de détection de somnolence de 98.3%, avec une sensibilité de 98.3%, et une spécificité de 98.2%. Ils ont aussi conclu que les canaux T6, P3, TP7, O1, Oz, T4, T5, FCz, FC3 et CP3 sont significatifs.
Il existe de nombreuses autres méthodes dans la littérature pour détecter les états cérébraux de conducteurs à partir de signaux EEG. Une grande partie de ces méthodes se base sur l'apprentissage automatique avec des vecteurs de paramètres très variés. Un état de l'art complet sur les techniques d'apprentissage automatique pour l'analyse de comportements de conducteurs peut être trouvé dans (Elassad et al., 2020).
Le réseau à mémoire court et long terme (LSTM) est un type de réseaux de neurones récurrents (RNN) capable d'apprendre des dépendances de long-terme dans les séries temporelles

(Hochreiter et Schmidhuber, 1997). La variante LSTM bidirectionnel (BiLSTM) est très efficace sur des données de longueur fixe (Schuster et Paliwal, 1997). Ses deux couches cachées, connectées aux entrées et sorties, permettent la circulation de l'information dans les deux sens. Les biLSTM ont été étudiés pour le traitement d'EEG par plusieurs chercheurs (Li et Jung, 2020; Hou et al., 2020; Yang et al., 2020; Fares et al., 2019).
Les méthodes d'analyse statistique monovariée et multivariée permettent l'étude de corrélation dans les données. Le sous-ensemble de techniques d'analyse robuste concernent les méthodes résilientes aux données aberrantes et applicables aux distributions non normales. Les méthodes classiques de moyenne et de covariance s'appliquent en situation de non-normalité mais ne sont pas robustes (Olive et al., 2017). Les méthodes d'analyse robuste ont été appliquées de nombreuses fois aux signaux EEG (Molinari et Dumermuth, 1986; Yong et al., 2008; Sameni et Seraj, 2017; Uehara et al., 2017). Ces signaux sont néanmoins considérés comme ayant une distribution normale (Quintero-Rincón et al., 2018), ce qui permet l'application de méthodes de statistiques optimales (Maronna et al., 2019). En neurosciences, l'analyse multivariée est employée pour l'étude simultanée d'états et d'interactions entre plusieurs régions cérébrales, alors que les méthodes monovariées se concentrent sur une région unique (Quintero-Rincón et al., 2016).
La présente étude propose une nouvelle méthode simple et rapide pour détecter la somnolence chez les conducteurs à l'aide d'un seul canal EEG. Elle se compose de deux étapes. La première consiste à déterminer le canal le plus pertinent à l'aide d'une méthode robuste d'analyse en maximum de covariance. La seconde étape, développe un réseau biLSTM pour différencier les états d'alerte et de somnolence, en utilisant des paramètres spectraux du canal déterminé (Hochreiter et Schmidhuber, 1997; Schuster et Paliwal, 1997). Plus précisément, on transforme la série temporelle de données EEG en une séquence de paramètres pour alimenter le réseau BiLSTM. Dans ce travail, nous avons analysé la somnolence à l'aide d'un ensemble de données composé de signaux EEG de 32 canaux provenant de 12 patients, fourni par Jiangxi University of Technology (Min et al., 2017). L'analyse en maximum de covariance a donné le canal TP7, situé dans la région temporo-pariétale, comme étant le plus pertinent (Fig. 1). Cette région concerne *la conscience spatiale et la navigation spatiale visuelle* en relation directe avec la conduite, et donc très pertinente pour l'analyse de somnolence. Ce canal est aussi relié à la faculté de prudence. Le réseau BiLSTM développé permet de détecter les changements entre alerte et somnolence durant des périodes longues de conduite. Les techniques mises en œuvre dans cette étude sont tous bien connues. Cependant, à notre connaissance, elles n'ont jamais été utilisées pour traiter le problème de la somnolence. La suite de l'article est organisée comme suit. La section 2 décrit la méthode proposée en 4 étapes : description des données (section 2.1), sélection du canal pertinent (section 2.2), transformation spectrale des données (section 2.3), et création du modèle prédictif (section 2.4). Dans la section 3 , nous analysons et discutons les résultats obtenus. Finalement, la section 4 esquisse des conclusions et des perspectives.

2 Processus de science de données

Un processus standard de science de données a été mis en œuvre pour cette étude (Fig.2). D'abord une phase de recherche et d'acquisition de données a été conduite. Puis, la seconde phase a consisté à prétraiter les données en deux étapes. D'abord, nous avons sélectionné l'électrode la plus pertinente pour l'analyse de somnolence. Ensuite, les données de ce canal ont été

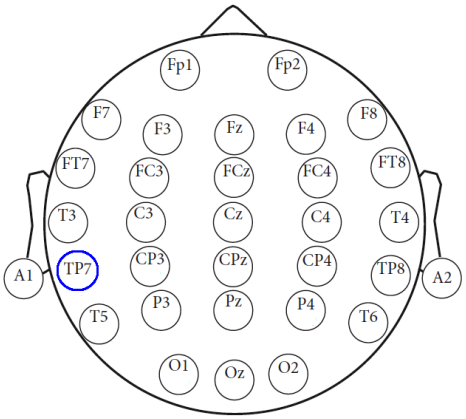


FIG. 1 – *Position et nomenclature des électrodes avec les lobes sous-jacents, (T) lobe temporal, (P) lobe pariétal, (O) lobe occipital et (F) lobe frontal.*

transformées en une séquence de paramètres spectraux. La troisième phase a consisté à apprendre un modèle prédictif pour détecter les états de somnolence. Ces phases sont présentées dans les sections suivantes.

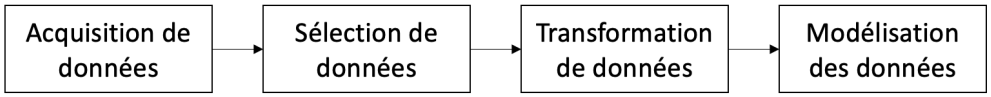


FIG. 2 – *Etapes de notre méthode.*

2.1 Description des données

Une base de données composée de 12 signaux unipolaires de somnolence et de 12 autres d'alerte a été fournie par Min et al. (2017). Les sujets concernés sont 12 étudiants âgés entre 19 et 24 ans. La tâche a consisté à conduire un véhicule statique à l'aide d'un simulateur de conduite dans un environnement contrôlé par logiciel. Pour chaque sujet, deux EEG ont été enregistrés : un correspondant à une conduite en état d'alerte, et le second à une conduite en état de somnolence. Les signaux correspondant à l'état d'alerte ont été enregistrés pendant une conduite de 15 minutes. Les signaux de conduite en état de somnolence ont été acquis quand les sujets ont montré des signes de fatigue selon l'échelle de Lee Lee et al. (1991) et l'échelle de Borg CR-10 Borg (1990), après une conduite entre 60 et 120 minutes. Les deux types de signaux consistent en 32 canaux de durée de 5 minutes, numérisés avec la fréquence $fs = 1000$ Hz.

2.2 Sélection de l'électrode pertinente

Soit $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ la matrice regroupant les M signaux EEG $\boldsymbol{x}_m \in \mathbb{R}^{1\times N}$ mesurés simultanément par différents canaux en N instants temporels. En appliquant l'algorithme MCD (Minimum cavariance determinant) (Rousseeuw et Driessen, 1999), nous estimons les paramètres $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ de la distribution uni-modale elliptique symétrique de nos données multivariées $\boldsymbol{X}$. Ces paramètres sont donnés par les expressions suivantes

$$\widehat{\boldsymbol{\mu}} = \frac{\sum_{i=1} W(d_i^2)\boldsymbol{x}_i}{\sum_{i=1} W(d_i^2)} \tag{1}$$

$$\widehat{\boldsymbol{\Sigma}} = c_0 \frac{1}{N} \sum_{i=1} W(d_i^2)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^*(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T \tag{2}$$

où $d_i = d(\boldsymbol{x}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ est la distance de Mahalanobis, $W(.) = I(. \leq \sqrt{\chi^2_{M,0.975}})$ est une fonction de poids avec Il'indicatrice, $*$ est le complexe conjugué, μ est la moyenne. Le paramètre c_0 appelé facteur de consistance est donné par $\alpha / F_{\chi^2_{M+2}}(q_\alpha)$, où q_α est le α-quantile de la distribution χ^2_M , avec $\alpha = \lim_{n\to\infty} h(n)/n$, et h choisi tel que $[(n+p+1)/2] \leq h \leq n$.
Afin d'estimer le canal le plus significatif $\boldsymbol{x}_p$, le maximum de covariance a été calculé pour des segments de 2 secondes à l'aide d'une fenêtre rectangulaire glissante avec 0.5 seconde de recouvrement, en résolvant

$$\boldsymbol{x}_p = \max_{[\boldsymbol{X}]^T} \widehat{\Sigma}.$$

Voir Hubert et Debruyne (2010); Rousseeuw et Driessen (1999); Maronna et al. (2019) pour plus de détail sur la méthode d'analys en maximum de covariance.

2.3 Calcul des paramètres spectraux

Nos analyses préliminaires ont montré que l'utilisation directe des données dans le domaine temporel ne permet pas une caractérisation efficace du phénomène de somnolence. Nous avons donc eu recours au domaine spectral.
Pour chaque état d'alerte et de somnolence, le signal $\boldsymbol{x}_p$ est partitionné en segments $\boldsymbol{x}_p(t)$ à l'aide d'une fenêtre rectangulaire glissante sans recouvrement. Pour chaque segment t, nous avons estimé deux paramètres, nommément la fréquence instantanée $f(t)$ et l'entropie spectrale instantanée $H(t)$ selon l'approche suivante.
Soit $S(t,f) = |X_p(t,f)|^2$ le spectrogramme du signal $\boldsymbol{x}_p(t)$ avec $\boldsymbol{X}_p(t,f)$ la transformée de Fourier discrète de $\boldsymbol{x}_p(t)$. La distribution de probabilité du spectrogramme à l'instant t est donnée par

$$P(t,m) = \frac{S(t,m)}{\sum_f S(t,f)}$$

On peut ainsi calculer la fréquence instantanée du signal à l'aide de la transformée de Fourier à temps court Boashash (1992a,b)

$$f(t) = \frac{\sum_m mP(t,m)}{\sum_m p(t,m)}$$

Par ailleurs, on calcule l'entropie spectrale instantanée selon Pan et al. (2009)

$$H(t) = -\sum_{m} P(t,m) log_2 P(t,m)$$

Nous avons ainsi transformé le signal du canal pertinent en une séquence de pairs de paramètres $(f(t), H(t))$. Cette séquence sera considérée pour l'apprentissage d'un modèle de classification en alerte et somnolence.

2.4 Modèle prédictif

Afin de différencier les états d'alerte et de somnolence, nous avons conçu un classifieur de réseau récurrent à mémoire court et long terme bidirectionnel (BiLSTM). Ce réseau prend en entrée la séquence de paramètres spectraux $\Theta_t = (f(t), H(t))$ extraits à l'étape précédente. Les paramètres ont tous été normalisés selon l'approche z-score, en soustrayant la moyenne et en divisant par l'écart type sur l'ensemble des segments.
Selon l'approche standard, l'architecture du réseau consiste en trois portes entrée, sortie et oublie (input, output, forget) (Calin, 2020). La porte d'oubli est la sigmoïde

$$F_t = \sigma(W_F \Theta_t + U_F h_{t-1} + b_F)$$

où les hyper-paramètres W_F and U_F sont des matrices, et b_F un vecteur de biais. σ est la fonction d'activation sigmoïde. Ainsi, $F_t \in (0,1)$ représente la fraction de l'état passé qui sera oubliée, dépendant de l'état précédent h_{t-1} et de l'entrée Θ_t à l'itération t. De la même manière, la porte d'entrée est définie par

$$I_t = \sigma(W_I \Theta_t + U_I h_{t-1} + b_I)$$

avec les hyper-paramètres W_I, U_I et b_I. Et la porte de sortie

$$O_t = \sigma(W_O \Theta_t + U_O h_{t-1} + b_O)$$

avec ses hyper-paramètres W_O, U_O et b_O. La fonction de mise à jour met à jour sélectivement la mémoire interne du réseau avec

$$c_t = F_T \circledast c_{t-1} + I_t \circledast tanh(W_c \Theta_t + U_c ht - 1 + b_c)$$

avec les hyper-paramètres W_c, U_c et b_c, avec $h_t = O_t \circledast tanh(c_t)$. Le réseau traite une séquence de 10520 pairs de paramètres. Il est composé de 100 cellules. Une couche bidirectionnelle a été ajoutée au réseau pour permettre la prise en considération de la dépendance avec les séquences futures. Une couche entièrement connectée permet quant à elle de réduire l'échelle de 100 à deux dimensions. La classification en alerte ou somnolence est obtenue à l'aide d'une couche de fonction exponentielle normalisée.

3 Résultats et discussion

Dans cette section nous rapportons les résultats de notre méthode appliquée au canal TP7 de la banque de données EEG décrites dans la section 2.1. La figure 3 illustre des exemples

de signaux des états alerte et somnolence du canal TP7, dans le domaine temporel. La grande dynamique des signaux EEG empêche la distinction des deux états visuellement. La première phase de notre méthode a établi que le canal TP7 contient les informations pertinentes pour surveiller la somnolence chez les conducteurs. Ce canal situé sur la région tempo-pariétale est lié au lobe associé à la conscience spatiale et la navigation visuelle spatiale (Fig. 1). Par ailleurs, il est connu que la structure physique du cerveau reflète son organisation mentale. En général, les processus mentaux supérieurs se déroulent dans les régions supérieures, alors que les régions inférieures se chargent des fonctions de support (Carter, 2019). Ainsi, notre résultat est tout à fait crédible, du fait que le canal déterminé est lié à la conscience et aux facultés affectives (Postle, 2020). Notre hypothèse est que la mélatonine est responsable de la grande variabilité des formes des ondes du signal dans le domaine temporel (Fig. 3). Cette hormone produite par la glande pinéale aide à réguler les cycles sommeil-éveil. Selon (Carter, 2019), le noyau suprachiasmatique (NSC) situé dans l'hypothalamus joue un rôle clé dans les cycles de sommeil-éveil. Le niveau de lumière est perçu par la rétine et cette information est reliée au NSC qui à son tour envoie un signal à la glande pinéale. Cela entraîne la sécrétion de la mélatonine, qui met le corps en état de sommeil.

Partant de ce résultat, les données du canal TP7 ont été utilisées pour la détection de la somnolence. Les données de tous les sujets ont été traitées selon la méthode décrite dans la section 2, et transformées en séquences spectrales. Les séquences par sujet ont été partitionnées en 80% pour l'entrainement du réseau biLSTM et 20% pour le test (les sujets utilisés pour le test sont entièrement différents des sujets utilisés pour l'entrainement). Ce processus de tirage aléatoire a été répété 100 fois et les performances ont été moyennées. Il est à préciser que nos données sont équilibrées, ayant le même nombre de sujets pour chaque classe et ne nécessite aucune pondération (Fernández et al., 2018; Quintero-Rincón et al., 2019).
Plusieurs métriques sont communément utilisées pour évaluer les performances de classifieurs. Récemment, Elassad et al. (Elassad et al., 2020) ont analysé ces métriques dans de nombreuses études et ont trouvé que la précision, la sensibilité et la spécificité sont les plus utilisées. Selon cette étude la précision a été utilisée par 65.85% des 82 études considérées publiées entre 2009 and 2019, alors que la sensibilité et la spécificité ont été utilisées respectivement à 35.36% et 32.92%. En conformité, et pour permettre une lisibilité de nos résultats et leur comparaison avec les méthodes de l'état de l'art, nous rapportons nos résultats en termes de précision, sensibilité et spécificité.
La phase de test de notre méthode a donné une précision de 75% dans la distinction des états d'alerte et de somnolence, avec une sensibilité de 67.7% et une spécificité de 86%. Un test ANOVA a donné une p-value < 0.01. Le temps de prédiction a été de 1.4 secondes.

Ces résultats ont été comparés avec des méthodes d'apprentissage automatique statistiques, arbre décision, Bayes naïf, machine à support de vecteurs, les k-plus proches voisins, et ensemble d'arbres de décision boostés. Ces méthodes ont donné de meilleures précisions, sensibilités et spécificités, mais un délai de prédiction nettement plus large que celui du biLSTM. La table 1 synthétise ces résultats comparés. Le paramètre k du k-NN a été fixé à $\sqrt{N}$ avec N le nombre de sujets dans les données, dans notre cas $N = 10$ et donc $k \approx 3$. La table 2 rapporte des exemples de canaux utilisés dans différents travaux. Par exemple, dans (Wei et al., 2012) l'analyse relationnelle de Grey (ARG) et l'analyse en composantes principales à noyau (kACP) ont été utilisées pour extraire les canaux pertinents et les données ont été classées à l'aide d'une régression linéaire. Hu (Hu, 2017) a comparé l'efficacité de 4 paramètres en utili-

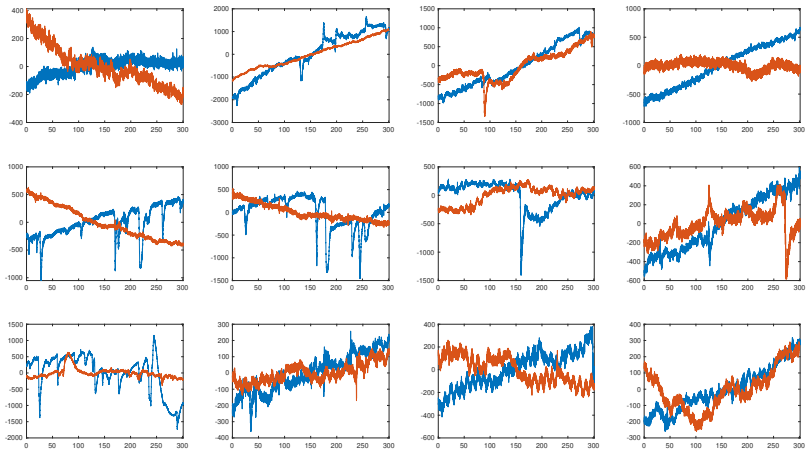


FIG. 3 – *Exemples de signaux EEG de l'électrode TP7 correspondant à des états d'alerte (bleu) et somnolence. On observe une grande variabilité du signal dans le domaine temporel.*

sant 10 classifieurs différents pour détecter la somnolence. Il a conclu que l'entropie floue avec un classifieur de forêt d'arbres décisionnels donne la meilleure performance avec une précision de 96.6%. Dans (Mu et al., 2017), une combinaison d'entropies avec un classifieur de machine à vecteurs de support a abouti à une précision de 98.75%, sensibilité de 97.50% et spécificité de 96%. Min et al. (Min et al., 2017) ont obtenu une précision de 98.3% avec leur fusion de plusieurs entropies et classifieur par réseau de neurones. Yeo et al. (Yeo et al., 2018) ont utilisé des paramètres du spectre de puissance à partir de 19 canaux avec un SVM. Ils ont rapporté une précision de 99.3% dans la distinction d'états d'alerte et de somnolence. Finalement, Chen et al. (Chen et al., 2009) ont fusionné deux paramètres, un paramètre du réseau cérébrale fonctionnel et un autre de la densité du spectre de puissance. Ils ont utilisé un classifieur de type extreme learning machine (ELM) pour détecter la somnolence, et ont rapporté une précision de 95%, une sensibilité de 95.71% et une spécificité de 94.29%

TAB. 1 – *Comparaison des résultats du biLSTM avec des méthodes conventionnelles. Précisions (Pré.) en %, Sensibilité (Sen.) en %, et Spécificité (Spé.) en %, et délai de prédiction (délai) en secondes*

Techniques	Pré.	Sen.	Spé.	Délai
biLSTM	75.0	67.7	86.0	1.40
Arbre de décision	84.8	83.0	86.0	6.60
Naive Bayes	85.1	84.0	86.0	19.80
SVM	84.2	80.0	89.0	16.40
k-NN	85.3	85.0	860	5.82
Ensemble d'arbres boostés	85.4	84.0	87.0	6.26

Malgré leur bonne performance, toutes ces méthodes ont une complexité calculatoire élevée, liée principalement à la transformation des données. A l'opposé, notre méthode a des résultats satisfaisants avec un seul canal. Elle est simple à implémenter en temps réel, ce qui est l'objectif ultime dans cette application.

4 Conclusions

Cet article a présenté la mise en œuvre complète d'un processus de science de données pour élaborer un modèle prédictif de la somnolence chez les conducteurs à partir de données EEG. La méthode a consisté à analyser les signaux EEG à l'aide d'une méthode d'analyse statistique robuste par maximum de covariance pour déterminer un canal unique porteur de l'information utile à la tâche. Cette analyse à abouti à l'identification du canal TP7 situé dans la région tempo-pariétale comme étant le plus pertinent. Cette région correspondant à la conscience spatiale et à la navigation visuelle spatiale, très pertinente pour surveiller la somnolence, rend le résultat très probant.
Une seconde contribution de l'article est la transformation de la série temporelle des signaux EEG en séquences de paramètres spectraux. Ces paramètres révèlent les caractéristiques relatives des états d'alerte et de somnolence pendant la conduite. Un classifieur de type réseau de neurones récurrent à mémoire de court et long terme a été entraîné avec ces séquences de paramètres. Des expérimentations ont été faites avec des données réelles avec 24 signaux EEG obtenus de Jiangxi University of Technology (Min et al., 2017). Les résultats ont montré une précision de 75% à distinguer les états d'alerte et de somnolence, avec un délai de prédiction de 1.4 seconde. L'avantage principale de la méthode proposée est sa simplicité lui conférant la capacité d'être implémentée en temps réel. En plus, l'utilisation d'un canal unique rend la mise en œuvre pratique plus facile.
La limitation principale de la méthode est le volume de données faible. Une des perspectives de ce travail est de développer une méthode d'augmentation de données à l'aide d'un réseau antagoniste génératif (GAN) (Radford et al., 2016). . Une autre perspective est d'intégrer la phase de transformation spectrale dans le réseau d'apprentissage profond en utilisant and LSTM convolutionnel (Abdelhameed et al., 2018).

Acknowledgement

Les auteurs remercient María Eugenia Fontecha pour ses commentaires utiles sur une versions préliminaire de cet article.

Références

Abdelhameed, A. M., H. G. Daoud, et M. Bayoumi (2018). Deep convolutional bidirectional lstm recurrent neural network for epileptic seizure detection. In *2018 16th IEEE International New Circuits and Systems Conference (NEWCAS)*, pp. 139–143. IEEE.

BalasuBramanian, V. et R. Bhardwaj (2018). Can cECG be an unobstrusive surrogate to determine cognitive state of driver ? *Transportation Part F 58*, 797–806.

TAB. 2 – *Quelques exemples de canaux utilisés dans différents travaux. La plupart des méthodes donnent une bonne précision (Pré.). Les techniques de sélection de carnaux sont diverses : Robust analysis (RUA), Grey relational analysis (GRA), Kernel Principal Component Analysis (kPCA). Various features : Sample Entropy (SA), Spectral entropy (SE), Approximate entropy (AE), Fuzzy entropy (FE), Entropy fusion (EF), Power spectral density (PSD), Functional brain network (FBN). And different classifiers : Long Short Term Memory (LSTM), Random Forest (RF), Support vector machine (SVM), Back propagation Neural network (BPNN), Extreme learning machine (ELM).*

Canaux	Méthode	Classifier	Pré.	Ref.
TP7	RA	biLSTM	75.0%	our
FP1, O1	GRA+KPCA	LRE	92.3%	(Wei et al., 2012)
CP4	SE+FE+ AE+PE	RF	96.6%	(Hu, 2017)
T5, TP7, TP8, FP1	SE+AE+SE+FE	SVM	98.7%	(Mu et al., 2017)
T6, P3, TP7, O1, Oz, T4, T5, FCz, FC3, CP3	EF	BP	98.3%	(Min et al., 2017)
tous	PSD	SVM	99.3%	(Yeo et al., 2018)
tous	FBN-PSD	ELM	95.0%	(Chen et al., 2009)

Boashash, B. (1992a). Estimating and interpreting the instantaneous frequency of a signal-part 1 : Fundamentals. *Proceedings of the IEEE 80*, 520–538.

Boashash, B. (1992b). Estimating and interpreting the instantaneous frequency of a signal-part 2 :algorithms and applications. *Proceedings of the IEEE 80*(4), 540–568.

Borg, G. (1990). Psychophysical scaling with applications in physical work and the perception of exertion. *Scandinavian Journal of Work 16*(1), 55–58.

Calin, O. (2020). *Deep learning architectures. A mathematical approach*. Springer.

Carter, R. (2019). *The Human Brain Book : An illustrated guide to its structure, function, and disorders*. Penguin.

Chen, J., H. Wang, C. Hua, et E. P. V. Wilder-Smith (2009). Electroencephalography based fatigue detection using a novel feature fusion and extreme learning machine. *Cognitive Systems Research 52*, 115–124.

Elassad, Z. E. A., H. Mousannif, H. A. Moatassime, et A. Karkouch (2020). The application of machine learning techniques for driving behavior analysis : A conceptual framework and a systematic literature review. *Engineering Applications of Artificial Intelligence 87*(10331), 2.

Fares, A., S.-h. Zhong, et J. Jiang (2019). EEG-based image classification via a region-level stacked bi-directional deep learning framework. *BMC medical informatics and decision making 19*(6), 1–11.

Fernández, A., S. García, M. Galar, R. C. Prati, B. Krawczyk, et F. Herrera (2018). *Learning from Imbalanced Data Sets*, Volume 11. Springer.

Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Horne, J. et L. Reyner (1995). Sleep related vehicle accidents. *BMJ Clinical Research 310*(6979), 565–567.

Hou, Y., S. Jia, S. Zhang, X. Lun, Y. Shi, Y. Li, H. Yang, R. Zeng, et J. Lv (2020). Deep feature mining via attention-based bilstm-gcn for human motor imagery recognition. *ArXiv*, 1–8. arXiv :2005.00777.

Hu, J. (2017). Comparison of different features and classifiers for driver fatigue detection based on a single EEG channel. *Computational and Mathematical Methods in Medicine 2017*(51095), 30.

Hubert, M. et M. Debruyne (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews : Computational statistics 2*(1), 36–43.

Jap, B. T., S. Lal, P. Fischer, et E. Bekiaris (2009). Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications 36*(2), 2352–2359.

Lal, S. et A. Craig (2002). Driver fatigue : Electroencephalography and psychological assessment. *Phychophysiology 39*(3), 313–321.

Lal, S. K., A. Craig, P. Boord, L. Kirkup, et H. Nguyen (2003). Development of an algorithm for an EEG-based driver fatigue countermeasure. *Journal of Safety Research 34*(3), 321–328.

Lawhern, V., S. Kerick, et K. Robbins (2013). Detecting alpha spindle events in EEG time series using adaptive autoregressive models. *BMC Neuroscience 14*(101), 1–9.

Lee, K. A., G. Hicks, et G. Nino-Murcia (1991). Validity and reliability of a scale to assess fatigue. *Psychiatry Research 36*(3), 291–298.

Li, G. et J. Jung (2020). Maximum marginal approach on EEG signal preprocessing for emotion detection. *Applied Sciences 10*(7677), 1–11.

Maronna, R. A., R. D. Martin, V. J. Yohai, et M. Salibián-Barrera (2019). *Robust statistics Theory and methods with R*. Wiley.

Min, J., P. Wang, et J. Hu (2017). Driver fatigue detection through multiple entropy fusion analysis in an EEG-based system. *Plos One 12*(12), e0188756.

Molinari, L. et G. Dumermuth (1986). Robust multivariate spectral analysis of the EEG. *Neuropsychobiology 15*(3-4), 208–218.

Mu, Z., J. Hu, et J. Min (2017). Driver fatigue detection system using electroencephalography signals based on combined entropy features. *Applied Sciences 7*(2), 150–167.

Olive, D. J., D. J. Olive, et Chernyk (2017). *Robust Multivariate Analysis*. Springer.

Pan, Y., J. Chen, et X. Li (2009). Spectral entropy : a complementary index for rolling element bearing performance degradation assessment. *Proceedings of the Institution of Mechanical Engineering Science, Part C : Journal of Mechanical Engineering Science 223*(5), 1223–1231.

Postle, B. R. (2020). *Essentials of Cognitive Neuroscience*. John Wiley & Sons.

Quintero-Rincón, A., M. Flugelman, J. Prendes, et C. d'Giano (2019). Study on epileptic seizure detection in EEG signals using largest lyapunov exponents and logistic regression.

Revista Argentina de Bioingeniería 23(2), 17–24.

Quintero-Rincón, A., M. Pereyra, C. D'Giano, M. Risk, et H. Batatia (2018). Fast statistical model-based classification of epileptic EEG signals. *Biocybernetics and Biomedical Engineering 38*(4), 877–889.

Quintero-Rincón, A., J. Prendes, M. Pereyra, H. Batatia, et M. Risk (2016). Multivariate bayesian classification of epilepsy EEG signals. In *2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pp. 1–5. IEEE.

Radford, A., L. Metz, et S. Chintala (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio et Y. LeCun (Eds.), *4th International Conference on Learning Representations*.

Rousseeuw, P. J. et K. V. Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*(3), 212–223.

Sameni, R. et E. Seraj (2017). A robust statistical framework for instantaneous electroencephalogram phase and frequency estimation and analysis. *Physiological Measurement 38*(12), 2141–2163.

Schomer, D. L. et F. H. L. da Silva (Eds.) (2017). *Niedermeyer's Electroencephalography : Basic Principles, Clinical Applications, and Related Fields* (7th ed.). Oxford University Press.

Schuster, M. et K. Paliwal (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing 45*(11), 2673–2681.

Simon, M., E. A. Schmidt, W. E. Kincses, M. Fritzsche, A. Bruns, C. Aufmuth, M. Bogdan, W. Rosenstiel, et M. Schrauf (2011). EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions. *Clinical Neurophysiology 122*(6), 1168–1178.

Uehara, T., M. Sartori, T. Tanaka, et S. Fiori (2017). Robust averaging of covariances for EEG recordings classification in motor imagery brain-computer interfaces. *Neural Computation 29*(6), 1631–1666.

Wei, L., H. Qi-chang, F. Xiu-min, et F. Zhi-min (2012). Evaluation of driver fatigue on two channels of EEG data. *Neuroscience Letters 506*(2), 235–239.

Yang, J., X. Huang, H. Wu, et X. Yang (2020). EEG-based emotion classification based on bidirectional long short-term memory network. *Procedia Computer Science 174*, 491–504.

Yeo, M. V. M., X. Li, K. Shen, et E. P. V. Wilder-Smith (2018). Can svm be used for automatic EEG detection of drowsiness during car driving ? *Safety Science 47*(1), 715–728.

Yong, X., R. K. Ward, et G. E. Birch (2008). Robust common spatial patterns for EEG signal preprocessing. In 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2087–2090. IEEE.

Annexe : Lobes du cerveau

Dans cette annexe courte, nous fournissons quelques informations sur la structure du cerveau. Ces informations sont nécessaires pour interpréter et évaluer les résultats de n'importe quel modèle d'apprentissage automatique sur l'EEG. Dans notre cas, il est important d'inter-

préter la localisation du canal trouvé pertinent par notre analyse statistique pour la surveillance de la somnolence chez les conducteurs.

La tâche principale du cerveau est d'aider à garder l'organisme dans un état optimal étant donné l'environnent afin de maximiser les chances de survie. Lors de ce processus, le cerveau classifie les informations pertinentes arrivant comme des impulsions électriques depuis les neurones au niveau des organes sensoriels. Cette activité électrique est amplifiée et représentée dans les régions cérébrales frontale, pariétale, occipitale et temporale. Quand cette activité persiste une durée significative, il en résulte une expérience consciente (Carter, 2019).

La figure 1 montre les positions d'un système EEG du standard international 10-20. Les électrodes sont étiquetées en fonction du lobe lié à leur position et à leur l'hémisphère. Par exemple F = frontal, P = pariétal, O = occipital, T = temporal, C = central, z : la fissure longitudinale, nombres impairs pour la gauche et nombres pairs pour la droite.
Les fonctions principales du lobe frontal sont le contrôle du mouvement, le raisonnement, l'émotion consciente, le langage, la planification, la pensée, le jugement, et la décision. Le lobe pariétal est associé à l'intégration des stimuli sensoriels et perceptuels aussi bien conscient qu'inconscient, comme le calcul spatial, l'orientation du corps, et l'attention. Le lobe pariétal est lié à la mémoire, l'ouïe, le langage, l'émotion, et la navigation visuelle spatiale. Le lobe occipital est responsable du traitement visuel par intégration des informations visuelles aussi bien conscientes qu'inconscientes. Le sillon central sépare les lobes frontal et pariétal. Il est relié au cortex moteur primaire et au cortex somesthésique primaire. Le sillon latéral sépare les lobes latéral et frontal du lobe temporal. Il est associé à la production du langage. Se référer à (Schomer et da Silva, 2017) pour plus d'information sur la structure du cerveau.

Summary

Driver fatigue is a major cause of traffic accidents. Electroencephalogram (EEG) is considered one of the most reliable predictors of fatigue. This paper proposes a novel, simple and fast method for driver fatigue detection that can be implemented in real-time by using a single channel on the scalp. The study consists of two objectives. First, determine the most significant EEG channel to monitor fatigue using maximum covariance analysis. And second, develop a machine learning method to detect fatigue from this single channel using a Long Short-Term Memory (LSTM) deep learning model based on spectral features. Experiments with 12 EEG signals were conducted to discriminate the fatigue stage from the alert stage. Our main discovery was that the most significant channel found (TP7) is located in the left tempo-parietal region where spatial awareness and visual-spatial navigation are shared. This channel is also related to the cautiousness faculty. In addition, despite the small dataset, the proposed method yielded 75% accuracy for fatigue prediction with a 1.4-second delay. These promising results provide new insights on relevant data for monitoring driver fatigue.
Keywords: Robust analysis, biLSTM, Driver fatigue, EEG

Analyse automatique du discours de patients pour la détection de comorbidités psychiatriques

Christophe Lemey*,**, Yannis Haralambous**, Philippe Lenca**,
Romain Billot**, Deok-Hee Kim-Dufor***

*Service hospitalo-universitaire de psychiatrie adulte, CHRU de Brest
christophe.lemey@chu-brest.fr,
**IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
prenom.nom@imt-atlantique.fr
***EA 7479 SPURBO, Université de Bretagne Occidentale, Brest, France

Résumé. Les comorbidités sont très fréquentes en santé mentale et représentent un enjeu thérapeutique majeur ainsi qu'un levier pour une meilleure compréhension des mécanismes physiopathologiques des pathologies. Certains maladies, comme la schizophrénie, s'installent progressivement et de façon variable d'un individu à l'autre, les symptômes augmentant progressivement en intensité et en spécificité. Dans ces cas, les cliniciens cherchent à identifier au plus tôt les symptômes annonciateurs et les comorbidités associés, afin de proposer des interventions maximisant les effets thérapeutiques. La parole, et donc le langage, est un élément-clé sur lequel ils s'appuient lors des consultations pour comprendre l'état psychique des patients, et les systèmes d'analyse automatique de la langue peuvent fournir une aide à l'évaluation. Nous proposons une telle aide, ciblant la détection de comorbidités et fondée sur les grammaires de dépendances et des indicateurs paralinguistiques comme les pauses et les interjections, qui s'avèrent être des choix pertinents.

Mots-clés : comorbidités psychiatriques, schizophrénie, traitement automatique du langage, grammaires de dépendances, paralinguistique

1 Introduction

Une comorbidité est l'association de deux ou plusieurs maladies ou troubles dans le même temps. En santé mentale, une étude ((Roca et al., 2009)) a revelé que jusqu'à 30% des patients dans une cohorte nationale présentent des comorbidités, L'exploration des comorbidités représente un double enjeu, à la fois thérapeutique et étiologique. En effet, les comorbidités compliquent la mise en œuvre des traitements et leur étude permet de mieux comprendre les mécanismes physiopathologiques mis en jeu dans la génèse des troubles.

Parmi les principales maladies psychiatriques, la schizophrénie touche environ 1% de la population et est l'une des principales maladies entraînant un nombre important d'années vécues avec un handicap (Rössler et al., 2005; Anderson, 2019), en grande partie en raison du jeune âge auquel elle se déclare (souvent à l'adolescence), du poids élevé des handicaps (fonctionnels et sociaux) et de l'évolution chronique fréquente de la pathologie. Début 2020, en

France, environ 600 000 personnes souffraient de cette maladie [1], et il est estimé que parmi elles, une sur deux fera au moins une tentative de suicide, à court ou moyen terme.

La schizophrénie s'installe progressivement. Le cours évolutif de la pathologie est caractérisé par les phases suivantes : phase prémorbide, de la naissance du patient jusqu'à l'apparition des premiers signes ; phase prodromique au cours de laquelle apparaissent les premiers symptômes peu spécifiques, ces symptômes augmentant progressivement en intensité et en spécificité au cours de la phase qui précède les symptômes psychotiques francs ; phase psychotique avec les premiers signes psychotiques avérés qui déterminent le premier épisode de psychose. Lors de la phase active de la schizophrénie on constate une multitude de symptômes très variables (syndromes positifs : idées délirantes et hallucinations ; syndromes négatifs : retrait social et déficits cognitifs ; syndrome de désorganisation : trouble du contact). Il s'agit d'une maladie complexe dont la physiopathologie reste peu connue. Le modèle explicatif dominant actuellement, est le modèle de diathèse-stress qui combine deux facteurs : la vulnérabilité intrinsèque et le stress provenant d'expériences vécues (Howes et McCutcheon, 2017; Pruessner et al., 2017). Néanmoins, les mécanismes sous-jacents doivent encore être explorés. Lors de ces phases précoces d'évolution de la maladie, la présence de comorbidités est très fréquente (notamment les comorbidités anxieuses, dépressives et addictives). Jusqu'à 50% des patients en phase prodromique peuvent présenter une comorbidité (Lim et al., 2015).

La durée entre l'apparition des premiers symptômes psychotiques francs et le premier accès aux soins est en moyenne de deux à cinq ans (avec d'importantes différences entre régions du monde). Cette période est communément appelée «durée de la psychose non traitée» (Fusar-Poli et al., 2013). Les efforts vont dans le sens d'un traitement précoce et d'une réduction de cette durée. En effet, l'identification précoce et les interventions rapides au cours de l'évolution d'un trouble psychotique semblent maximiser les effets thérapeutiques et améliorer la qualité de vie des patients (McGlashan et Johannessen, 1996). Durant cette phase, des signes d'alerte avant la phase active de la maladie peuvent être détectés, ce qui permet d'optimiser les soins et de réduire la durée de la psychose non traitée (Olsen et Rosenbaum, 2006) en orientant les patients vers des centres de détection précoce des troubles psychotiques utilisant des outils spécifiques d'évaluation (Olsen et Rosenbaum, 2006; Yung et al., 2005).

Parmi ces outils d'aide à l'évaluation, et de façon générale en psychiatrie, se trouvent les systèmes d'analyse automatique de la parole et, plus particulièrement, du discours des patients (Le Glaz et al., 2021). En effet, le langage est l'un des éléments clés sur lequel les cliniciens peuvent s'appuyer lors des consultations pour mieux comprendre l'état psychique des patients (Mota et al., 2012). Ainsi, les psychiatres étudient tout le spectre des caractéristiques linguistiques du discours des patients, en tant que reflet des pathologies qu'ils présentent. Les techniques d'analyse informatisée du langage telles que l'analyse sémantique latente et l'analyse structurelle du discours indiquent une diminution de la cohérence chez les patients atteints de schizophrénie en corrélation avec les évaluations cliniques et une précision identique ou supérieure dans l'évaluation diagnostique (Hoffman et al., 1986; Elvevåg et al., 2007; Mota et al., 2012). Une combinaison d'analyses sémantiques et syntaxiques peut prédire avec une précision raisonnable la transition vers la schizophrénie et semble être plus efficace que l'évaluation clinique utilisant des outils standardisés (Bedi et al., 2015; Corcoran et al., 2018).

1. https://www.inserm.fr/information-en-sante/dossiers-information/schizophrenie, consulté le 23/03/2021

La psychose est accompagnée de comorbidités, en particulier les troubles de l'humeur, l'anxiété et la dépendance (Bazziconi et al., 2017). Il est donc important de les identifier afin de proposer des soins adaptés. Les analyses prosodiques des comorbidités psychiatriques se sont principalement concentrées sur la fréquence fondamentale (F0) et le débit de parole (Scherer et Bänziger, 2004; Audibert et al., 2005; van den Broek, 2004; Moore et al., 2003). Silber-Varod et al. (2016) considèrent les pauses et les disfluences dans les comorbidités anxieuses en étudiant principalement les caractéristiques prosodiques. Notre étude considère les mêmes facteurs en se concentrant sur la syntaxe.

Nous présentons, section 2, les dépendances syntaxiques, au cœur de notre approche, et introduisons un nouveau concept, le croisement interstitiel de dépendances. Dans la section 3 nous décrivons la façon dont les comorbidités sont évaluées, la constitution du corpus, notre méthodologie et nos résultats. Une conclusion et des perspectives sont proposées section 4.

2 Grammaires de dépendances et croisements interstitiels

À la fin des années trente, le linguiste français Lucien Tesnière a commencé à travailler sur une nouvelle théorie syntaxique fondée sur les relations entre les mots, théorie qui n'a été publiée qu'à titre posthume (Tesnière, 1959). Ses travaux resteraient méconnus au niveau international si un chercheur de la Rand Corporation, David Hays, n'avait pas présenté les idées de Tesnière à la communauté encore jeune des «linguistes computationnels» par une présentation au célèbre symposium de l'UCLA sur la traduction automatique (Hays, 1960), suivie d'un article dans la revue *Language* (Hays, 1964) et, enfin, d'un livre qui se trouve être *le premier livre consacré à la linguistique computationnelle* (Hays, 1967). C'est Hayes qui a introduit les termes de *grammaire de dépendances* et de *relation de dépendance*. Par la suite, l'utilisation des grammaires de dépendances a continué de se répandre et elles semblent avoir aujourd'hui supplanté les méthodes basées sur les constituants (Osborne, 2019). Les grammaires de dépendances ont déjà été utilisées dans le domaine psychiatrique, par exemple dans Tanana et al. (2016) où les séances d'entretiens de motivation ont été codées par ordinateur.

Dans une grammaire de dépendances, chaque phrase a une *tête* (généralement le verbe) qui est la racine d'un arbre orienté de *relations de dépendance*. Les arêtes sont orientées de manière que l'on puisse tracer des chemins (orientés) de chaque feuille à la racine. Chaque arête possède une étiquette, appelée *nature de dépendance*, qui décrit la relation entre la *dépendance* (source de l'arête) et le *gouverneur* (cible de l'arête).

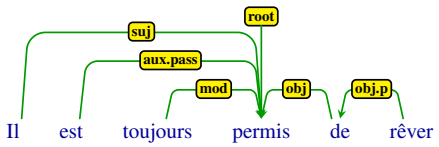


FIG. 1 – *Arbre de dépendance de «Il est toujours permis de rêver», tiré du* French Treebank Corpus *(Abeillé et al., 2003)*

Nous remarquons dans l'arbre de la figure 1 que le participe «permis» est la tête de l'arbre, et qu'il gouverne : le pronom «il» en tant que sujet (suj) ; le verbe «est» en tant qu'auxiliaire

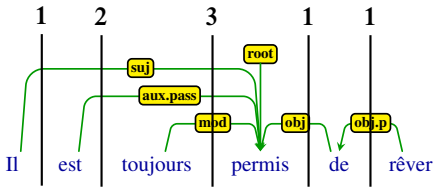


FIG. 2 – *Croisements interstitiels de dépendances*

(aux.pass) ; l'adverbe «toujours» en tant que modificateur (mod) ; et la préposition «de» en tant qu'objet (obj). De plus, nous remarquons que «rêver» est gouverné par «de» à travers une relation de dépendance de complément d'objet indirect prépositionnel (obj.p).

Liu (2008) explore les dépendances d'un point de vue cognitif et définit une mesure de complexité du langage, la *distance de dépendance moyenne* (DDM), qui quantifie le fait qu'une phrase anglaise telle que «The man the boy the woman saw heard left», bien que grammaticale, est bien plus difficile à comprendre que la phrase sémantiquement équivalente «The woman saw the boy that heard the man that left» (leurs valeurs DDM sont resp. de 3 et de 1,4). Si l'on définit DDM comme la distance moyenne entre gouverneur et gouverné, plus les dépendances sont «à longue distance», plus le DDM est élevé.

Par ailleurs, les dépendances ne se chevauchent pas, de sorte que nous avons une relation binaire irréflexive, asymétrique et transitive $\prec$ entre elles : $(a \rightarrow b) \prec (c \rightarrow d)$ lorsque $(\mathrm{pos}(a) < \mathrm{pos}(c)$ et $\mathrm{pos}(d) \geq \mathrm{pos}(b))$ ou $(\mathrm{pos}(a) \leq \mathrm{pos}(c)$ et $\mathrm{pos}(d) > \mathrm{pos}(b))$, où pos représente l'ordre linéaire des mots de la phrase.

Dans l'exemple de la figure 2, nous avons (toujours $\rightarrow$ permis) $\prec$ (est $\rightarrow$ permis) $\prec$ (Il $\rightarrow$ permis). La relation $x \prec y$ implique également que $\mathrm{longueur}(x) < \mathrm{longueur}(y)$, et est un ordre partiel de sorte que nous pouvons construire un treillis dont les nœuds sont des dépendances et les arêtes représentent $\prec$. Les chemins dans ce treillis peuvent être visualisés dans l'arbre de dépendance en traçant des lignes verticales entre les mots. Le fait que (toujours $\rightarrow$ permis) $\prec$ (est $\rightarrow$ permis), qui est un chemin de longueur 2 dans le treillis, est représenté par le fait que la deuxième ligne verticale traverse deux dépendances. De même, le chemin (toujours $\rightarrow$ permis) $\prec$ (est $\rightarrow$ permis) $\prec$ (Il $\rightarrow$ permis), qui est d'ordre 3, est représenté par le fait que la troisième ligne verticale croise trois dépendances. Comme nous le voyons, le nombre de croisements augmente lorsque nous nous approchons de la racine par la gauche puisque de nombreuses dépendances ciblant la racine s'accumulent, tandis qu'à droite, en raison de l'adjacence entre les nœuds, le nombre de croisements reste faible.

Outre les mots, notre corpus contient également des *éléments paralinguistiques*, tels que les *interjections* et les *pauses* qui, par définition, ont lieu dans des positions interstitielles. Pour les traiter de manière appropriée, nous avons introduit (Haralambous et al. (2020)) une nouvelle notion : le ***croisement interstitiel de dépendances***. Notre hypothèse est que les positions interstitielles ayant une valeur de croisement élevée sont stratégiques et que le fait d'y placer des «intrus» (interjections, pauses) peut être indicateur de trouble. Nous verrons qu'en combinant le nombre et la nature des dépendances croisées sur une pause ou une interjection nous obtenons des indicateurs de comorbidité.

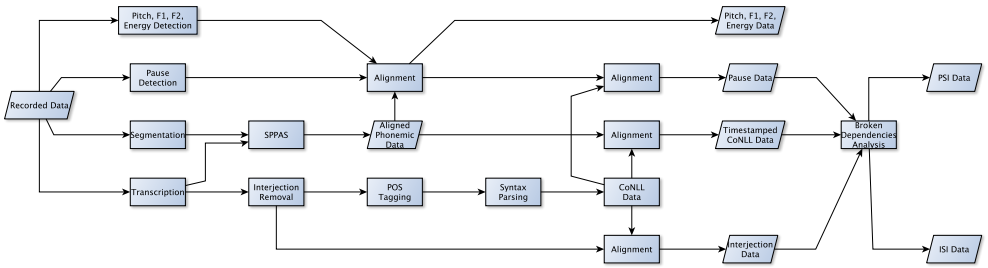


FIG. 3 – *Le processus d'extraction de données*

3 Vers une aide au diagnostic de comorbidités chez des patients courant le risque de développer une psychose

Par la suite nous allons décrire notre processus de traitement de données extraites d'entretiens psychiatriques visant à détecter, de manière précoce, un patient présentant un trouble psychotique (en général) et notamment la schizophrénie (en particulier). Notre corpus est relativement petit (des *small data*), comme c'est souvent le cas pour les données médicales – dans notre cas, une patienthèque, avec plusieurs centres de détection précoce, est en cours de constitution et plusieurs centaines d'entretiens seront disponibles d'ici deux ans. Nous avons fait une recherche d'indicateurs linguistiques et paralinguistiques à large spectre, qui nous a permis de conclure que certaines propriétés syntaxiques des pauses et interjections peuvent être corrélées avec certains groupes de comorbidités. Si nous nous limitons, pour le moment, aux comorbidités, c'est parce que ce n'est qu'après deux années de suivi, lorsque l'état clinique des interviewés aura évolué, que nous pourrons véritablement évaluer des prédictions de transition vers la psychose. En effet, ces patients bénéficient d'un suivi rapproché pendant deux ans afin de surveiller leur évolution clinique et de leur apporter les soins appropriés en cas d'aggravation de leur état.

3.1 Évaluation du risque de psychose et constitution de corpus

Les patients reçus au sein de la consultation de détection et d'intervention précoce du CHU de Brest (programme CEVUP = consultation d'évaluation de la vulnérabilité psychologique, Bazziconi et al. 2017) sont évalués par une équipe pluridisciplinaire comprenant un psychiatre, un psychologue, un infirmier et un neuropsychologue. L'évaluation initiale permet d'identifier leur niveau de risque et d'établir un protocole de soins personnalisé. Une réévaluation semestrielle est proposée pendant deux ans afin d'identifier les aggravations potentielles des troubles et l'apparition éventuelle d'une psychose. Cette transition du statut de patient «à risque de développer un trouble psychotique» à l'apparition d'une pathologie psychotique confirmée est appelée «transition vers la psychose» (24% des patients à risque développent un trouble psychotique dans les deux années qui suivent et 33% dans les trois années suivantes, cf. *loc. cit.*). Les résultats présentés dans cet article s'inscrivent dans le cadre d'un projet de recherche sur l'analyse informelle de la parole impliquant tous les patients orientés vers le centre de détection et d'intervention précoce (protocole de recherche NCT03525054 validé par le Comité de Pro-

ID	Genre	Durée	A_1	A_2	B	C	D_1	D_2	E	F	G	H	I	J	K_1	K_2	K_3	K_4	L	M	N	O	P	**THY**	**ANX**	**ADD**
15	F	47′29″	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
21	M	47′45″	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	3	0
23	F	43′50″	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
25	M	30′59″	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
27	M	25′05″	1	0	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	3	2	1
28	M	27′26″	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2	0
30	M	63′08″	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	1	1
44	M	43′13″	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	1	2

TAB. 1 – *Valeurs de comorbidités de notre corpus de patients*

tection des Personnes Est-III –N CPP : 18.04.03–). Il prévoit un enregistrement de l'entretien médical clinique initial et un suivi de deux ans.

Les comorbidités suivantes, représentées par les lettres A à P (variables oui (1) / non (0) sauf pour la suicidalité, qui est graduée selon trois intensités 0/1/2), sont évaluées lors de l'entretien selon le standard *Mini International Neuropsychiatric Interview* (Sheehan et al., 1998) :

A. Trouble dépressif majeur (A_1 : trouble dépressif majeur sans désordres psychotiques ; A_2 : trouble dépressif majeur avec caractéristiques psychotiques)
B. Dysthymie
C. Suicidalité
D. Épisode (hypo)maniaque (D_1 : hypomaniaque ; D_2 : maniaque)
E. Trouble de panique
F. Agoraphobie
G. Phobie sociale
H. Trouble obsessionnel compulsif
I. Syndrome de stress post-traumatique
J. Dépendance/abus d'alcool
K. Dépendance aux drogues (K_1 : opioïdes ; K_2 : cocaïne ; K_3 : cannabis ; K_4 : sédatifs)
L. Troubles psychotiques
M. *Anorexia Nervosa*
N. *Bulimia Nervosa*
O. Trouble d'anxiété généralisée
P. Trouble de la personnalité antisociale

Nous avons regroupé ces comorbidités en trois groupes selon la nature des troubles, afin de permettre des analyses statistiques sur un nombre limité de patients : troubles thymiques (**THY** : A, B, C, D) ; troubles anxieux (**ANX** : E, F, G, H, I, O) ; et troubles de dépendance/addiction (**ADD** : J, K, M, N). Les comorbidités L et P, ne concernant presque aucun patient de notre corpus, ont été omises par notre étude.

Les enregistrements sont retranscrits par un personnel médical, en respectant les conventions d'interjections et de respiration paralinguistique établies par Bigi (2015). Ces transcriptions sont ensuite relues et éditées par un correcteur indépendant (pour une deuxième vérification, garantissant, entre autres, l'anonymat au sein des retranscriptions). Le processus de constitution du corpus nécessite donc des ressources conséquentes (personnels, temps). Par ailleurs, la schizophrénie touche une très petite part de la population (environ 1%). De ce fait nous ne disposons, pour le moment, que de corpus de petite taille. Le corpus utilisé dans cette étude se compose uniquement de huit entretiens – nous verrons néanmoins qu'il permet d'exhiber des indicateurs concluants. Le tableau 1 présente les principales caractéristiques du corpus «brut» (genre des patients ; durée de l'entretien ; valeurs des comorbidités).

Chaque entretien enregistré est segmenté en répliques entre le soignant et le patient. Une fois la transcription soigneusement vérifiée, les deux flux de données (son et texte) sont fournis

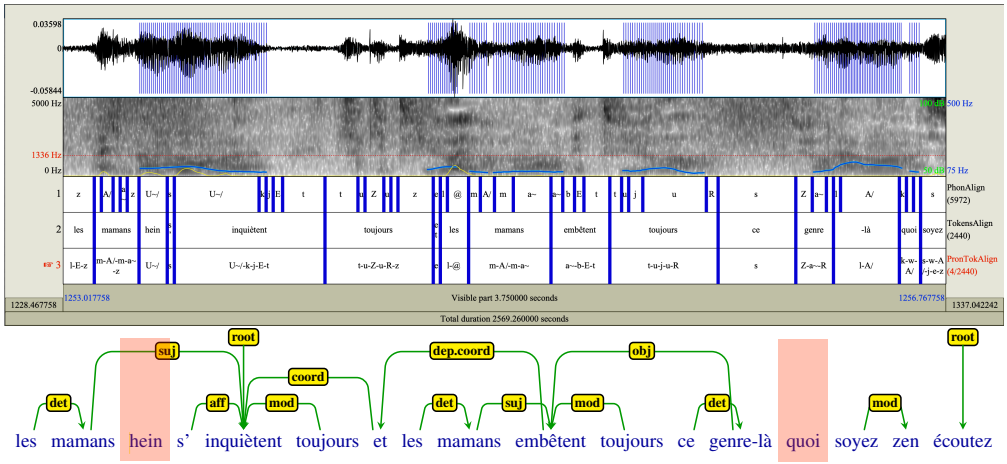


FIG. 4 – *Un énoncé (patient #44) visualisé dans PRAAT (alignement phonémique) et annoté par des dépendances syntaxiques. L'interjection primaire «hein» croise la dépendance de type sujet «mamans» → «inquiètent». L'interjection secondaire apposée «quoi» ne croise aucune relation de dépendance.*

au logiciel SPPAS (Bigi, 2015), qui produit un fichier comportant des phonèmes horodatés, des mots en orthographe standard et des mots en représentation phonémique. Cependant, SPPAS ne détecte pas les pauses. Nous utilisons donc PRAAT (Boersma et Weenink, 2001) sur une version préparée du fichier sonore pour détecter les pauses et les écarts. Puis nous les injectons dans les données de phonèmes et de mots horodatés. PRAAT fournit également des données d'énergie, de hauteur, de F1 et de F2, que nous alignons avec les phonèmes et les mots.

Dans un flux de données parallèles (voir fig. 3), nous supprimons les interjections du texte transcrit et effectuons un marquage POS sur le résultat avec Talismane (Urieli, 2013; Urieli et Tanguy, 2013), suivi d'une analyse syntaxique des dépendances effectuée par Grew (Guillaume et Perrier, 2015). Ce processus nous fournit des données CoNLL relativement propres. Nous alignons ensuite les deux flux de données (données fournies par SPPAS et données sous format CoNLL) en utilisant l'algorithme de Needleman-Wunsch tel qu'implémenté dans bioPython. Nous obtenons ainsi des données CoNLL horodatées. L'horodatage des pauses et des interjections nous sert à étudier leurs croisements avec les dépendances syntaxiques. Il est prévu que les résultats des analyses soient confrontés aux données cliniques recueillies pour chacun des patients et validés par un psychiatre clinicien de l'équipe de détection et d'intervention précoce.

3.2 Définition et justification du PIDC et du IIDC

Considérons la forêt syntaxique de dépendances[2] d'un énoncé donné. Comme on peut le voir, figure 4, dans *«les mamans s'inquiètent toujours et les mamans embêtent toujours*

2. Nous utilisons le terme de *forêt* en raison de la co-présence de plusieurs arbres syntaxiques dans le même énoncé.

nature	fréquence	DDM dép./gouv.	nature	fréquence	DDM dép./gouv.
mod	120,741	4,1937	det	85,154	1,1987
obj.p	90,400	1,7511	suj	35,402	4,2315

TAB. 2 – *Les quatre relations les plus fréquentes dans le corpus French Treebank*

ce genre-là quoi soyez zen écoutez», les mots «quoi» et «écoutez» ne sont pas connectés à l'arbre syntaxique des deux phrases coordonnées «les mamans s'inquiètent toujours» et «les mamans embêtent toujours ce genre-là» («quoi» et potentiellement «ce genre-là» pourraient aussi être considérés comme des interjections secondaires). Nous ne disposons donc pas d'un arbre syntaxique unique mais de fragments d'arbre de taille variable.

Nous supprimons toutes les interjections primaires afin d'obtenir des dépendances plus proches de l'intention du locuteur et d'éviter une mauvaise interprétation par l'analyseur syntaxique qui a été entraîné sur un corpus sans interjections. Nous introduisons la mesure IIDC (croisements d'interjections), dont le but est de quantifier le croisement des interjections, en tant qu'interstices entre les mots, avec les relations de dépendance qui relient les mots. Comme le lecteur peut le voir dans la figure 4, l'interjection primaire «hein» croise une relation de dépendance entre le nom «mamans» agissant comme sujet, et le verbe «s'inquiètent», qui est la racine du fragment d'arbre. Une autre interjection (secondaire, cette fois), «quoi», ne croise aucune relation de dépendance puisqu'elle est située entre des arbres syntaxiques distincts dans la forêt. Nous faisons de même pour les pauses : le PIDC (croisements de pauses) est une mesure du croisement des pauses (c'est-à-dire des silences internes aux répliques de chaque patient) en tant qu'interstices entres les mots avec les relations de dépendance.

Notre hypothèse est la suivante : les croisements d'interjections et les croisements de pauses peuvent servir d'indicateurs de la désorganisation linguistique du patient. Nous nous intéressons donc (1) au nombre de dépendances traversant des interjections ou des pauses et (2) aux étiquettes des relations de dépendance croisées. Nous définissons ainsi quatre mesures :

$$\text{PIDC} = (\#c)\times\frac{\text{durée de pause}}{\text{durée de l'énoncé}}, \qquad \text{PIDC}_S = (\#c \text{ dans S})\times\frac{\text{durée de pause}}{\text{durée de l'énoncé}},$$
$$\text{IIDC} = (\#c)\times\frac{\text{durée de l'interjection}}{\text{durée de l'énoncé}}, \qquad \text{IIDC}_S = (\#c \text{ dans S})\times\frac{\text{durée de l'interjection}}{\text{durée de l'énoncé}},$$

où #c est le nombre de croisements, et S est un ensemble de relations de dépendance.

Nous calculons par la suite les valeurs de PIDC_S et d'IIDC_S pour quatre ensembles spécifiques de relations de dépendances : {det,suj}, {det}, {obj.p} et {suj}. Ces dernières sont les quatre relations les plus fréquentes dans le corpus French Treebank (Abeillé et al., 2003) (voir tableau 2). Malgré sa fréquence élevée, nous n'avons pas sélectionné la dépendance «mod» (modificateur) pour la raison suivante : les mots gouvernés (dans 26% des cas, un adjectif ; dans 22% des cas, une préposition ; dans 20% des cas, un adverbe ; dans 18% des cas un nom) peuvent être assez éloignés de leur gouverneur et donc l'existence d'une pause ou d'une interjection entre gouverné et gouverneur n'est pas nécessairement significative. Au contraire, la dépendance «obj.p» (objet prépositionnel) est en fait l'équivalent d'un *cas de gouvernance* (pour les langues à cas) et donc, selon Osborne (2019, p. 142), elle serait techniquement plutôt morphologique que syntaxique. Elle est très stable en termes de partie de discours (86% de ses gouvernés sont des noms) et la distance entre gouverneur et gouverné est assez faible (1,7511 en moyenne). Sa nature morphologique et ses caractéristiques positionnelles nous amènent à

	pauses			interjections		
dépendances	comorbidités	ρ	p-valeur	comorbidités	ρ	p-valeur
{obj.p}	**ADD**	**0,8660**	0,0054	**ADD**	0,8248	0,0117
{det}	**ADD**	0,7735	0,0254	**ADD**	0,7285	0,04
{det,suj}	**ADD**	0,5770	0,1340	**ANX**	−0,8247	0,0117
{suj}	**ANX**	−0,5086	0,1980	**ANX**	**−0,8450**	0,0080
toutes	**THY**	0,7042	0,0512	**THY**	−0,6730	0,0671

TAB. 3 – *Corrélation de Spearman sur les comorbidités des trois groupes **THY**, **ANX** et **ADD***

formuler l'hypothèse que le croisement d'une interjection ou d'une pause avec une dépendance obj.p est susceptible de révéler une désorganisation mentale.

La dépendance «det» (déterminant) est également candidate à révéler une désorganisation : la liste des déterminants est très réduite et ils sont très proches de leur gouverneur (1,1987 en moyenne, plus petite distance moyenne des relations). Enfin, la relation «suj» est importante, malgré sa distance élevée entre gouverneur et gouverné (4,2315 en moyenne), puisque (sauf à l'impératif) les verbes français possèdent nécessairement des sujets.

3.3 Résultats et analyse

Nous avons effectué un test de corrélation de Spearman sur les valeurs de comorbidité des trois groupes **THY**, **ANX** et **ADD** par rapport aux croisements de pauses/interjections calculés. Le tableau 3 présente les résultats les plus pertinents avec leur p-valeur.

Pour {obj.p} et {det} nous obtenons un comportement similaire pour les pauses et les interjections, même si ces deux phénomènes paralinguistiques sont bien distincts (et mesurés de manière différente, cf. fig. 3). Nous remarquons également que les pauses ou interjections traversant la dépendance {obj.p} constituent un indicateur très fort ($\rho > 0{,}82$) du groupe **ADD**, avec une significativité élevée ($p = 0{,}012$). La dépendance {det} a également un comportement cohérent ($\rho \approx 0{,}75$, avec $0{,}025 \leq p \leq 0{,}04$) et cible, de nouveau, le groupe **ADD**. Les valeurs des autres dépendances révèlent des comportements différents : alors que les pauses croisant {det,suj} ou {suj} donnent des résultats non significatifs ($p > 0{,}13$), les interjections croisant {det,suj} et {suj} donnent des résultats très élevés, mais ciblent négativement le groupe **ANX** ($\rho < 0{,}824$ avec $p = 0{,}012$). Ces résultats peuvent être résumés comme suit :

— les membres du groupe **ADD** ont tendance à placer des pauses ou des interjections entre la préposition et le nom gouverné ou entre le déterminant et le nom qui le gouverne ;

— les membres du groupe **ANX** ont tendance à placer des interjections entre le déterminant et le nom qui le gouverne, ou entre le sujet et le verbe qui le gouverne.

Le premier résultat peut refléter la forte prévalence des comportements addictifs chez les patients à risque de psychose (Valmaggia et al., 2014). Il montre que le croisement d'une interjection ou d'une pause entre préposition et nom ou entre déterminant et nom est susceptible de révéler une désorganisation mentale qui est un des symptômes psychotiques souvent retrouvés chez les patients à risque, elle est caractéristique de la schizophrénie (Fusar-Poli et al., 2013). Le second résultat peut s'expliquer par une tendance des patients anxieux à éviter de laisser des blancs, notamment dans le cadre d'une conversation où l'individu est soumis au jugement de son interlocuteur, de manière similaire aux patients bègues (Iverach et Rapee, 2014).

4 Conclusion

Ces résultats montrent qu'il est possible d'utiliser le traitement du langage naturel combiné avec des données paralinguistiques pour explorer les comorbidités psychiatriques. Les dépendances et leurs croisements avec les pauses et les interjections semblent être particulièrement indiquées à cette fin. Nous comptons poursuivre l'exploration de marqueurs linguistiques et paralinguistiques afin d'identifier des marqueurs pertinents pour la pratique clinique.

Références

Abeillé, A., L. Clément, and F. Toussenel (2003). Building a treebank for French. In *Treebanks*, pp. 165–187. Kluwer.

Anderson, K. (2019). Towards a public health approach to psychotic disorders. *Lancet Public Health 4*(5), e212-3.

Audibert, N., V. Aubergé, and A. Rilliard (2005). The prosodic dimensions of emotion in speech: the relative weights of parameters. In *European Conference on Speech Communication and Technology*, pp. 525–528. ISCA.

Bazziconi, P., C. Lemey, L. Bleton, and M. Walter (2017). CEVUP program: An analytical epidemiological cohort study. *European Psychiatry 41*, S729.

Bedi, G., F. Carrillo, G. Cecchi, D. Fernández-Slezak, M. Sigman, N. Mota, S. Ribeiro, D. Javitt, M. Copelli, and C. Corcoran (2015). Automated analysis of free speech predicts psychosis onset in high–risk youths. *npj Schizophrenia 1*, 15030.

Bigi, B. (2015). SPPAS – Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician – International Society of Phonetic Sciences 111–112*, 54–69.

Boersma, P. and D. Weenink (2001). PRAAT, a system for doing phonetics by computer. *Glot International 5*(9-10), 341–347.

Corcoran, C., F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. Javitt, C. Bearden, and G. Cecchi (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry 17*(1), 67–75.

Elvevåg, B., P. W. Foltz, D. R. Weinberger, and T. E. Goldberg (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr. Res 93*, 304–316.

Fusar-Poli, P. et al. (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry 70*, 107–120.

Guillaume, B. and G. Perrier (2015). Dependency parsing with graph rewriting. In *International Conference on Parsing Technologies*, pp. 30–39.

Haralambous, Y., C. Lemey, P. Lenca, R. Billot, and D.-H. Kim-Dufor (2020). Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities. In *Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments*, pp. 142–150.

Hays, D. (1960). Grouping and dependency theories. In *Proceedings of the National Symposium on Machine Translation*, pp. 257–266. UCLA.

Hays, D. (1964). Dependency theory: A formalism and some observations. *Language 40*, 159–525.

Hays, D. (1967). *Introduction to computational linguistics*. Macdonald & co.

Hoffman, R. E., S. Stopek, and N. C. Andreasen (1986). A comparative study of manic vs schizophrenic speech disorganization. *Arch. Gen. Psychiatry 43*, 831–838.

Howes, O. D. and R. McCutcheon (2017). Inflammation and the neural diathesis-stress hypothesis of schizophrenia: A reconceptualization. *Transl. Psychiatry 7*, 1024.

Iverach, L. and R. M. Rapee (2014). Social anxiety disorder and stuttering: current status and future directions. *J. Fluency Disord. 40*, 69–82.

Le Glaz, A., Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, R. Taylor, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, and C. Lemey (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research 23*(5), e15708.

Lim, J., G. Rekhi, A. Rapisarda, M. Lam, M. Kraus, R. Keefe, et al. (2015). Impact of psychiatric comorbidity in individuals at ultra high risk of psychosis - findings from the longitudinal youth at risk study (lyriks). *Schizophr. Res. 164*, 1–3.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science 9*, 159–191.

McGlashan, T. H. and J. O. Johannessen (1996.). Early detection and intervention with schizophrenia: rationale. *Schizophr. Bull. 22*, 201–222.

Moore, E., M. Clements, J. Peifer, and L. Weisser (2003). Analysis of prosodic variation in speech for clinical depression. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, Volume 3, pp. 2925–2928.

Mota, N. B., N. A. P. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE 7*(4), e34928.

Olsen, K. A. and B. Rosenbaum (2006). Prospective investigations of the prodromal state of schizophrenia: assessment instruments. *Acta Psychiatr. Scand. 113*, 273–282.

Osborne, T. (2019). *A Dependency Grammar of English*. John Benjamins.

Pruessner, M., A. E. Cullen, M. Aas, and E. F. Walker (2017). The neural diathesis–stress model of schizophrenia revisited: An update on recent findings considering illness stage and neurobio. and methodol. complexities. *Neurosci. Biobehav. Rev. 73*, 191–218.

Roca, M., M. Gili, M. Garcia-Garcia, J. Salva, M. Vives, J. Garcia Campayo, and A. Comas (2009). Prevalence and comorbidity of common mental disorders in primary care. *J. Affect. Disord. 119*(1–3), 52–58.

Rössler, W., H. Joachim Salize, J. van Os, and A. Riecher-Rössler (2005). Size of burden of schizophrenia and psychotic disorders. *European Neuropsychopharmacology 15*, 399–409.

Scherer, K. R. and T. Bänziger (2004). Emotional expression in prosody: a review and an agenda for future research. In *The Speech Prosody Conference*.

Sheehan, D. V., Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, and G. C. Dunbar (1998). The Mini–International Neuropsychiatric Interview

(M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM–IV and ICD–10. *J. Clin. Psychiatry 59 Suppl 20*, 22–33.

Silber-Varod, V., H. Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir (2016). Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. In *Proceedings of the Speech Prosody Conference*, pp. 1211–1215.

Tanana, M., K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *J Subst Abuse Treat. 65*, 43–50.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph. D. thesis, Université de Toulouse II le Mirail.

Urieli, A. and L. Tanguy (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions: études de cas avec l'analyseur Talismane. In *Actes de la 20^e conférence sur le Traitement Automatique des Langues Naturelles*, pp. 188–201.

Valmaggia, L. R., F. L. Day, C. Jones, S. Bissoli, C. Pugh, D. Hall, S. Bhattacharyya, O. Howes, J. Stone, P. Fusar-Poli, M. Byrne, and P. McGuire (2014). Cannabis use and transition to psychosis in people at ultra–high risk. *Psychological Medicine 44*(12), 2503–2512.

van den Broek, E. L. (2004). Emotional prosody measurement (EPM): a voice–based evaluation method for psychological therapy effectiveness. *Stud. Health Technol. Inform. 103*, 118–125.

Yung, A. R., H. P. Yuen, P. D. McGorry, L. J. Phillips, D. Kelly, M. Dell'olio, S. M. Francey, E. M. Cosgrave, E. Killackey, C. Stanford, K. Godfrey, and J. Buckby (2005). Mapping the onset of psychosis: The comprehensive assessment of at–risk mental states. *Aust N Z J Psychiatry 39*(11-12), 964–971.

Summary

Co-morbidities are very frequent in mental health and represent a major therapeutic issue as well as a lever for a better understanding of physiopathological mechanisms. Some diseases, such as schizophrenia, develop progressively and at different rates from one individual to another, with symptoms gradually increasing in intensity and specificity. In these cases, clinicians seek to identify early on the warning symptoms and aggravating comorbidities, in order to propose interventions that maximize the therapeutic effects. Speech, and thus language, is a key element they use during consultations to understand the psychological state of patients, and automatic language analysis systems can provide an aid to assessment. We propose such an aid, targeting the detection of comorbidities and based on dependency grammars and paralinguistic indicators such as pauses and interjections, which are shown to be relevant.

Keywords: psychiatric comorbidities, natural language processing, dependency grammars, schizophrenia, paralinguistics

Prédiction des maladies chroniques: cas de l'insuffisance rénale

Basma Boukenze

Basma.boukenze@gmail.com
Computer, Networks, Mobility and Modelling laboratory FST,
Hassan 1st University, Settat, Morocco

Résumé. L'insuffisance rénale est une maladie chronique silencieuse, sans symptômes, et qui évolue rapidement vers des phases plus compliquées. Malheureusement, la majorité des patients découvrent cette maladie à des stades avancés où l'hémodialyse ou la greffe devient une nécessité. Prédire l'insuffisance rénale permettra de contrôler la progression de la maladie et pourra même stabiliser l'état des patients. Les techniques d'analyse et de fouille de données, ainsi que l'apprentissage automatique ont montré un grand succès dans la découverte de l'information et l'extraction des connaissances afin d'aider à la bonne prise de décision. En effet, le Big data médical a rendu l'application de ces technologies primordiale, compte tenu l'utilité des connaissances extraites à partir des données patients. Ces connaissances contribueront à améliorer la prise de décision et à sauver des vies humaines. Dans cet article, nous présentons l'importance de l'exploration des données médicales dans la prédiction des maladies chroniques. Nous montrons également l'importance des outils et techniques de fouille et d'apprentissage automatique dans les applications au secteur médical à travers la prédiction de la progression de la maladie d'insuffisance rénale à partir d'une base de données médicales.

Mots-clés : Analyse et fouille de données, Apprentissage automatique, Prédire, Big data Médicale, Secteur Médical, Prise de décision, insuffisance rénale

1. Introduction

Le nombre de patients atteints d'insuffisance rénale terminale augmente dans le monde entier, avec une augmentation des cas de 5 à 7% chaque année; l'insuffisance rénale devient l'une des priorités de la loi sur la santé publique [1]. Bien qu'essentiels à notre vie, les reins sont des organes assez particuliers du corps humain: ils filtrent chaque jour 180 litres de notre sang et éliminent les déchets du corps [2]. Mais un danger silencieux pouvait troubler cette harmonie vitale; c'est l'insuffisance rénale. Cette diminution progressive du fonctionnement des reins, dépourvue de symptômes, correspond à une destruction progressive et irrémédiable des canaux (les néphrons) qui constituent le rein. Il apparaît lorsque seul un tiers de ces canaux reste en état de fonctionnement. Elle peut être détectée par un bilan biologique appelé dosage de la créatinine en cas de diabète, d'infection des voies urinaires, d'hypertension artérielle, de calculs rénaux, d'albumine dans les urines, d'infection sévère, d'ané-

mie inexpliquée ou d'intoxication médicamenteuse [3]. Ces manifestations peuvent être la cause ou la conséquence d'une insuffisance rénale. Au stade final (90% des néphrons inefficaces), cette maladie nécessite un traitement urgent comme la dialyse ou la greffe.

Les personnes atteintes d'insuffisance rénale chronique (IRC) sont plus susceptibles de mourir d'une maladie cardiovasculaire (CV). Ce résultat est prouvé par une vaste étude de cohorte comprenant plus de 130 000 sujets âgés et qui a montré que l'augmentation de l'incidence des événements CV pourrait être en partie liée au fait que les personnes souffrant d'insuffisance rénale sont moins susceptibles de recevoir des traitements cardio-protecteurs appropriés. Il est clair que la MCV accélérée est répandue chez les sujets atteints d'IRC [4].

Dans le diagnostic, l'insuffisance rénale chronique est définie comme une atteinte rénale ou un débit de filtration glomérulaire inférieur à 60 ml / min pendant trois mois ou plus [5]. Il s'agit invariablement d'un processus progressif qui aboutit à une maladie rénale terminale. La créatinine sérique est couramment utilisée pour estimer la clairance de la créatinine, mais elle est un mauvais prédicteur du taux de filtration glomérulaire, car elle peut être influencée de manière imprévisible par les techniques de dosage, les substances endogènes et exogènes, la manipulation tubulaire rénale de la créatinine et d'autres facteurs (âge, sexe, poids corporel, masse musculaire, alimentation, médicaments). Le taux de filtration glomérulaire est

«L'étalon-or» pour déterminer la fonction rénale, mais sa mesure reste lourde. À des fins pratiques, la clairance de la créatinine calculée est utilisée comme corrélat du taux de filtration glomérulaire et est généralement estimée en utilisant la formule de Cockcroft-Gault [6] [7] comme suit:

$$Cer = \frac{(140-\text{age})(\text{wt Kg})}{72*\text{Ser}\,(\frac{\text{mg}}{100ml})} \qquad (1)$$

Même si l'insuffisance rénale est une maladie silencieuse, elle ne l'empêche pas de générer un ensemble de symptômes et de signes qui nécessitent des dosages par analyse. Parfois, les médecins sur la base de cette analyse commettent un faux diagnostic qui entre dans le cas d'erreurs médicales commises par les médecins lors de l'examen pour déterminer la nature des symptômes présentés par le patient. De telles erreurs surviennent lorsqu'un médecin s'abstient de rechercher des informations sur l'état de santé du patient, ne sollicite pas l'avis d'un confrère plus spécialisé lorsqu'un tel avis est requis ou ne prescrit pas un examen pourtant nécessaire à l'établissement d'un diagnostic. Une étude a montré que les erreurs médicales seraient la troisième cause de décès aux États-Unis, après les maladies cardiovasculaires et le cancer. C'est la conclusion à laquelle sont parvenus deux médecins américains, Martin Makary et Michael Daniel, tous deux du département de chirurgie de l'Université Johns Hopkins de Baltimore, Maryland, qui ont conclu que les erreurs médicales causent

251 000 décès par an [8].

L'insuffisance rénale chronique est une maladie qui n'a pas encore de médicament efficace. Mais une détection précoce suivie d'un bon diagnostic conduira à une meilleure prise de décision et joue un rôle important dans l'identification précoce, la prévention des patients, l'anticipation du traitement, l'arrêt de la progression de la maladie, la stabilisation de la situation du patient et par conséquent l'amélioration des résultats [9].

Les technologies de méga-données connues par le Big Data, les techniques d'exploration de données et l'apprentissage automatique ont connu un grand succès dans la découverte et l'extraction des connaissances, d'autant plus que le domaine médical est un domaine de production de données.

Le potentiel de l'analyse Big Data permet de ralentir les coûts de soins toujours croissants, d'aider les prestataires à pratiquer une médecine plus efficace, de responsabiliser les patients et les prestataires de soins de santé et de rêver d'une médecine plus personnalisée et prédictive. L'utilisation des médias sociaux, du cloud computing, associée à des procédures intelligentes de gestion, d'analyse et d'extraction d'informations à partir des données, transformera le système de santé et donnera le pouvoir d'explorer, de prédire et pourquoi pas d'anticiper la guérison. L'analyse des Big data médical promet et affirme que les dimensions de la médecine changeront complètement.

Dans cet article, nous explorerons la prédiction de la maladie d'insuffisance rénale; un résultat obtenue par l'application de différentes méthodes de fouille de données sur des bases de données patients. Notre objectif est, d'une part de montrer le grand intérêt de ce sujet dans la communauté scientifique et les efforts fournis dans ce sens, et d'autre part le plus grand besoin d'appliquer ces techniques dans le domaine de la néphrologie en raison des promesses et des résultats obtenus.

2. Revue de littérature

Les auteurs de [10] ont développé un modèle pour prédire le risque de progression de la maladie rénale chronique (IRC) de la phase III à la phase IV, qui comprend des données longitudinales et des caractéristiques cliniques.

La base de données d'étude a été dérivée de l'entrepôt de données cliniques, NewYork-Presbyterian Hospital (NYP). Cette base de données était composée de 2908 patients en soins primaires qui ont eu au moins trois visites avant le 1er janvier 2013 et ont développé une maladie de IRC de phase III au cours de leurs antécédents documentés. Les groupes de test et de validation ont été sélectionnées au hasard dans la base de données de l'étude.

L'ensemble de données de l'étude comprenaient des données longitudinales pour les populations hospitalisées et non-hospitalisées. La cohorte de l'étude a été divisée au hasard en cohortes de test (90%) et cohortes de validation (10%).

Les auteurs ont utilisé deux types de variables: des variables indépendantes comme l'âge et le sexe, et des variables dépendant du temps, y compris tous les paramètres vitaux, comme la pression artérielle, la créatinine, le magnésium, les protéines, etc. Ces variables sont incluses dans les cinq modèles statistiques considérés: le modèle de débit de filtration glomérulaire estimé (eGFR), le filtre de Kalman de test de laboratoire (LKF), le filtre de Kalman de texte (TKF), de test de laboratoire et de filtre de Kalman de texte (LTKF) et les tests de laboratoire récents (RLT)). Les patients ayant développé une IRC de phase III (définie comme le débit de filtration glomérulaire estimé (DFGe) avaient systématiquement <60 ml / min / 1,73 m2 pendant 3 mois). Le résultat d'intérêt a été défini comme la progression vers la phase IV de la CKD lorsque (DFGe toujours <30 ml / min / 173 m2 pendant 3 mois).

Les auteurs ont combiné l'analyse des séries chronologiques (filtre de Kalman) avec l'analyse de survie (modèle cox) et le résultat a montré que LTKF a la plus grande précision, car il prend en compte les résultats des tests de laboratoire longitudinaux et de la documentation clinique longitudinale.

Pour [11] Les auteurs proposent un système d'inférence neurofuzzy adaptatif (ANFIS) pour prédire le début de l'insuffisance rénale en se basant sur sur une base de données cliniques réelles et sur le taux de filtration glomérulaire (TFG), et ils le comparent avec un TFG

réel avec MATLAB afin de fournir une méthode avec une précision acceptable pour trouver la base d'un système d'aide à la décision dans les soins de santé et pour soutenir les systèmes de décision clinique. Les données utilisées proviennent des dossiers cliniques d'une étude de cohorte de patients nouvellement diagnostiqués d'IRC qui ont été admis en série à la clinique de néphrologie; l'hôpital Imam Khomeini (Téhéran, Iran), entre octobre 2002 et octobre 2011. Au total, 465 patients atteints d'IRC ont été inclus dans l'étude. Le groupe de test était composé de 389 patients. Le groupe validation était composé de 76 patients. Les variables utilisées comprennent des données cliniques et de laboratoire.

Les critères d'inclusion pour la définition de l'IRC comprennent un rein de petite taille dans les images échographiques ou un DFG inférieur à 60 cc / kg / min / 1,73 m2 pendant plus de 3 mois. La valeur seuil de 15 cc / kg / min / 1,73 m2 de débit de filtration glomérulaire (DFG) a été utilisée comme marqueur de l'insuffisance rénale.

Les comparaisons des valeurs prédites avec les données réelles ont montré que le modèle ANFIS pouvait estimer avec précision les variations du DFG dans toutes les périodes séquentielles (erreur moyenne absolue normalisée inférieure à 5%).

Concernant [12] Les auteurs ont tenté dans cette étude d'identifier les facteurs prédictifs de la récupération rénale chez les patients présentant une insuffisance rénale secondaire due à une lithiase urinaire obstructive bilatérale. Les données utilisées provenaient des dossiers électroniques de patients adultes atteints de lithiase urinaire obstructive bilatérale entre janvier 2007 et avril 2011, des dossiers médicaux électroniques de l'urologie, Christian Medical College en Inde.

Les variables étudiées étaient l'âge, le sexe, la durée des symptômes, l'emplacement de la pierre, le nombre et la taille, l'infection, l'épaisseur maximale du parenchyme rénal, le temps de créatinine et la présence de facteurs co-morbids. La récupération rénale a été définie comme un nadir de créatinine ≤ 62 mg / dL.

L'analyse de survie a été utilisée pour évaluer le temps nécessaire pour atteindre le nadir de la créatinine. Les courbes de survie ont été obtenues à l'aide des estimations de Kaplan-Meier pour l'absence d'hypertension, la durée des symptômes et l'épaisseur maximale du parenchyme rénal. Des statistiques de log-rank ont été utilisées avec SPSS ont été utilisés pour analyser les données.

Les résultats ont montré qu'une durée plus courte des symptômes (≤ 25 jours) est prédictive de la récupération rénale en cas d'insuffisance rénale secondaire liée à une lithiase urinaire obstructive bilatérale.

Les auteurs de [13] proposent un nouveau modèle génératif probabiliste qui peut fournir des prédictions individualisées de la progression future de la maladie tout en modélisant conjointement le modèle des événements indésirables récurrents associés. Ils ajustent leur modèle à l'aide d'un algorithme d'inférence variationnelle évolutif et appliquent la méthode à un vaste ensemble de données de dossiers de santé électroniques longitudinaux des patients. Le modèle donne des performances supérieures en termes à la fois de prédictions des trajectoires futures de la maladie et des futurs événements graves par rapport aux modèles non articulaires. Les prévisions sont actuellement utilisées par l'organisation de soins responsable locale lors de l'examen des dossiers des patients à haut risque. L'objectif est de développer des méthodes statistiques qui modélisent à la fois les risques de perte future de la fonction rénale et les risques de complications futures ou d'événements de santé indésirables et les prévisions de ces modèles, qui peuvent ensuite être utilisées par les organisations de soins de

santé pour connecter les patients à haut risque de manière appropriée et permettre une intervention

Plus ciblée.

L'ensemble de données utilisé était une grande cohorte de patients atteints d'IRC du système de santé de l'Université Duke; l'approche est formulée comme un modèle de variable latente hiérarchique. Chaque patient est représenté par un ensemble de variables latentes caractérisant à la fois la trajectoire de sa maladie et le risque d'avoir des événements. Cette approche capture les dépendances entre la trajectoire de la maladie et le risque d'événement. Cette étude est caractérisée par l'utilisation d'un mécanisme d'inférence via un algorithme d'inférence pour mettre à l'échelle le grand ensemble de données.

Pour [14] , Les auteurs de cette étude ont développé un cadre pour améliorer la précision des modèles d'induction de règles et d'arbres de décision pour prédire l'insuffisance rénale. Les données utilisées ont été collectées auprès des hôpitaux Apollo, du Tamil Nadu et de l'Inde. Avec 24 variables, y compris la variable de classe patient avec CKD ou NON CKD, l'ensemble de données initial n'était pas équilibré. Et si l'ensemble de données est déséquilibré, les modèles traditionnels ne peuvent pas produire de résultats précis. Ainsi, le cadre proposé améliore la précision des modèles en équilibrant l'ensemble de données déséquilibré. Pour cela, les auteurs ont appliqué un algorithme de rééquilibrage SMOTE sur un jeu de données déséquilibré pour le rendre équilibré. Plusieurs classificateurs sont utilisés comme l'algorithme d'arbre de décision et la méthode basée sur des règles (Rule Based Method), les résultats produits prouvent une précision croissante. Les résultats sont également comparés à un jeu de données équilibré et déséquilibré. Cette méthode atteint une précision moyenne de 98,73%.

Pour [15], et en utilisant des techniques de fouille de données, les auteurs tentent de spécifier les paramètres efficaces dans la prédiction de la maladie d'insuffisance rénale chronique et aussi de déterminer leurs relations les uns avec les autres chez les patients iraniens atteints d'IRC. 23 variables sont utilisées; la population statistique de l'étude comprend 31996 enregistrements de données enregistrés dans la base de données, de l'hôpital SINA de l'Université des sciences médicales de Téhéran. Après la compréhension, la préparation et le nettoyage des données avec la méthodologie CRISP-DM (Cross Industry Standard Process for Data Mining), des outils d'exploration de données ont été utilisés pour trouver les règles cachées et les relations entre les paramètres dans les données collectées. L'algorithme appliqué aux données de cette étude est les règles d'association (Association Rules). Après avoir exécuté des algorithmes de fouille de données sur la base de données, les relations entre les paramètres effectifs ont été spécifiées.

Les auteurs de [16] tentent de détecter une association entre les paramètres du patient et l'échec précoce de la Fistule artérioveineuse AVF, car il s'agit d'un accès vasculaire important pour le traitement par hémodialyse (HD) mais a un taux d'échec précoce de 20 à 60%. Il est également important de réduire sa prévalence et les coûts correspondants. Ils essaient également de prédire l'incidence de cette complication chez les nouveaux patients, ce qui est une procédure de contrôle bénéfique. La sécurité des patients et la préservation de l'échec précoce de la FAV sont le but ultime. Les recherches ont eu lieu au Hasheminejad Kidney Center (HKC) de Téhéran, qui est l'un des plus grands hôpitaux rénaux d'Iran. Les données de 193 patients ont été analysées en utilisant des techniques supervisées de l'approche d'exploration de données.

Il y avait 137 hommes (70,98%) et 56 femmes (29,02%) patients introduits dans cette étude. L'âge moyen de tous les patients était de 53,87 ± 17,47 ans. Vingt-huit patients etaient fumeurs et le nombre de patients diabétiques et non diabétiques était de 87 et 106, respectivement. Une relation significative a été trouvée entre «diabète», «tabagisme» et «hypertension» avec échec précoce de la FVA dans cette étude. La technique utilisée est l'apprentissage automatique pour prédire l'échec précoce de l'AVF et déterminer les facteurs de risque qui y jouent un rôle important. Les algorithmes d'apprentissage utilisés dans ce travail sont JRIP et J48, et ont été testés sur l'environnement Waikato pour l'analyse des connaissances (weka).

Ils ont constaté que ces facteurs de risque mentionnés ont un rôle important dans l'issue de la chirurgie vasculaire, par rapport à d'autres paramètres tels que «l'âge». Ensuite, ils ont prédit cette complication dans les futures chirurgies AVF et évalué les méthodes de prédiction conçues avec des taux d'exactitude de 61,66% à 75,13%.

3. Discussion

Il existe deux types d'insuffisance rénale: l'insuffisance rénale aiguë due à une défaillance soudaine de la fonction rénale, et l'insuffisance rénale chronique, qui résulte du déclin lent et progressif de la fonction rénale [17], on constate qu'il existe plusieurs méthodes pour estimer ce DFR comme: La créatinine sérique, molécule produite par le métabolisme musculaire, éliminée principalement par filtration glomérulaire, qui constitue un marqueur rénal couramment dosé mais manquant de sensibilité. La mesure de la clairance, par dosage de la créatinine sérique et urinaire, présente l'inconvénient de nécessiter un prélèvement urinaire de 24 heures, sous réserve d'incertitudes.

Pour cela plusieurs formules sont développées, dont la formule The Cockcroft et Gault, qui prend en compte le poids, l'âge et la créatinine, et permet à l'adulte d'estimer la clairance de la créatinine [18]. Cette clairance est très proche du DFG et fournit des informations sur l'état de la fonction rénale et plus récemment un algorithme a été développé pour estimer le taux de filtration glomérulaire basé sur la créatinine sérique, l'âge, le sexe et l'origine ethnique du patient baptisé MDRD (Modification of Diet in Renal Disease) [19]. La formule MDRD rend le GFR normalisé sur la surface corporelle en ml / min / 1,73 m2. Il estime souvent les DFG supérieurs à 602.

Le Programme national d'éducation sur les maladies rénales (NKDEP) recommande l'équation MDRD comme la meilleure approche pour évaluer la fonction rénale chez les patients atteints d'insuffisance rénale chronique et pour identifier les patients à risque [20].

La fonction rénale est décrite en plusieurs niveaux selon le DFG [21]:

Niveau 1: ≥ 90 ml / min / 1,73 m2, DFG normal ou augmenté;

Niveau 2: 60-89mL / min / 1,73m2, DFG légèrement diminué;

Niveau 3: 30-59mL / min / 1,73m2, insuffisance rénale chronique modérée;

Niveau 4: 15-29mL / min / 1,73m2, insuffisance rénale chronique sévère;

Niveau 5: <15 ml / min / 1,73 m2, insuffisance rénale terminale chronique.

L'insuffisance rénale chronique est généralement définie par un débit de filtration glomérulaire inférieur à 60 mL / min. Une dialyse ou, si possible, une transplantation rénale doit être effectuée si le DFG est inférieur à 10 ou 15 mL / min.

Toutes les recherches effectuées dans le cadre de la prédiction de l'insuffisance rénale étaient basées sur ces constantes médicales. Ce qui diffèrent, ce sont juste les techniques utilisées; il y a des auteurs qui ont appliqué des techniques d'analyse de données telles que le mécanisme d'inférence par un algorithme pour la simulation du raisonnement déductif, donc un moteur d'inférence permet aux systèmes experts de conduire un raisonnement logique et de tirer des conclusions basées sur des faits et une base de connaissances [22] ou filtre de Kalman [23] [24] qui estime les états d'un système dynamique à partir d'une série de mesures [25] [26].

Nous remarquons que les variables dépendantes du patient, telles que la glycémie rapide, la créatinine, le cholestérol, les lipoprotéines de haute densité, les triglycérides, l'hémoglobine, le nombre de globules rouges, le nombre de globules blancs… etc., ne dépassent pas 24 au total, y compris les habitudes quotidiennes variables (fumeurs ou non-fumeurs) sont primordiales dans le diagnostic de la maladie et bien évidemment pour la prédire.

La prévision de l'insuffisance rénale est basée sur des données médicales précises constituées de tous les antécédents médicaux du patient et dans un intervalle de temps donné, puis sur l'application des nouvelles techniques de fouille de données et d'apprentissage automatique.

4. Expérimentation

La maladie de l'insuffisance rénale chronique (IRC) comprend un large éventail de processus physiopathologiques qui seront observés avec une fonction anormale des reins et une diminution progressive du taux de filtration glomérulaire (DFG). Afin de prédire l'insuffisance rénale, il est nécessaire d'utiliser l'ensemble des données chronologiques concernant toutes les variables dépendantes de l'IRC, et aussi des informations précises sur les antécédents médicaux du patient.

De préférence, la base de données doit provenir d'une organisation médicale (hôpital, clinique) et contenir toutes les informations et données médicales sur des patients (hommes ou femmes) d'âges différents et sur un intervalle de temps de dix ans ou plus. Les patients doivent avoir fréquenté l'organisation médicale plus d'une fois après leur premier diagnostic.

Nous avons choisi d'implémenter notre algorithme avec Spark Databrikcs, une plateforme qui combine toutes les composants de gestion et d'analyse du Big Data en Cloud, ce qui permettra de bénéficier de tous les avantages du travail en mode Cloud.

Notre implémentation s'est déroulée en étapes :

4.1 Préparation des données

La base de données médicales réelles que nous avons reçue comprend des informations sur une période de 10 ans, tous les paramètres des laboratoires et cliniques sont enregistrés sur des intervalles de 6 mois pour chaque patient, nous avons pour chaque paramètres des

valeurs allant de t = 0 jusqu'à t = 120 mois. La base de données comprend 503 patients en totalité.

Le modèle de prédiction que nous allons établir, va prédire les valeurs de GFR ainsi que sa progression. Pour réaliser cet objectif, nous allons prédire les valeurs de GFR à l'instant t = 0, et après six mois avec t = 1, sur la base des valeurs des variables influençant (inputs) le Taux de Filtration Glomérulaire. Ces variables sont : (âge, sexe, poids, cause, dbp, sbp, creat, gfr).

Toutes les valeurs des variables (inputs) doivent être numériques, et puisque le sexe et la cause logiquement signifient des termes avec une chaîne de caractères, nous leur avons attribué des codes, [0.1] pour la variable sexe [féminin, masculin]. Et pour la cause 'underlying disease', les valeurs numérique sont associée à chaque type de cause comme suit :

1= (diabetes mellitus (type 2 diabetes)), 2= (glomerulopathy), 3= (hypertension), 4= (urologic), 5= (unknow) , 6= (pkd= 'polycystic kidney disease) , 7=(dm+urologic (diabetes mellitus + urologic disease)), 8= (Renovascular stenosis).

4.2 Protocole d'Expérimentation

Nous allons utiliser la régression linéaire avec une régularisation. La méthode choisie est selon le type da la base de données que nous avons, une base de données en série temporelles, et aussi pour une raison scientifique, tous les travaux de recherche faits dernièrement se focalisent plus sur d'autres méthodes d'apprentissage automatique, et que l'implémentation de la régression linaire régularisé sur Spark constitue une nouveauté.

4.3 Environnement matériel et logiciel

le langage Python est utilisé pour l'exploitation de la bibliothèque d'apprentissage automatique de Spark MLlib .

Après la création du (Cluster) et du (Notebook) dans le (Workspace) dédié, les données doivent être soigneusement préparées et importées en format .csv, et puis chargé dans DBFS (DataBricks File System), et bien préciser le type de chaque colonne car c'est primordiale pour la visualisation par la suite.

Les lignes qui manquent certaines valeurs doivent être supprimées lors de la préparation des données et avant l'implémentation de l'algorithme d'apprentissage, pour avoir une meilleure valeur de précision.

4.4 Métriques d'évaluation

Pour évaluer la performance d'un modèle de régression pour la prédiction d'une variable continue à partir d'un certain nombre de variables indépendantes, il faut faire référence aux métriques suivantes :

L’erreur quadratique moyenne (MSE)

Le carré moyen des erreurs ou erreur quadratique moyenne (MSE pour Mean Square Error) c’est la moyenne arithmétique des carrés des écarts entre les prévisions et les observations. C’est la valeur à minimiser dans le cadre d’une régression simple ou multiple :

$$MSE = SCR \div n$$

Dans notre cas la régression est multiple donc SCR / (n – k – 1) avec (k = nombre de variables explicatives), SCR est la somme des carrés des résidus (SCR ou Sum of Squared Error), et n = nombre de variables.

L'erreur-type (RMSE)

L’erreur-type (RMSE) ou Root-Mean-Square Error (RMSE), est une règle de notation quadratique qui mesure également la grandeur moyenne de l'erreur. C'est la racine carrée de la moyenne des différences carrées entre la prédiction et l'observation réelle.

$$\text{RMSE}= \sqrt{1 / n \sum_{j=1}^{n} (y - \gamma)\ 2}$$

L’erreur absolue moyenne (MAE) L’erreur absolue moyenne (MAE pour Mean Absolute Error) : moyenne arithmétique des valeurs absolues des écarts.

$$\text{MAE} = 1 / n \sum_{j=1}^{n} | y - \lambda |$$

Coefficient de détermination

Le coefficient de détermination (R2) ou encor (R-squared) est une mesure de la qualité de la prédiction d'une régression linéaire. Il détermine aussi à quel point l'équation de régression est adaptée pour décrire la distribution des points.

$$R^2 = 1 - \frac{SCR}{SCT}$$

5. Résultats

Après l’implémentation de la régression linaire sur la base de données contenant les variables à l’instant t = 0, Spark nous a permis de visualiser la distribution de gfr0m , selon les variables prédictives (features) voir FIG.1. Ou bien visualiser la distribution de gfr0m selon chaque variable prédictive.

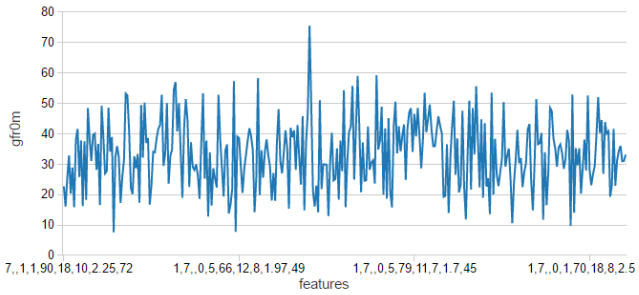


FIG. 1 – *Variation de gfr0m en fonction de tous les prédicteurs*

Le modèle de prédiction est formé sur un ensemble qui constitue 60 % de la base de données initiale, et un ensemble de test de performance de prédiction de 40%.

Notre modèle de prédiction ne doit pas apprendre aveuglement les données, pour ne pas tomber dans des situations de sur-apprentissage connues par (overfitting), où le modèle aura de la peine à généraliser les caractéristiques des données. Il va se comporter par la suite comme une table contenant tous les échantillons utilisés lors de l'apprentissage et perd ses pouvoirs de prédiction sur de nouveaux échantillons, malgré sa trop grande capacité à stocker des informations.

A ce niveau la régularisation fait référence à un processus consistant à ajouter de l'information à un problème pour éviter le sur-apprentissage. Cette information prend généralement la forme d'une pénalité envers la complexité du modèle.Le paramètre de régularisation le plus optimal dans ce contexte est $\lambda = 0.1$

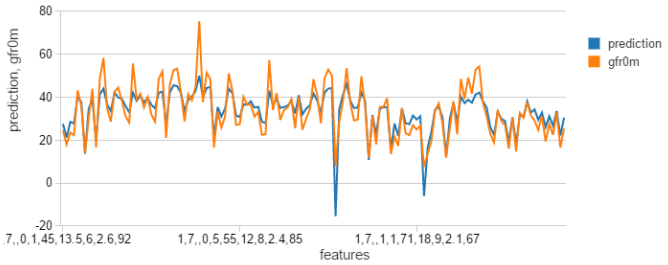


FIG. 2 – *Variation de gfr0m et gfr prédite (prédiction) sur l'ensemble de test*

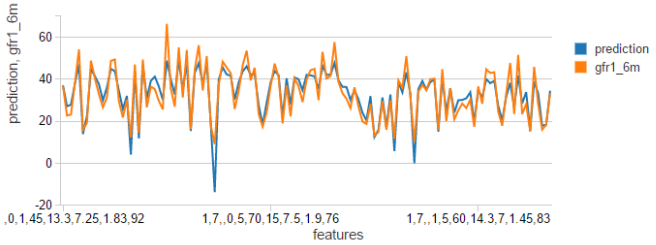


FIG. 3 – *Variation de gfr6m et gfr prédite (prédiction 6m) sur l'ensemble de test*

	MSE		RMSE		MAE		R-Squared	
	Ensemble de Formation	Ensemble de Test	Ensemble de Formation	Ensemble de Test	Ensemble de Formation	Ensemble de Test	Ensemble de Formation	Ensemble de Test
t=0m	15.500	31.308	3.937	5.595	2.960	4.020	0.848	0.707
t=6m	31.720	23.501	5.632	4.847	4.296	3.733	0.746	0.806

TAB. 1 – *Comparaison des métriques de performance pour les ensembles de données de formation / test*

La visualisation des résultats de l'implémentation de la régression linaire régularisée, sur la base de données CKD, a montré que les techniques d'apprentissage automatique via les algorithmes dédiés, permettent vraiment de prédire la variation du taux de filtration glomérulaire (gfr), afin de décider par la suite les stades de la progression de la maladie.

Sur la période t = 0, le modèle de prédiction a montré une performance sur l'ensemble de test (voir FIG.2) avec un taux de coefficient de détermination respectivement de 0.7 (voir TAB.1).

Le modèle aussi a montré sa performance sur la période de six mois, en marquant une forte prédiction pour l'ensemble de test (voir FIG.3) avec un taux de coefficient de détermination 0.80 (voir TAB.1).

Le MSE permet de répondre à la question « quelle est la magnitude de l'erreur de la prévision», mais n'indique pas la direction des erreurs. Parce qu'il s'agit d'une quantité au carré, MSE est influencée plus par les grandes erreurs que par les petites erreurs. Sa portée est de 0 à l'infini, un score de 0 étant un score parfait n dans notre étude cela varie entre 15.5 et 31 pour l'ensemble de formation (training set) et entre 23 et 31 .

Il est important de rappeler que RMSE a la même unité que la variable dépendante. Cela signifie qu'il n'y a pas de bon ou mauvais seuil absolu. Pour un point de référence qui varie de 0 à 1000, un RMSE de 0,7 est petit, mais si la plage passe de 0 à 1, cette grandeur aura une importance. Cependant, bien que plus le RMSE est petit, mieux vaut. Dans notre étude le RMSE varie entre 3.93 et 5.63 pour l'ensemble de formation et entre 4.84 et 5.59 pour l'ensemble de test.

MAE ou l'erreur absolue moyenne, la quantité utilisée pour mesurer à quel point les prévisions sont proches des résultats éventuels. Nous avons obtenu un taux qui varie entre 2.96 et 5.29 pour l'ensemble de formation et 3.73 et 4.02 pour l'ensemble de test.

Finalement quoique cela soit les valeurs de MSE, RMSE, MAE, ces valeurs sont tenues en compte et considérées comme meilleures, que lorsqu'elle offre une précision élevée. Et Il n'y a pas de plage de valeurs acceptables pour le calcul des erreurs en général, les valeurs dépendent au dataset et de l'objectif souhaité. Ce qui détermine la qualité de la prédiction du modèle et son efficacité et le coefficient de détermination.

Notre modèle de prédiction a montré sa performance en termes de prédiction du taux de filtration glomérulaire (GFR) ce qu'est prouvé avec les valeurs du coefficient de détermination entre 0,7 et 0,8.

6. Conclusion

La prévision des maladies rénales est un domaine fertile pour les chercheurs qui souhaitent développer des mécanismes prédictifs capables de lutter contre cette maladie.

En effet plus les données sont précises, plus la prédiction est ciblée et mènera à la meilleure prise de décision. La régression linaire a montré son efficacité dans le contexte de prédiction de la variation de progression de GFR.

Les techniques appliquées d'exploration des données et, en particulier, l'arrivée du Big data et d'analyse prédictive de données donnent plus d'espoir et plus de chances d'affaiblir cette maladie, arrêter sa progression, anticiper le traitement et de minimiser les dépenses financières vue coût élevé du traitement.

Ces promesses sont prouvées par les résultats obtenus dans la prédiction de la maladie, quoique ce soit la technique ou la méthode utilisée, l'objectif reste unique, sauver des vies humaines.

Références

[1] Josef C., Elizabeth S., Lesley A. S., and all, ''Prevalence of Chronic Kidney Disease in the United States'', JAMA, November 7, 2007—Vol 298, No. 17 , pp 2038-2047

[2] Malvinder,S.P.,"Chronic renal disease",clinical review,BMJ VOL. 325 N.13 JULY 2002, pp 85-90

[3] Ernesto L. S., Mark L. L., Johannes F.E. M.," Chronic Kidney Disease, Effects on the Cardiovascular System",Circulation,July 2007,pp.85-97

[4] Javed ,B., Daniel E, F., William T A., and all ''Relationship between heart failure treatment and development of worsening renal function among hospitalized patients'',American Heart Journal, February 2004Volume 147, Issue 2, pp 331–338

[5] Martin M., Michael D.,'' Medical error—the third leading cause of death in the USA '', British Medical Journal ,May 2016,

[6] Alan S. G, Glenn M. C., Dongjie F., and all ''Chronic Kidney Disease and the Risks of Death, Cardiovascular Events, and Hospitalization'', The new england journal of medicine, september 23, 2004, pp 1296-1305

[7] Cockcroft D.W. · Gault M.H. ''prediction of Creatinine Clearance from Serum Creatinine'', 1976, Vol.16, No. 1, pp 31-41

[8] ANAES Agence Nationale d'Accréditation et d'Évaluation en Santé / Service des recommandations et références professionnelles ''Diagnostic de l'insuffisance rénale chronique chez l'adulte Recommandations'' Septembre 2002 pp.1-27

[9] Andrew S. L., KAI-UWE E., YUSUKE T., and all , '' Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO)'', Kidney International by International Society of Nephrology, Vol. 67 (2005), pp. 2089–2100

[10] Adler P.,, Rajesh R., Jamie S H., and all ,'' Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis'', Published by Oxford University Research and Applications ,2015, pp 872-880
[11] Jamshid N.,Ali Y., Seyed Ahmad M., and all ,'' Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System'''' Hindawi Publishing Corporation , Computational and Mathematical Methods in Medicine, 2016, Article ID 6080814, 9 pages
[12] Muthukrishna P. R., Chandrasingh J.B., Grace J. R., '' Predictors of renal recovery in renal failure secondary to bilateral obstructive urolithiasis'', Arab Journal of Urology ,2016, vol 14, pp. 269–274
[13] Joseph F.,, Mark Sendak C. Blake C., '' Scalable Joint Modeling of Longitudinal and Point Process Data for Disease Trajectory Prediction and Improving Management of Chronic Kidney Disease'', UAI'16 Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, June 25 - 29, 2016 , Pages 222-231
[14] Sai Prasad P., Sreedevi, M.,''An Improved Prediction of Kidney Disease using SMOTE'', Indian Journal of Science and Technology, Vol 9(31)
[15] Shahram T., Marjan G., Mostafa L., and all, '' Applying data mining techniques to determine
important parameters in chronic kidney disease and the relations of these parameters to each other'', Journal of Renal Injury Prevention, 2017, Vol, 6, N 2, pp. 83-87
[16] Mohammad R., Morteza Khavanin Z.,Mohammad Mehdi S., ''Implementation of Predictive Data Mining Techniques for Identifying Risk Factors of Early AVF Failure in Hemodialysis Patients'', Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine , 2013, Article ID 830745, 8 pages
[17] john, S., '' Acute-on-chronic kidney disease: prevention, diagnosis, management and referral in primary care'',Best Practice Journal , September 2012, pp. 10-15
[18] Kathleen D. L,Glenn M. C, '' ACUTE RENAL FAILURE'',Jameson Nephrology,,McGraw-Hill Medical, January 2011 ,chapter 10, pp 97-112.
[19] Earley A, Miskulin D, Lamb EJ, Levey AS, Uhlig K, Estimating equations for glomerular filtration rate in the era of creatinine standardization: a systematic review [archive], Ann Intern Med, 2012;156:785-95
[20] Gary L. M., Greg M., Josef C., ''Recommendations for Improving Serum Creatinine Measurement: A Report from the Laboratory Working Group of the National Kidney Disease Education Program'', Clinical Chemistry Journal January 2006, Vol .52.
[21] Rinaldo B., Claudio R.,John A K., Ravindra L M., and all,'' Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group'', Critical Care Journal , May 2004, Vol. 8, No .4,pp. 204-212
[22] Matsushita K, Mahmoodi BK, Woodward M et al. Comparison of risk prediction using the CKD-EPI equation and the MDRD study equation for estimated glomerular filtration rate, JAMA, 2012;307:1941-51
[23] Stephan,W.,Gerhard,W., '' Evaluating the Inference Mechanism of Adaptive Learning Systems'', june 2003, International Conference on User Modeling, pp154-162
[24] Paul Zarchan; Howard Musoff (2000). Fundamentals of Kalman Filtering: A Practical Approach. American Institute of Aeronautics and Astronautics, Incorporated. ISBN 978-1-56347-455-2.
[25] Kalman, R. E. (1960). "A New Approach to Linear Filtering and Prediction Problems". Journal of Basic Engineering. 82: 35. doi:10.1115/1.3662552.
[26] Steffen L. Lauritzen. "Time series analysis in 1880. A discussion of contributions made by T.N. Thiele". International Statistical Review 49, 1981, 319–333. JSTOR 1402616

Une analyse NLP du flux Twitter Covid/Corona – Confinement 1 : la montée du masque.

Christophe Benavent*, Mihai Calciu**
Julien Monnot***, Sophie Balech****

*Université Paris Nanterre, Christophe.benavent@parisnanterre.fr
**Université de Lille,mihai.calciu@univ-lille.fr
*** Julien Monnot, jmonnot52@gmail.com
**** Université de Picardie, sophie.balech@u-picardie.fr

Résumé. La pandémie de Covid-19 qui frappe la planète propose un cas spectaculaire de management du désastre. L'atténuation de l'impact de la catastrophe, la qualité de la préparation et la résilience de la société, facilitent la reconstruction, mais dépendent de la participation des populations. Pouvoir observer et mesurer l'état de santé mentale des populations sont des nécessités pour accompagner les mesures destinées à l'encourager. Les médias sociaux, et en particulier Twitter, offrent des ressources précieuses pour explorer ce discours. Le résultat principal repose sur l'identification du caractère central de la figure du masque et vise à établir l'importance du phénomène. Nous l'établissons de manière quantitative par les méthodes de NLP en exploitant un corpus de 550 K tweets extraits sur la période de février à fin mai 2021.

Mots-clés : Covid-19, disaster management, NLP, STM, social media.

1. Introduction

L'épisode de pandémie de la Covid-19 en illustre l'importance. Sans l'adoption scrupuleuse de la distance sociale, il est difficile de contenir l'épidémie : les pays qui ont réussi à casser la (première) vague l'ont fait au prix d'un confinement strict. Dans l'attente d'un traitement et de la vaccination, la capacité à maintenir une distance entre les sujets biologiques par l'observance des gestes barrières, est un enjeu essentiel. La participation des populations est nécessaire pour le relever. Ecouter ce qu'elle en pense, comprendre comment le discours se construit est le rôle traditionnel des enquêtes d'opinion que les méthodes modernes de "social listening" complètent aujourd'hui en apportant de nouveaux matériaux provenant des médias sociaux.

Cette étude participe au mouvement de " la science immédiate", puisque les données exploitées couvrent la période de février à mai 2020, soit la totalité de la période du premier confinement en France. L'approche largement descriptive et longitudinale, vise essentielle-

ment à construire une microhistoire, par l'analyse des contenus de Twitter, en traitant par des outils NLP (sentiment, annotations, POS, dépendance syntaxique) un corpus de 566 000 posts répartis sur une période de 22 semaines.

La position centrale du masque dans le flux discursif est le résultat empirique principal : une figure polaire, qui condense le matériel et le symbolique. On en observe les variations sémantiques au cours du temps. Nous dessinerons d'abord un cadre d'analyse issu d'une discipline de gestion née dans le drame de la guerre et dans l'expérience des catastrophes naturelles, qui emprunte à l'anthropologie de la technique l'idée que les conventions matérielles, incarnées dans des objets quotidiens offrent moins une capacité technique (filtrage pour le masque), qu'un espace social commun qui tisse et coordonne l'effort collectif d'atténuation de l'épidémie. Le masque serait ainsi ce par quoi les relations sociales se redéfinissent et qui en forme une réalité partagée.

2. Cadre conceptuel

Si la catastrophe semble être singulière pour ceux qui la vivent, elle est un phénomène finalement ordinaire : tempêtes et ouragans, séismes, inondations, accidents industriel et les épidémies en sont quelques-unes des manifestations fréquentes, à chaque fois unique par leur sévérité et leurs conséquences matérielles. Elles varient essentiellement par leur échelles, elles posent à la question de la décision, celle de la contingence : celle de la crise dont Ian Mittrof est un des pionniers (Mitroff 1986).

2.1 Quelques éléments de "disaster management"

C'est à la frontière de plusieurs disciplines, avec l'expérience d'instituts comme la Croix Rouge et la sécurité civile, la médecine d'urgences, la médecine de guerre, la gestion du risque, que s'est développée une discipline : le "disaster management" dont l'ouvrage de Havidán Rodríguez, Quarantelli, et Dynes (2007) est devenu un classique. En 2005, l'Epidemiologic Reviews publiait un numéro spécial d'état de l'art de la discipline (Noji 2005), soulignant l'accroissement de la complexité, et rappelant que l'on ne sort jamais de la crise de manière immédiate. Dans cette littérature trois concepts sont clés et nécessaires à tous les stades du désastre : le premier est l'état de préparation aux conséquences du phénomène et à ses répliques, le second concerne celui de l'atténuation de ces conséquences qui nécessite la participation des populations, le troisième est lui relatif à la capacité qu'a chacun de rebondir et de s'engager dans la voie de la reconstruction : la résilience.

La conception moderne du management du désastre, en mettant l'accent sur la participation des populations, interroge les facteurs qui l'encouragent ou la réfrènent. Les ressources matérielles, cognitives et organisationnelles sont évidentes, l'entraînement et la préparation sont déterminants, mais in fine, c'est sans doute la santé mentale des populations et leur niveau d'engagement et de participation qui font la différence. Il semble qu'un cycle de vie anime la dynamique de la santé mentale pendant les catastrophes. Le modèle de Zunin et Myers est souvent cité, il suppose une succession de plusieurs phases.

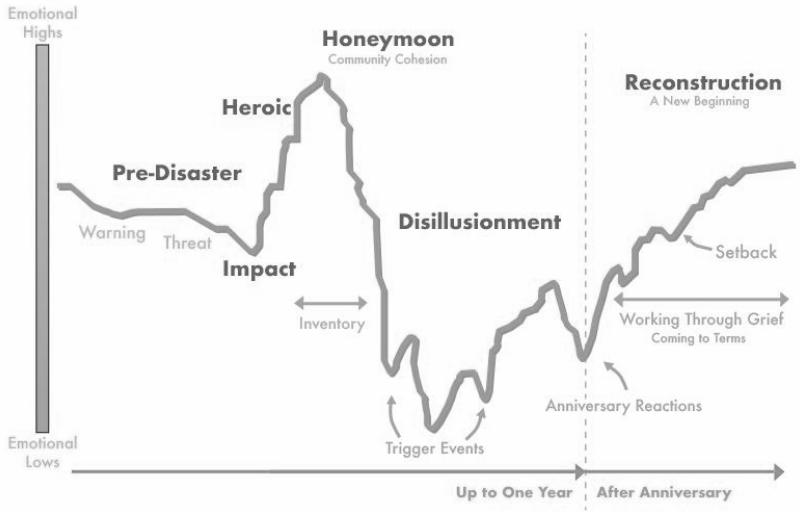


FIG. 1 – Evolution de la densité des termes de contexte et de geste barrière dans le corpus

La première est la phase de l'impact qui s'accompagne de stupéfaction, d'autant plus que ce dernier est brutal. Cette phase dure le temps que des catégories de pensée communes surgissent pour comprendre, donner un sens à l'événement et orienter les décisions. Une deuxième phase, "héroïque", dure peu de temps et se caractérise par de fortes réactions émotionnelles mais peu productives pour faire face à la catastrophe. La troisième, qualifiée de "lune de miel", débute lorsque les forces de lutte sont en place, qu'une stratégie d'atténuation est engagée, et qu'apparaît une réorganisation dans laquelle chacun trouve une sorte d'apaisement. La quatrième séquence est celle des désillusions qui se prolonge d'autant que la phase de reconstruction n'est pas encore en vue. Elle évolue en fonction des événements, passant par la désespérance (quand la circulation du virus s'accélère) et des lueurs d'espoirs (quand des annonces de vaccin et de tests sont proclamées). La dernière et ultime phase est celle de la reconstruction à moins qu'une réplique vienne à nouveau enclencher le cycle en fonction du degré de préparation qui aura été acquis précédemment.

2.2 Un objet singulier dans les stratégies d'atténuation : le masque.

Le masque fait partie de la panoplie des gestes barrières. Sa particularité est que s'il est une des pratiques d'atténuation les moins observées au début du confinement, c'est pourtant celle qui s'est le plus généralisée. Il a été l'objet de polémiques multiples (au sujet de son absence, de son efficacité, de son prix, de sa disponibilité ou encore de son obligation) le constituant en quelques semaines comme controverse, puis comme norme.

Symboliquement, le masque protège, mais son port est en lui-même un signal, qui marque le degré de participation à l'effort d'atténuation, s'il cache le visage, c'est pour maintenir l'interaction sociale, c'est une nouvelle frontière dont l'ambiguïté est de séparer pour réunir. Phénoménologiquement, le masque est un objet matériel ambivalent : fonctionnellement, il est, dans l'usage chirurgical, destiné à protéger les autres plutôt que soi. Au tournant

de mars 2020, quand le confinement a été imposé à pas accélérés, en dépit des élections municipales, et que l'impact de l'épidémie ne s'était pas encore produit, la polémique du masque a surgi, avec la pénurie qui a frappé les soignants. L'histoire du masque et de son stock se noue dans les alertes épidémiques de 2005 et 2009, et se développe dans l'interaction de la production scientifique qui se penche sur son efficacité avec la doctrine des décideurs et des glissements administratifs de la gestion du stock stratégique (Steyer 2020).

C'est surtout un objet social, dont une réglementation négociée depuis le début de l'épidémie règle autant ses modes de production par des normes techniques que son usage par des mesures différenciées d'obligation, ce qui lui confère une existence au-delà de son utilité individuelle et collective, celle d'une norme, au sens juridique, mais aussi sociologique, il se trouve au cœur des interactions sociales, et de la redéfinition des scripts d'interactions interpersonnelles au travail, à l'école, dans les transports, les commerces et la sphère des relations privées.

Dans cette perspective, la notion d'objet intermédiaire et celle proche d'objet frontière apporte un éclairage intéressant. Ces concepts ont été développé dans un cadre de sociologie des sciences où, dans un esprit latourien, les objets matériels participent à la vie sociale, en s'appuyant sur un double processus de représentation et de traduction, dans la mesure où l'objet ne véhicule pas seulement l'intention des acteurs, mais transforme les intentions et produit autre chose que ce qui était anticipé (Vinck,2009). Il est à la fois le médiateur et le cadre de l'action. Le concept d'objet frontière proposé Star & Griesemer (1989) se définit ainsi comme ce qui assure la coordination entre des mondes sociaux différents. De tels objets sont caractérisés par leur flexibilité interprétative, ils supportent des définitions, des connaissances, des conceptions multiples, sans consensus préalable, tout en les articulant. Ils se caractérisent aussi par une organisation et une certaine échelle d'action. Faire l'hypothèse que le masque est un de ces objets offre une grille de lecture intéressante, dans la mesure où, il est le moyen d'associer des mondes sociaux bien différents : scientifiques, médecins, personnages politiques, commerçants, citoyens, en leur permettant de construire un discours pour orienter et coordonner leurs actions.

Une autre manière d'appréhender le masque est l'idée du fétiche qui nous offre une meilleure conceptualisation d'un a) objet matériel b) pourvu de puissance sociale, sans que sa légitimité soit rationnelle, c) transitionnel et, d) qui concrétise pour permettre l'action. Le fétiche est ce qui s'impose à tous et avec ce que chacun transige. L'anthropologie de Lemmonier s'intéresse aux objets, tambours, barrières ou pièges à anguille et au motif périssologique, pléonastique, répétitif du rituel (Lemonnier, 2012). Il observe que l'objet est plus grand, plus travaillé, plus ouvragé que n'en réclame sa fonction. C'est qu'il est investi des forces sociales, de ses mythes et ses catégories, d'un travail conjoint dont le produit n'est pas que l'objet mais le tissage des relations par l'activité commune et l'incorporation dans la forme de l'objet des récits communs. La frontière entre technique et rituel s'estompe. Dans cette perspective, une anthropologie du masque viserait à rechercher les éléments de récits qui s'y attachent, les pratiques communes qui renouent les liens sociaux en dépit de la distance biologique qui rompt le lien habituel en instaurant des écrans (masque, hygiaphone, gants, gel) ou en accroissant la distance entre les corps.

Au travers de ces perspectives théoriques se dégage une grille de lecture de la question du masque comme objet au cœur de la redéfinition, de la reconstruction, de la renégociation des relations sociales. Il devient une chose admise par tous autant pour son intérêt que par un gain de légitimité qui lui donne une force normative, qui s'institutionnaliserait dans sa ritualisation. Quand bien même serait-il peu efficace de manière fondamentale, par son mésusage, ou ses qualités intrinsèques, il est ce par quoi l'ensemble des fractions de la société reconsidère les opérations les plus élémentaires de leurs activités : travailler avec des clients et des collègues, enseigner à des étudiants, recevoir des patients, participer à des réunions publiques, aller au spectacle, boire un verre.

3. Méthodes

L'objectif du travail empirique est d'établir un fait : la place centrale et croissante de la figure du masque dans la conversation sociale telle qu'elle apparaît à l'observation. L'approche est essentiellement quantitative et descriptive, dans la mesure où l'on souhaite tester cette hypothèse factuelle. La méthode générale employée s'inscrit dans un nouveau paradigme (Cambria et White 2014) qui se construit entre des données abondantes (web, réseaux sociaux, …) et des techniques de traitement nouvelles issues du champ du traitement automatique du langage naturel. Cela permet d'aller plus loin que l'analyse lexicale traditionnelle en incorporant des éléments syntaxiques, sémantiques, et pragmatiques.

Cette nouvelle approche méthodologique prend place entre l'analyse qualitative et les traditionnelles enquêtes par questionnaire, et se révèle capable de traiter des corpus d'une taille inédite. Dans le champ des sciences sociales, et en particulier du management, on trouvera des synthèses pour la recherche en éthique (Lock et Seele 2015), en comportement du consommateur (Humphreys et Wang 2018), en management public (Kozlowski, Taddy, et Evans 2019) ou en organisation (Kobayashi et al. 2018), sans compter en sociologie avec l'utilisation des word vectors pour analyser l'évolution de la définition des classes sociales (Kozlowski, Taddy, et Evans 2019).

3.1 Données.

Les données sont constituées par un jeu de données est un corpus de contenu Twitter élaboré par (Banda et al. 2020) sur la base d'un ensemble de mots-clé tournant autour de "covid", "corona" et autres mots associés. Ce corpus global intègre près de 3 millions de tweets (en français) produits durant l'année 2020 sur la plateforme Twitter. Plusieurs étapes ont été nécessaires pour reconstituer le corpus et le pré-traiter, avant de pouvoir l'analyser. Le corpus est finalement constitué de 565 662 contenus contributifs : tweet original, réponse et citation dont la distribution de la production dans le temps est donnée dans la figure 1.

L'hétérogénéité du contenu se manifeste pleinement dans la distribution du nombre de tweets émis par compte. Sur un corpus de production primaire de 565 662 tweets, 202 000 comptes y ont contribué, mais de telle matière à ce qu'une ultra minorité (environ 1000) produit 25% du volume total des posts. Le fait principal est une inégalité profonde de la contribution, avec une forte concentration (indice de Gini = 0,50). On y trouve les grands acteurs :

médias, politiques, mais aussi militants, journalistes, chroniqueurs qui peuvent s'y affronter. Sur un jeu de données analogue focalisé sur l'expérience du confinement, des chercheurs approfondissent déjà cette question Boulet et Lebraty (2020). L'extrême hétérogénéité de la population et de sa production pose question, tout autant que la faiblesse de sa taille (le nombre d'utilisateurs quotidiens de Twitter est de l'ordre de 6 millions en France dont moins de la moitié sont actifs).

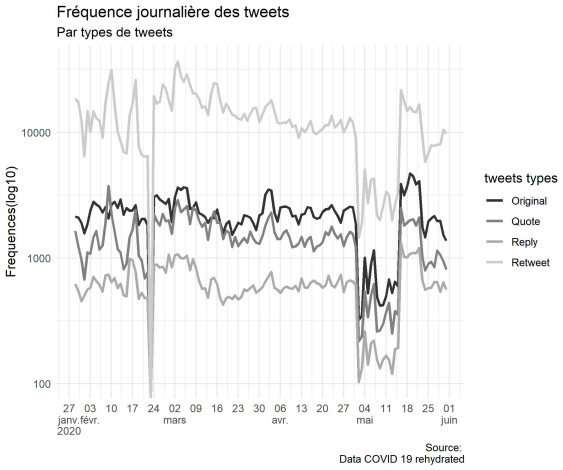


FIG. 1 – Distribution temporelle du corpus par types de posts

3.2 Techniques et modèles d'analyse

Ce corpus fait l'objet d'un prétraitement : nettoyage du corpus, par suppression des Url et autres mention, mise en minuscule, elimination des emojis. Puis d'une annotation par des catégories morpho-syntaxiques (les "part of speech", POS), et l'identification des dépendances syntaxiques, via l'annotateur *Udpipe* de *CleanNLP* (Arnold, 2017). L'ensemble du corpus représente 10 millions de termes dont on filtre les seuls noms communs pour analyser les sujets discutés dans le flux des messages.

Trois types de méthodes sont employées de manière complémentaire pour mettre en évidence les différents aspects de la chronologie des variations des thématiques du discours :

• Une analyse de l'évolution quotidienne des catégories focales de l'étude : un certain nombre de termes cibles ont été identifiés, ceux relatifs à l'épidémie (corona, covid), au confinement et au déconfinement, et ceux naturellement liés aux gestes d'atténuation (masque, gel, gestes barrières, télétravail). Les termes cibles sont identifiés via des regex pour couvrir l'ensemble des variations morphologiques (ie, corana et coronavirus) et leur fréquences relatives par jour sont lissé par une moyenne mobile.

• La production de cartes sémantiques dynamiques : on utilise des cartes sémantiques simples, basées sur les co-occurrences entre les mots les plus fréquents au niveau du document (le tweet), pour explorer l'évolution des discours à l'échelle de la semaine en utilisant une méthode de projection selon l'algorithme de Fruchterman and Reingold. On utilise les

ressources de *igraph*. Sont conservées les cooccurrences supérieures à une valeur proportionnelle au nombre de documents.

• Analyse de thématiques structurales : nous employons ici une méthode d'analyse de topics (Blei, Ng, et Jordan 2003), qui vise à identifier un nombre de k sujets dans un corpus, en prenant en compte le caractère longitudinal des données par la mesure de la prévalence dans le temps de chacun des topics et en supposant une corrélation entre les topics . C'est une application du modèle *STM* proposé par (Roberts et al. 2014)). Le nombre k a été déterminé en explorant les qualité d'ajustement dans une fenêtre de 5 à 50 topics.

4. Résultats

L'objectif de l'étude empirique est de reconstituer une micro histoire des réactions des utilisateurs à l'égard de l'épidémie telles qu'ils les partagent sur les réseaux sociaux. En dépit du volume de données et de la technique utilisée, l'approche est descriptive. Nous développons les résultats des analyses NLP de manière progressive pour de démontrer quantitativement la transformation du rapport au phénomène de la pandémie et la montée corrélative de l'intérêt pour le masque. Avec un modèle de topics structurels, on confirme de manière plus qualitative et longitudinale cette hypothèse de transformation structurelle des représentations. Enfin, en employant une technique de réseaux sémantiques, on montre l'évolution qualitative de la question du masque et on confirme la centralité croissante de ce terme.

4.1 Du Coronavirus à la Covid-19, l'endogénéisation de l'épidémie

À partir des fréquences d'occurrences d'une série de termes représentatifs des débats et des sujets d'intérêts, on calcule une densité quotidienne, avec un lissage sur 7 jours. L'évolution temporelle de la fréquence de ces termes est représentée dans la figure 3.

L'évolution est d'une clarté lumineuse. Si de janvier à février le coronavirus était la star (bien noire), la Covid-19 prend le dessus courant mars et reste en tête pour tout le reste de la période. Autrement dit, ce sont les conséquences de la diffusion épidémique du coronavirus qui dominent les débats. La maladie pèse directement par la pression mise sur la santé publique et par ses victimes, mais aussi par les mesures qu'elle suscite pour en atténuer son impact. En ce sens, la Covid-19 est l'être social et politique du Coronavirus, cet être biologique. Le changement de terme marque un changement de discours : la menace qui était extérieure est rapidement endogénéisée : elle devient moins le virus que les perturbations qu'il génère : l'excès de mortalité, l'ébranlement du système de santé, la redéfinition des relations sociales, le choc économique, le questionnement sur les institutions. Le moment du confinement est un basculement de perspective du dehors vers le dedans.

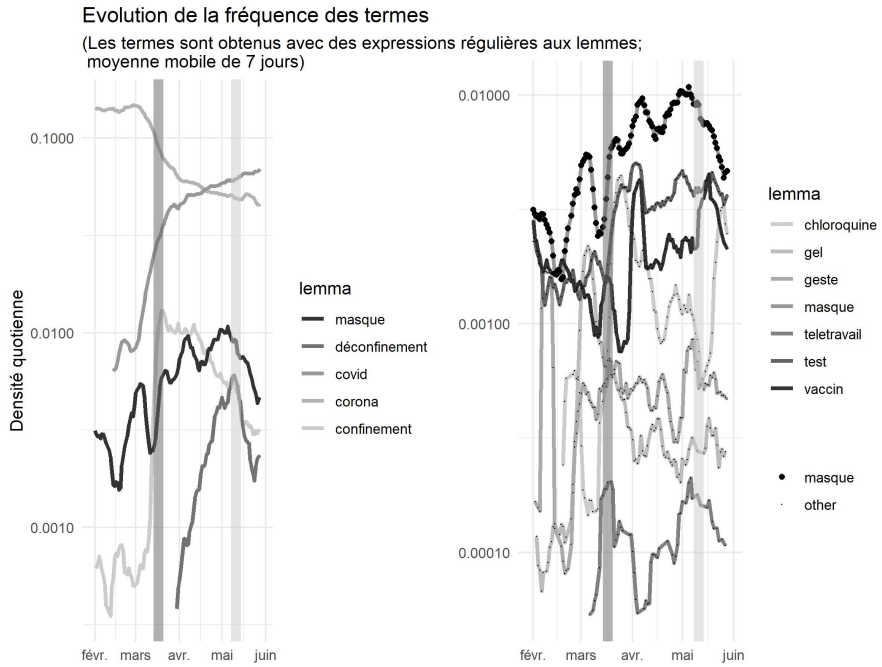


FIG. 2 – *Evolution de la densité des termes de geste barrière dans le corpus*

Dans le registre des questions de santé, on note une substitution parallèle à celle du Corona/Covid entre l'hospitalier et le sanitaire. Si la réponse au virus a été médicale, celle à la maladie qui touche le corps social devient sanitaire. L'impact du désastre venu d'ailleurs s'accompagne d'un changement de perspective, et dans ce renversement, le masque devient une préoccupation de plus en plus fréquente.

On confirme ce résultat avec l'analyse des thématiques(modèle STM). Une solution satisfaisante du modèle semble supporter 20 thématiques, dont on peut difficilement donner ici une analyse exhaustive, mais plus facilement une présentation synthétique à partir de la représentation la plus visuelle que fournit le modèle. La figure 6 en fournit la synthèse. Les topics sont d'autant plus proches qu'ils sont corrélés, les corrélations (>.2) sont représentées par l'épaisseur des segments et la taille des cercles est proportionnelle à leur fréquence. Cette structure se caractérise par une sorte de dualité, un macro segment semble s'articuler autour de deux composants. L'un est centré sur le coronavirus et la Chine, on le caractérisera d'exotique ; l'autre sur la Covid et la question publique. On retrouve cette idée d'un basculement de perspective. Ce qui était un corps étranger, devient une douleur intérieure.

L'avantage du modèle est de permettre de représenter la prévalence temporelle, qui est indiquée dans la figure 4. On identifie clairement les topics favorisés en première période puis qui déclinent, tout autant que ceux qui montent en deuxième période.

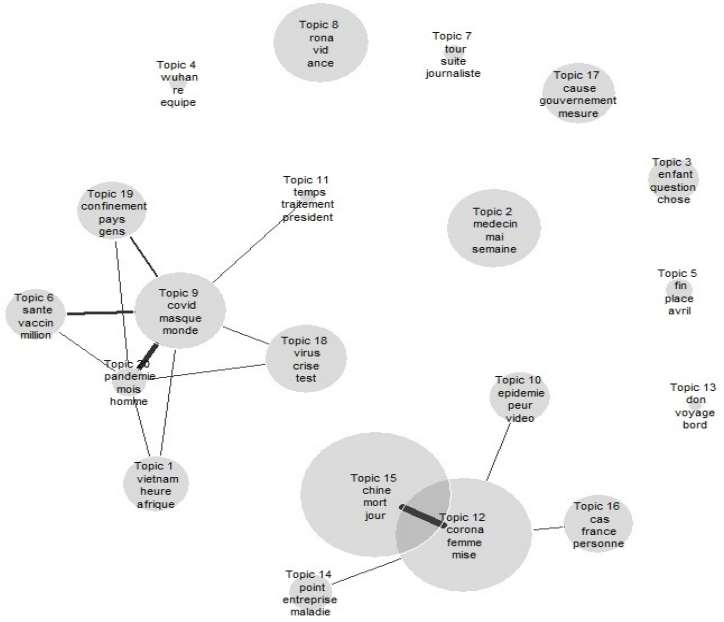


FIG. 3 – *Réseaux des topics en fonction de leur corrélations.*

Le début de la période étudiée se caractérise par la prévalence des questions liées à la découverte du coronavirus, de ses effets visibles à travers le nombre de morts en Chine, et à la situation de la France face à un virus inédit. On voit ensuite émerger de nouvelles thématiques, en lien avec la Covid-19 : le développement de la pandémie, le cas des enfants, les annonces concernant de futurs vaccins et les traitements potentiels (avec au cœur la chloroquine, évidemment), pour finir par l'état de crise dans le pays et sa gestion par le pouvoir politique. Le mouvement général est que les discussions évoluent vers une endogénéisation de l'épidémie, qui commence par un virus inconnu de Wuhan pour se transformer en une maladie qui dévaste la France et les Français (l'économique et le médical).

Quant au masque, l'analyse de ses occurrences montre une montée progressive, par vagues successives qui concernent sans doute ses différentes polémiques et sa domination en terme quantitatif sur les termes qui évoquent les autres méthodes d'atténuation : le test, le gel, les gestes barrières, le télétravail pour en mentionner les plus significatifs. À travers l'analyse des topics se précise un schéma : on ne retrouve pas le thème du masque dans un grand nombre de thématiques, mais dans une thématique centrale qui s'articule étroitement à la maladie. Ce topic est lui-même associé à quelques autres, et cette configuration définit le nouveau paradigme du discours qui se noue dans le premier confinement.

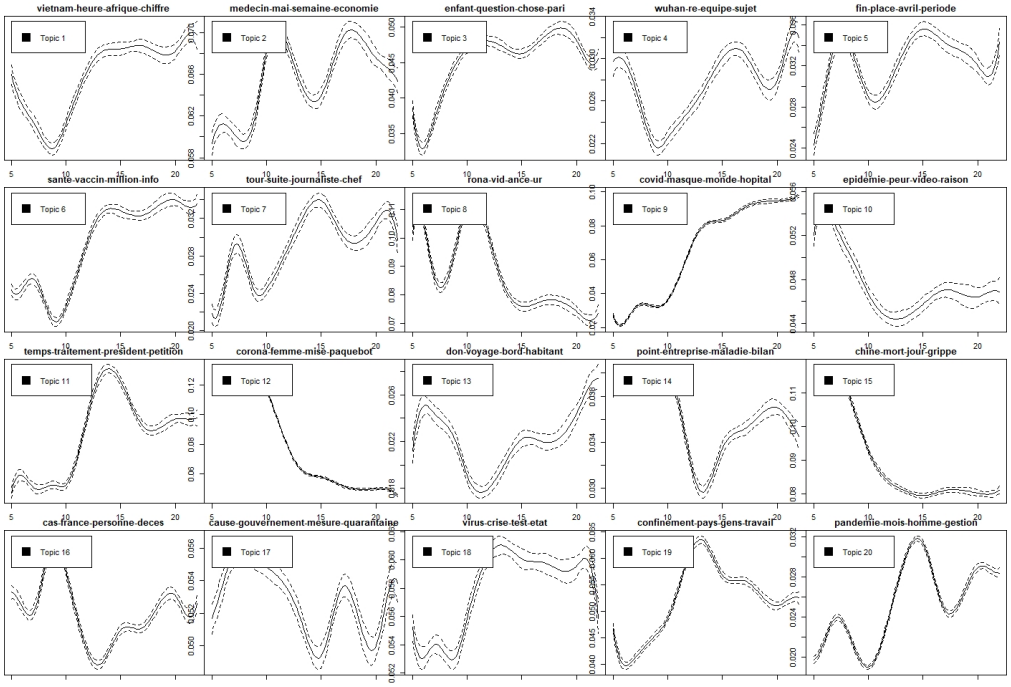


FIG. 1 – Prévalence des topics au cours du temps

4.2 Le masque au centre des débats

Dit-on les mêmes choses du masque aux différentes phases de l'épisode épidémique ? À cette fin on utilise les annotations de dépendance syntaxique, pour identifier quels sont les termes associés grammaticalement au nom commun "masque" : des adverbes, des adjectifs, et d'autres substantifs. On compare les plus fréquents pour chacun des 4 mois d'observation. Pour chacun des termes obtenus on calcule leur densité et on obtient le spectre des significations, représenté sur la figure 5.

Les résultats sont clairs : 1) le masque protège, c'est trivial ; 2) sa forme est le chirurgical plutôt que FFP2, en dépit d'une hésitation courant mars ; 3) sa distribution et sa commercialisation deviennent plus importantes avec le temps. Mais l'essentiel est dans l'association à "obligatoire". Le masque fût un temps, un moyen, il est en mai 2020, une norme. Cette norme est autant juridique que sociale. En effet, le masque a été rendu obligatoire progressivement, d'abord dans les transports, à l'école puis à l'initiative des maires et des organisations privées, mais il joue aussi par l'invisible pression du regard d'autrui, auquel pourtant l'on échappe, et les croyances construites quant à son efficacité. Avec le temps, son sens s'enrichit, même si en mai, il semble s'affiner, les traits mineurs ayant une fréquence moindre. Dans cette évolution on retrouve l'idée d'une flexibilité interprétationnelle : au

cours du confinement, l'idée de masque sans cesser d'être sociale change de signification. De l'objet fonctionnel qui manque et dont on se pose la question de l'efficacité relative, on passe au registre de l'obligation, du sentiment d'un impératif, le masque devient une norme au-delà de sa signification évidente : la protection. On est tenté de voir dans ce résultat une confirmation sur l'hypothèse d'un basculement de la nature de la convention : ce qui s'impose dans un débat à propos de l'efficacité en raison de l'importance du moyen, devient ce qui possède un caractère obligatoire, s'imposant de lui-même, comme évidence.

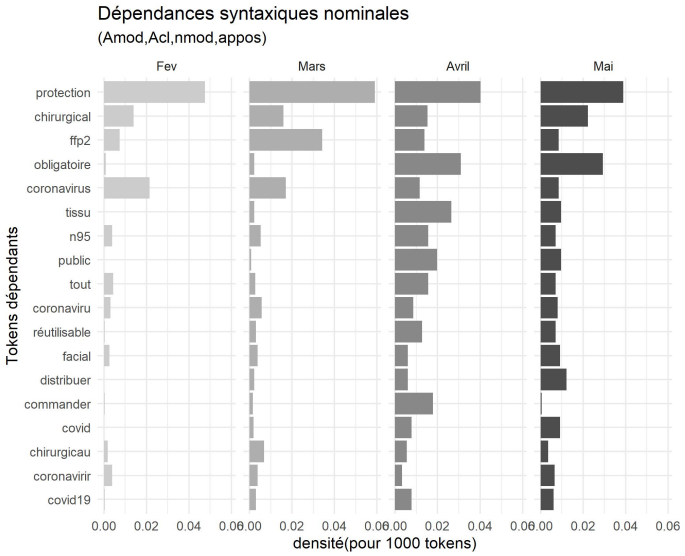


FIG. 5 – Analyse des dépendances syntaxique au cours du temps.

Pour rendre compte des discussions et de leurs évolutions, on utilise une méthode de cartes sémantiques, calculées pour chacune des périodes (22 semaines consécutives). On représente les mots dont les cooccurrences sont supérieures à une fréquence déterminée en fonction du corpus de chacune des périodes et correspondant à une proportion constante à travers les périodes (environ 30%). La taille des nœuds correspond à leur densité dans le corpus, celle des arcs à la fréquence de la cooccurrence. Les positions relatives dans le plan sont calculées en fonction de la similarité des termes.

Une problématique centrale est manifestée par un macro-composant du réseau de termes, des thématiques plus spécifiques, déconnectées et positionnées en périphérie structurent l'espace du discours. La figure 6 en donne une illustration pour le jour des élections alors que le premier confinement s'engageait. Les conversations périphériques traitent de sujets divers : le système de santé, la question pratique de l'attestation de déplacement, la fuite des gens des villes à la campagne, le décès de Manu Dibango, et la pétition adressée au chef de l'État pour l'emploi de la chloroquine. Le macro-composant, quant à lui, se structure de clairement. Sur un axe presque horizontal, le confinement pose la question du « pour combien de temps » ? À l'opposé se dessine nettement la crise de la pénurie de masques, en particulier, pour les

soignants. Au centre, la France et le gouvernement font le pont entre les sujets. Sur le sud du composant, on retrouve la question de l'état d'urgence, au nord le thème de l'ampleur de la crise traduite en nombre de morts quotidien.

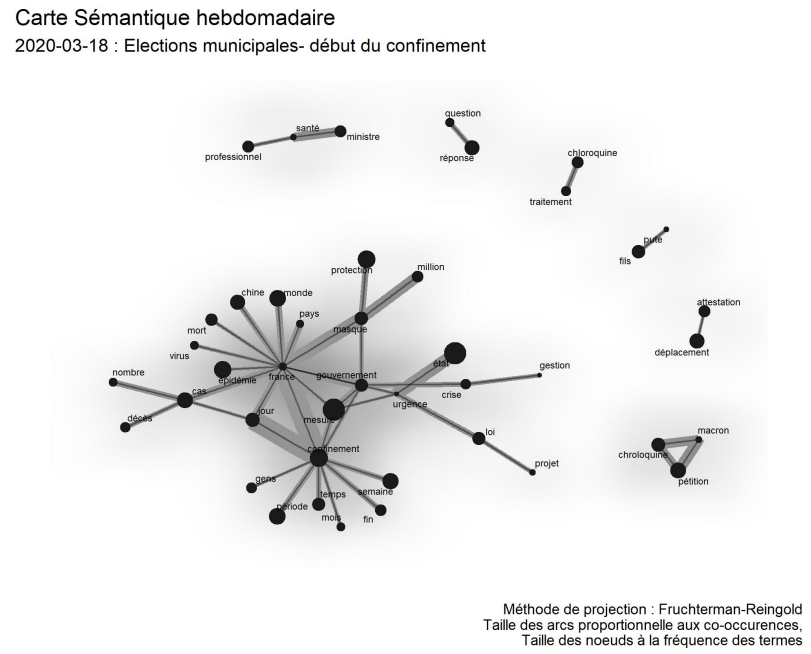


FIG. 6 – Réseau sémantique (exemple de la semaine du 18 mars 2020)

En répétant cette analyse interprétative sur l'ensemble des 22 semaines, et donc 22 cartes, on peut reconstruire schématiquement l'évolution des discours produits sur la thématique Coronavirus/Covid. Pour s'assurer de cette intuition ou hypothèse d'une évolution au cours de la période de l'objet masque dans le discours du média social, un test de centralité peut être engagé, il permet de systématiser l'analyse. Trois indicateurs sont choisis (cf. figure 7) : le degré de centralité (degree) ; la centralité d'intermédiarité (betweenness) , la centralité de proximité (closeness).

La tendance générale va à la hausse de la centralité. le masque est connecté à un nombre plus grand de conversations. Devenant un plus petit dénominateur commun, il devient la clé principale par laquelle on peut accéder aux différents cheminements de la pensée collective. Observons-nous là l'installation dans le discours d'une convention ? Naturellement la fréquence de citation du masque, plus haute que celle des autres moyens d'atténuation, favorise le fait qu'il soit associé plus fréquemment à une plus grande variété d'objets. C'est le point de vue statistique. On peut envisager aussi l'hypothèse qu'étant associé à plus de thématiques, il soit cité plus souvent.

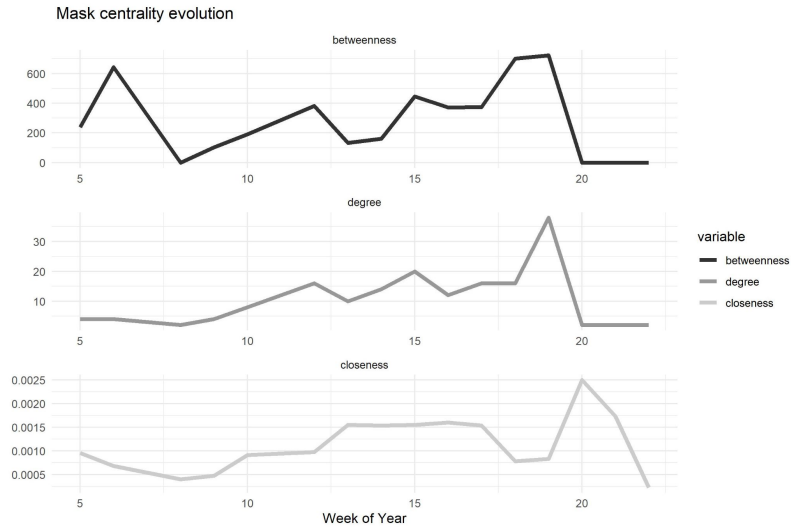


FIG. 7 – Evolution des indicateurs de centralité du masque.

5. Conclusion

Notre étude se limite à l'analyse d'un discours dont il faut rappeler la dimension fragmentaire. Twitter ne diffuse pas des textes mais des bouts de phrases. Le discours qu'il produit est peu articulé, peu développé, il est un flux de mots et de leurs associations. Nous avons essayé d'en quantifier les flux et nous en obtenons des résultats factuels qui méritent considération.

La perspective théorique de cette recherche se tient dans une perspective d'anthropologie matérielle qui précise et souligne l'importance des objets matériels dans l'élaboration des interaction sociale. Les résultats soulignent la nécessité de considérer les moyens de l'atténuation (gestes barrières, distance sociale, utilisation des artefacts …) non seulement sur le plan de leur efficacité intrinsèque et du respect de l'observance, mais aussi dans une perspective plus anthropologique qui donne aux objets matériel de la vie (extra)ordinaire un pouvoir nourri par leur capacité à ritualiser les interactions sociales, à porter de manière symbolique l'engagement des acteurs et plus encore à leur donner un pouvoir sur le mal invisible de l'épidémie. Même s'il ne protège guère, le masque est efficace comme l'est le fétiche, il maintient un ordre social quand on ne sait rien des batailles qui se produisent. Pour obtenir la participation des populations les arguments rationnels ne sont peut être pas suffisants.

Sur le plan empirique, le résultat principal est que le masque émerge comme figure polaire du discours dès le mois d'avril ce qui converge avec des données d'enquête plus conventionnelles (ie : CoviPrev). C'est le terme qui articule les polémiques (pénurie de protection pour les agents de santé, problématique de l'approvisionnement, diffusion de l'obli-

gation de son port) tout en restant le fil conducteur de la conversation alors que sur la période son obligation est limitée, et l'observance de son port est largement volontaire. Ce résultat n'était pas évident de prime abord, d'autres figures étaient candidates : la visioconférence dont des millions de travailleurs et d'étudiants ou de professeurs ont appris rapidement à manipuler les interfaces mais qui n'apparaît pas dans les contenus. Les tests surtout, dont l'utilité a été mise en question avant de devenir un argument central de l'action gouvernementale et un facteur de réussite dans certains pays asiatiques, où simplement le vaccins.

Le masque vaut moins par ses qualités fonctionnelles que par sa capacité à fixer l'attention et à organiser les conditions de vie sous la menace épidémique. L'hypothèse défendue est que son ambivalence phénoménologique nourrit une flexibilité d'interprétation suffisante pour articuler des mondes sociaux mal accordés (le savant, le politique et le citoyen), et donner un cadre commun où peuvent se renégocier les conventions et les normes de la vie matérielle, il forme paradoxalement le tissu des relations sociales, le nœud matériel, rituel et symbolique par lequel l'activité sociale se réorganise dans la catastrophe.

Références

Arnold, Taylor. 2017. « A Tidy Data Model for Natural Language Processing Using CleanNLP ». The R Journal 9 (2): 248.

Banda, Juan M., Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, et Gerardo Chowell. 2020. « A large-scale COVID-19 Twitter chatter dataset for open scientific research -- an international collaboration ».

Blei, David M., Andrew Y. Ng, et Michael I. Jordan. 2003. « Latent Dirichlet Allocation ». J. Mach. Learn. Res. 3 (mars): 993–1022.

Cambria, Erik, et Bebo White. 2014. « Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article] ». IEEE Computational Intelligence Magazine 9 (2): 48-57.

Humphreys, Ashlee, et Rebecca Jen-Hui Wang. 2018. « Automated Text Analysis for Consumer Research ». Édité par Eileen Fischer et Linda Price. Journal of Consumer Research 44 (6): 1274-1306.

Kobayashi, Vladimer B., Stefan T. Mol, Hannah A. Berkers, Gábor Kismihók, et Deanne N. Den Hartog. 2018. « Text Mining in Organizational Research ». Organizational Research Methods 21 (3): 733–765.

Kozlowski, Austin C., Matt Taddy, et James A. Evans. 2019. « The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings ». American Sociological Review 84 (5): 905-49.

Leigh Star, Susan. 2010. « Ceci n'est pas un objet-frontière !: Réflexions sur l'origine d'un concept ». Revue d'anthropologie des connaissances Vol 4, 1 (1): 18.

Lemonnier, Pierre. 2012. Mundane objects: materiality and non-verbal communication. Critical cultural heritage series. Walnut Creek, CA: Left Coast Press.

Lock, Irina, et Peter Seele. 2015. « Quantitative Content Analysis as a Method for Business Ethics Research ». Business Ethics: A European Review 24 (juillet): S24-40.

Mitroff, Ian I. 1986. « TEACHING CORPORATE AMERICA TO THINK ABOUT CRISIS PREVENTION ». Journal of Business Strategy 6 (4): 40-47.

Noji, Eric K. 2005. « Disasters: Introduction and State of the Art ». Epidemiologic Reviews 27 (1): 3–8. https://doi.org/10.1093/epirev/mxi007.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, et David G. Rand. 2014. « Structural Topic Models for Open-Ended Survey Responses: ». American Journal of Political Science 58 (4):

Rodríguez, Havidán, E. L. Quarantelli, et Russell Rowe Dynes, éd. 2007. Handbook of disaster research. Handbooks of sociology and social research. New York: Springer.

Star, Susan Leigh, et James R. Griesemer. 1989. « Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39 ». Social Studies of Science 19 (3): 387-420.

Steyer, Véronique. 2020. Le masque de 2004 à 2020 : fausse bonne-idée ou arme absolue ? LE VIRUS DE LA RECHERCHE, JUIN 2020. Pug.

Trompette, Pascale, et Dominique Vinck. 2009. « Retour sur la notion d'objet-frontière ». Revue d'anthropologie des connaissances 3, 1 (1): 5.

Vinck, Dominique. 2009. « De l'objet intermédiaire à l'objet-frontière: Vers la prise en compte du travail d'équipement ». Revue d'anthropologie des connaissances 3, 1 (1): 51.

Summary

The Covid-19 pandemic that has hit the world offers a spectacular case study in disaster management. In this literature, the participatory paradigm is fundamental: the mitigation of the impact of the disaster, the quality of the preparation and the resilience of the society, facilitate the reconstruction, but depend on the participation of the populations. Being able to observe and measure the mental health of the population and the content of the discourse are necessary to accompany measures to encourage this participation. Social media, and in particular Twitter, offer valuable resources for exploring this discourse. The main result is based on the identification of the centrality of the mask figure and aims to establish the importance of the phenomenon. We show this in a quantitative way using NLP methods. We exploit here a database of 500k tweets extracted from a corpus over the period from February to the end of May 2021.

Keywords: Covid-19, disaster management, NLP, STM, social media.

Index