Actes de la 26e Conférence Nationale en Intelligence Artificielle

Sandra Bringay

AfIA

Association française
pour l'Intelligence Artificielle

CNIA

Conférence Nationale en Intelligence Artificielle

PFIA 2023

Table des matières

Éditorial

Conférence Nationale en Intelligence Artificielle

La Conférence Nationale en Intelligence Artificielle (CNIA), soutenue par le Conseil d'Administration de l'AFIA, est hébergée par la plateforme PFIA, conjointement avec d'autres conférences francophones dans le domaine de l'intelligence artificielle (IA). Elle s'adresse à l'ensemble de la communauté de recherche en IA. CNIA se veut un lieu privilégié pour faire connaître les dernières avancées en IA. Elle se veut aussi un forum destiné à renforcer les liens et les interactions entre les différentes sous-disciplines de l'IA et les disciplines faisant appel à l'IA. À ce titre, CNIA encourage les soumissions à la frontière entre sous-branches de l'IA, ainsi que les soumissions à la frontière de l'IA et d'autres disciplines.

Alors que l'IA se trouve aujourd'hui au cœur de nombreux développements, il est important d'avoir un forum qui réunisse l'ensemble des acteurs intéressés de près ou de loin par l'IA. L'objectif de CNIA est d'aborder à la fois les problématiques de recherche, les enjeux technologiques et les enjeux sociétaux liés à l'utilisation de l'IA, à travers l'ensemble des disciplines de l'IA :

— recherche heuristique et résolution de problèmes,
— incertitude et intelligence artificielle,
— logique, satisfiabilité et satisfaction de contraintes,
— apprentissage automatique,
— extraction, ingénierie et gestion des connaissances,
— représentation des connaissances et raisonnement,
— planification, contrôle,
— aide à la décision,
— causalité,
— agents autonomes et systèmes multi-agents,
— reconnaissance des formes et vision par ordinateur,
— traitement automatique des langues naturelles et de la parole, recherche d'information,
— interactions avec l'humain,
— perception et robotique,
— environnements informatiques d'apprentissage humain et apprentissage à distance,
— IA responsable, IA de confiance (incluant explicabilité, certification, équité, ...),
— éthique de IA, droit et IA,
— IA & X (X= société, web, santé, environnement, énergie, transport, défense, agriculture, matériaux, ...),
— ...

Suite à l'appel à contributions, la conférence CNIA a reçu 17 soumissions d'articles de travaux originaux, de travaux déjà publiés dans une conférence ou revue internationale de renom et de posters/démonstrations. Grâce au travail conséquent des membres du comité de programme, chaque article a reçu entre 3 et 4 relectures comportant des critiques argumentées et constructives pour les auteurs. Sur la base de ces critiques, 13 articles ont été retenus pour une présentation longue de 30m : 9 articles longs et 3 articles déjà publiés et 1 poster. Le programme de la conférence réparti sur 3 jours est organisé en 9 sessions dont le contenu est détaillé dans ces actes.

Pour cette édition 2023 de la conférence, nous avons l'honneur de recevoir deux invités :

— Leman Akoglu (Carnegie Mellon University, États-Unis) interviendra le Lundi 3 juillet pour une présentation intitulée "Automating Unsupervised Learning" dans le cadre des conférenciers invités à la PFIA ;
— Gauvain Bourgne (Sorbonne Université, LIP6, Paris) interviendra le Mardi 4 juillet pour une présentation intitulée " Éthique computationnelle et Causalité" comme conférencier invité de CNIA ;

Dans cette édition de CNIA 2023, les thèmes abordés par les auteurs couvriront différents aspects de l'IA. Ainsi, les approches décrites s'intéressent à des thèmes variés comme les règles, trajectoires et IA neuro-symbolique, les motifs et leur sémantique, l'explicabilité et l'équité des méthodes, les méthodes d'apprentissage....

Il nous reste à remercier chaleureusement l'ensemble des acteurs de la communauté francophone en IA qui ont contribué au succès de CNIA 2023, ainsi que le comité d'organisation de la plateforme PFIA 2023 qui a été d'une aide précieuse.

Sandra Bringay

Comité de programme

Présidence

— Sandra Bringay, Université Paul-Valéry Montpellier.

Membres

— Jérome Azé, Université de Montpellier
— Isabelle Bloch, Sorbonne Université
— Olivier Boissier, Mines Saint-Etienne, LIMOS
— Grégory Bonnet, Université de Caen Normandie
— Robert Bossy, INRAE Centre de Jouy en Josas, MaIAGE
— Armelle Brun, Université de Lorraine
— Sylvie Coste-Marquis, Université d'Artois
— Benjamin Dalmas, Centre de Recherche Informatique de Montréal
— Yves Demazeau, CNRS, LIG
— Arnaud Doniec, IMT Lille Douai
— Jean-Gabriel Ganascia, Sorbonne Université
— Eric Gaussier, Université Grenoble Alpes
— Guillaume Gravier, CNRS, IRISA
— Andreas Herzig, CNRS, IRIT
— Nathalie Hernandez, Université Toulouse Jean Jaurès
— Camille Kurtz, Université Paris Cité
— Nicolas Lachiche, Université de Strasbourg
— Frederique Laforest, INSA Lyon
— Florence Le Ber, École Nationale du Génie de l'Eau et de l'Environnement de Strasbourg
— Philippe Lenca, IMT Atlantique
— Marie-Jeanne Lesot, Sorbonne Université
— Pascal Poncelet, Université de Montpellier
— Catherine Roussey, INRAE Centre Occitanie-Montpellier, MISTEA
— Ana Roxin, Université Bourgogne Franche-Comté
— Nicolas Sabouret, Université Paris-Saclay
— Pascale Sébillot, INSA Rennes
— Nazha Selmaoui, Université de la Nouvelle Calédonie
— Catherine Tessier, ONERA
— Laurent Vercouter, INSA Rouen Normandie
— Bruno Zanuttini, Université de Caen Normandie

Session 1 : Règles, trajectoires et IA neuro-symbolique

IA neuro-symbolique pour l'interprétation granulaire des données des bases STUPS© et OTARIES© : défis applicatifs et perspectives

S. Guillemin[1], L. Dujourdy[2], L. Journaux[3], A. Roxin[1]

[1]Université de Bourgogne, Laboratoire d'Informatique de Bourgogne (LIB) EA 7534.
[2]Institut Agro Dijon – Cellule d'Appui à la recherche en science des données.
[3]Institut Agro Dijon, Laboratoire d'Informatique de Bourgogne (LIB) EA 7534.

sebastien.guillemin@u-bourgogne.fr, laurence.dujourdy@agro-dijon.fr, ludovic.journaux@agro-dijon.fr, ana-maria.roxin@u-bourgogne.fr

Résumé

Historiquement, l'intelligence artificielle (IA) s'est divisée en 2 courants selon les hypothèses faites pour modéliser l'intelligence humaine : l'IA symbolique, supposant que des symboles sont nécessaires, et l'IA statistique (plus particulièrement l'IA connexionniste) affirmant le contraire. Dernièrement, l'IA neuro-symbolique tente de réconcilier les 2. Les travaux présentés dans cet article sont en en collaboration avec la Police Scientifique française, dans le contexte du Plan National Stup. Nous présentons les problématiques métiers en lien avec le projet, puis en déduisons les problématiques scientifiques. Après un rappel des domaines de l'IA et leurs limites, nous présentons un état des lieux du domaine de l'IA neuro-symbolique. Nous discutons les défis applicatifs avant de finir par une présentation des premiers travaux qui ont été conduits. Des pistes d'approches basées sur l'état des lieux de l'IA neuro-symbolique sont aussi présentées pour la suite des travaux.

Mots-clés

Intelligence artificielle, IA neuro-symbolique, IA symbolique, IA connexionniste, aide à la décision.

Abstract

Historically, artificial intelligence (AI) has been divided into two streams based on the assumptions made to model human intelligence: symbolic AI, which assumes that symbols are necessary, and statistical AI (more specifically, connectionist AI), which asserts the opposite. Recently, neuro-symbolic AI has attempted to reconcile the two. The work presented here is in collaboration with the French forensic police, in the context of the Stup National Plan. We present the domain-related issues involved, and then the related scientific issues. After reviewing the above-mentioned sub-domains of AI along with their limitations, we present an overview of neuro-symbolic AI. We discuss the application challenges before ending with a presentation of the first work that has been conducted. Possible approaches based on the state of the art of neuro-symbolic AI are also presented for further work.

Keywords

Artificial intelligence, Symbolic AI, Connectionist AI, Neuro-symbolic AI, decision support.

1 Introduction

La lutte contre le trafic de drogue est une priorité pour le gouvernement français. Le ministère de l'intérieur a érigé cette lutte en l'une de ses trois priorités dès juillet 2020 [31]. Le Plan Stup français, publié en septembre 2019, prévoit 55 mesures, parmi lesquelles l'utilisation de nouveaux indicateurs pour mieux comprendre les usages des consommateurs et les méthodes des trafiquants.

Nos travaux se déroulent dans le cadre d'une thèse financée par le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI) visant à combiner des techniques de l'intelligence artificielle (IA) symbolique et de l'IA connexionniste. Ces travaux concernent le domaine de l'identification des filières de distribution de stupéfiants, plus précisément le rapprochement (ou appariement) automatique entre échantillons et la prédiction de tendances de dosage de principes actifs et de produits de coupage. Ils seront appliqués aux données contenues dans les bases STUPS© et OTARIES©. Pour en encadrer l'accès et l'usage, le projet de collaboration AI4NP (*Artificial Intelligence for Narcotic Prediction*) a été signé. Ce projet réunit le Laboratoire d'Informatique de Bourgogne (LIB), l'Institut Agro Dijon et le Service National de la Police Scientifique (SNPS).

La base de données nationale STUPS© (Système de Traitement Uniformisé des Produits Stupéfiants) du Ministère de l'Intérieur recueille des informations sur les drogues illicites circulant en France, incluant des données macroscopiques (e.g. logos de distributeurs de drogue, dimensions des produits etc.), qualitatives (e.g. noms des composants d'un échantillon de drogue) et quantitatives (e.g. dosages) ainsi que des données d'enquêtes non confidentielles (e.g. date, lieu de saisie). Cette base a été créée en 1986 et contient environ 10 millions d'entrées provenant des laboratoires de Police Scientifique du SNPS et de l'Institut de Recherche Criminelle de la

Gendarmerie Nationale. Cette base sert aux experts pour consulter des informations sur des échantillons analysés principalement pour faire le lien entre des affaires hors profilage chimique.
La base de données OTARIES© (Outil de Traitement Automatisé pour le Rapprochement InterEchantillons de Stupéfiants), créée en 2001, contient les profils chimiques des échantillons de cocaïne et d'héroïne analysés. Une méthode de profilage chimique exploite les informations issues de ces profils pour effectuer des rapprochements entre échantillons.
Cet article est structuré comme suit : la section 2 présente le contexte et les problématiques métier, par rapport auxquelles sont identifiées les problématiques scientifiques, exposées dans la section 3. La section 4 rappelle des définitions des domaines de l'IA symbolique et connexionniste ainsi que leurs limites, justifiant le choix de l'IA neuro-symbolique dont la section 5 présente un état de l'art. La section 6 discute des défis applicatifs du projet et la section 7 présente les premiers travaux ainsi que des pistes de futures approches basées sur les notions introduites dans la section 5.

2 Contexte et problématiques métiers

Lors d'une saisie de substances supposées illicites, des échantillons sont prélevés et analysés afin de déterminer le profil physico-chimique de la substance. L'analyse d'un échantillon s'effectue en 3 étapes, selon le processus illustré dans la Figure 1.

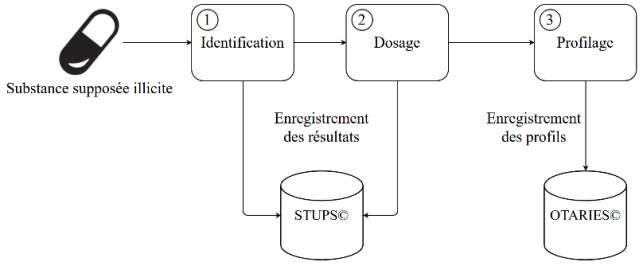


Figure 1. Processus d'analyse d'un échantillon de substance supposée illicite.

L'**identification** est la première étape. L'objectif de cette étape qualitative est de déterminer le classement des éventuelles substances illicites et d'identifier les produits de coupage (avec ou sans effet psychoactif) contenus dans l'échantillon. En complément du processus analytique pour l'identification de substance, est aussi réalisé un examen macroscopique afin de répertorier les caractéristiques physiques de l'échantillon e.g. la forme, la couleur, le logo du distributeur, etc. Cette première étape permet de statuer sur la mise en cause ou non d'une personne pour Infraction liée à la Législation sur les Stupéfiants (ILS).
La seconde étape est quantitative et vise le **dosage** du ou des principes actifs et des principaux produits de coupage identifiés à l'étape précédente. Cette étape fournit des données fiables sur la pureté des produits circulant sur le territoire national à l'ensemble des acteurs impliqués dans la prévention ou dans la répression du trafic de stupéfiants. Dans certains cas, ces données permettent de renseigner le niveau de trafic auquel appartient la drogue e.g. rue, grossiste ou laboratoire. Ces données sont utilisées à des fins statistiques par les services sociaux et sanitaires e.g. OFDT (Observatoires Français des Drogues et des Tendances addictives), EMCCDA (European Monitoring Centre for Drugs and Drug Addiction), et répressifs e.g. OFAST (Office antistupéfiants), Mission de lutte antidrogue (MILAD).
Afin de réduire le temps d'obtention des résultats, ces deux premières étapes sont souvent fusionnées en une seule pour la plupart des laboratoires nationaux. Les résultats obtenus au cours de ces étapes sont enregistrés dans la base de données STUPS©.
La troisième et dernière étape est l'étape de **profilage**. Pour un stupéfiant donné, il s'agit d'extraire les composés cibles du produit, afin d'en établir le profil chimique pour permettre le rapprochement entre saisies (notamment au travers de comparaisons de chromatogrammes de différents échantillons [11]). Plus particulièrement, deux types de profils peuvent être établis :

- Un *profil chimique* contenant les caractéristiques chimiques des composants e.g. impuretés (in)organiques, solvants résiduels, produits de coupage,
- Un *profil physique* contenant les caractéristiques macroscopiques des échantillons e.g. leur couleur, forme, emballages.

Une métrique de ressemblance entre deux échantillons est déterminée par la Police Scientifique et s'exprime en pourcentage (0% pour des échantillons totalement différents et 100% pour des échantillons identiques). Un seuil fixe le pourcentage minimal à partir duquel deux échantillons sont considérés proches. Dans ce cas, ils sont dits « liés ». Les données de profilage chimique sont enregistrées dans la base de données OTARIES©, qui intègre un algorithme de comparaison et les liens entre échantillons sont signalés ([9] et [11]). Le profilage physique est quant à lui réalisé à l'aide des données stockées dans STUPS©.
Bien que le processus décrit ci-dessus soit en place et fonctionnel, certains problèmes subsistent. Tout d'abord, concernant la base de données STUPS© très peu d'études rétrospectives ont été conduites jusqu'à présent (la dernière [10] date de 2017), principalement parce que les données antérieures à la création de la base STUPS© (1986) ne sont disponibles que sous forme manuscrite. De plus, aucune étude prospective utilisant des analyses prédictives n'a été conduite jusqu'à présent. Ces études pourraient porter, par exemple, sur l'évolution du dosage d'un produit de coupage ou l'évolution d'un réseau de distribution sur une période future et une zone géographique données. Les résultats de telles études pourraient permettre la prise de mesures de prévention proactives.
Ensuite, concernant la base de données OTARIES©, à l'instar de la base STUPS©, aucune étude prospective n'a été menée pour identifier des tendances temporelles ou géographiques sur les profils observés, l'apparition ponctuelle de profils atypiques, ou encore l'existence de catégories par origine ou recette employée. De plus, lorsque deux échantillons sont proches, selon la métrique de la Police Scientifique, les experts vérifient manuellement qu'il n'y a pas d'aberrations sur les échantillons qui réfuteraient le rapprochement. Par exemple, si la différence entre les dates de saisie des deux échantillons est supérieure à une certaine durée, les experts peuvent conclure que les échantillons ne sont en fait pas proches. Cette démarche est alors une démarche complexe, difficile à harmoniser et nécessite l'intervention d'experts métier.
D'après le contexte présenté ci-dessus, les Problématiques

Métiers (PM) suivantes ont été identifiées :

- **PM1** : Mener une étude approfondie des bases STUPS© et OTAIRES© pour en extraire une base de connaissances.
- **PM2** : Établir des modèles exploitant la base de connaissances afin de rapprocher automatiquement des profils selon les caractéristiques macroscopiques et les compositions des échantillons.
- **PM3** : Établir des modèles prédictifs, robustes statistiquement et exploitant la base de connaissances afin de caractériser au mieux les produits stupéfiants circulant en France.
- **PM4** : Comparer des compositions de produits par années, niveaux de trafic, territoires etc.

Les résultats fournis par les modèles utilisés doivent être explicables afin de servir de preuves lors d'un procès. Cela signifie qu'un expert doit pouvoir comprendre les causes amenant aux résultats d'un algorithme d'aide à la décision.

3 Problématiques scientifiques

Plusieurs problématiques scientifiques (PS) découlent des objectifs visés et des problématiques métiers du projet AI4NP, présentées dans la section précédente.

La première problématique (**PS1**) porte sur *la conception d'une base de connaissances réunissant les données et informations contenues dans les bases de données STUPS© et OTARIES©*. Cette problématique scientifique est liée à la problématique métier PM1. Il s'agira d'étudier comment les approches d'IA symbolique permettent de représenter ces éléments de manière explicite et formelle.

Les problématiques métiers PM2 et PM3 soulèvent la seconde problématique scientifique (**PS2**), qui concerne *le couplage d'approches d'IA symbolique et d'IA connexionniste*. Plus précisément, il s'agit d'investiguer comment les résultats produits par des modèles d'IA connexionniste peuvent être validés et contrôlés, en utilisant les connaissances des experts. Cela permettrait de fournir des résultats conformes aux connaissances des experts, d'éviter les résultats aberrants et pourrait réduire le temps d'entraînement de ces modèles. De plus, les modèles d'IA connexionniste seraient utilisés pour enrichir la base de connaissances. En effet, la construction d'une base de connaissances est un processus itératif nécessitant des échanges avec des experts et la compréhension d'un domaine métier complexe. De ce fait, elle n'est pas complète dès la première itération. Les modèles d'IA connexionniste pourraient alors extraire des connaissances à partir des données pour faciliter le processus de création de la base de connaissances. De plus, ces nouvelles connaissances pourraient être utilisées par les experts pour identifier de nouveaux axes d'analyse. Cependant, ces connaissances doivent pouvoir être acceptées ou réfutées par les experts pour éviter de polluer la base de connaissances déjà existante.

Pour finir, la troisième et dernière problématique scientifique (**PS3**) découle de PM4 et se focalise sur *la prise en compte de différentes échelles d'analyse (ou niveaux de granularité) par les modèles d'IA connexionniste et d'IA symbolique*. L'objectif est d'étudier comment différents modèles d'IA peuvent gérer efficacement les changements d'échelle sans nécessiter de lourdes modifications telles qu'un réentraînement complet des modèles ou une nouvelle spécification de la base de connaissances. La prise en compte de plusieurs niveaux de granularité peut permettre d'améliorer les performances des différents modèles, en éliminant les informations non pertinentes pour un niveau spécifique. En considérant, par exemple, les niveaux de distribution de drogue il est possible de faire des analyses à l'échelle de la vente en gros (avant que la drogue ne soit altérée avec des produits de coupage) jusqu'à la vente au détail (où la drogue est le plus coupée). La représentation de la connaissance à ces différents niveaux ne serait pas la même car les acteurs et les échelles temporelles et spatiales entrants en jeu seraient différentes.

Ces trois problématiques relèvent des domaines de l'IA symbolique et connexionniste ainsi que l'informatique granulaire. La section suivante donne des rappels sur les notions d'IA symbolique et connexionniste et expose les limites de ces approches. L'informatique granulaire n'a pas encore été étudié dans le cadre du projet et n'est donc pas abordé dans cet article.

4 IA Symbolique et connexionniste

L'intelligence artificielle est le domaine ayant pour objectif de concevoir des techniques informatiques permettant de simuler l'intelligence humaine et dont l'origine remonte au milieu des années 40 [38]. Les problématiques de la section 3 font référence à deux courants de l'IA : l'IA symbolique et l'IA connexionniste. Cette section présente succinctement ces deux courants ainsi que leurs limites justifiant le choix d'une autre voie de recherches - l'IA neuro-symbolique - dont la section 5 dresse l'état des lieux.

4.1 IA Symbolique

L'IA symbolique a pour objectif de représenter et reproduire le raisonnement cognitif humain via un système de représentation des connaissances [18]. L'IA symbolique nécessite des représentations formelles et explicites d'un domaine de connaissances, couplées à des mécanismes permettant d'en déduire des connaissances implicites. Selon Studer et al [42], une ontologie est une « spécification formelle et explicite d'une conceptualisation partagée d'un domaine de connaissance ». La définition d'une ontologie se fait à l'aide de langages logiques.

La revue de l'ensemble des langages logiques sort du cadre de cet article mais l'on peut toutefois en noter deux principaux : la logique du premier ordre (FOL pour *First Order Logic*) [32] et les logiques de description (DL pour *Description Logics*)[2]. La FOL permet la modélisation de systèmes déductifs : ses règles de syntaxe permettent de définir une interprétation (une sémantique) d'une théorie (un ensemble de formules) et de déterminer si une formule est une conséquence logique d'une théorie. Toutefois elle n'est pas décidable : une procédure de décision (ou algorithme de décision) en FOL peut ne pas se terminer en un temps fini. Il existe alors les DL qui sont des sous-ensembles de FOL et sont (généralement) décidables. Un langage de la DL est caractérisé par un ensemble de constructeurs permettant d'enrichir l'expressivité du langage. Il faut noter que le gain d'expressivité d'un langage se fait au détriment du temps d'exécution des procédures de décision. Parmi les différents constructeurs on peut citer la conjonction, la négation, la disjonction ou encore le quantifieur existentiel. Les langages informatiques utilisés pour la définition d'ontologies reposent donc sur les logique de description [25]

car ils sont plus efficaces (en termes de temps d'exécution) des procédures de décision associées.
Une ontologie comprend une signature S (l'ensemble des classes, instances et propriétés utilisés pour représenter la connaissance), et un ensemble d'axiomes associés A (exprimé en DL par exemple) [17]. Généralement les termes « base de connaissances » et « ontologie » sont utilisés comme synonymes. Pour une base de connaissances, on fait la distinction entre *Terminological Box* (TBox – concepts et relations entre concepts) et la *Assertional Box* (ABox – instanciation des éléments de la Tbox). Pour une ontologie on distingue les entités conceptuelles (équivalent de TBox) et les entités concrètes (équivalent de ABox) [17]. Les procédures de décision peuvent alors être appliquées au niveau de la TBox (par exemple, pour inférer des relations entre concepts non explicitement spécifiées) ou au niveau de la ABox (par exemple, pour déduire l'appartenance d'un individu à une classe d'un concept). Les inférences produites par les algorithmes d'IA symboliques sont explicables au travers d'arbres de décision. Il est donc possible de connaître les causes amenant à un résultat.
Ainsi, l'IA symbolique est pertinente par rapport aux problématiques métiers, présentées ci-dessus, notamment pour modéliser la connaissance des experts métier et permettre l'explicabilité des déductions faites. Cette explicabilité est un point essentiel dans un contexte d'aide à la décision.

4.2 IA Connexionniste

Le second courant de l'IA présenté dans cette section est l'IA connexionniste (ou simplement connexionnisme) [38]. Apparu au milieu des années '80, ce courant se place en opposition à celui de l'IA symbolique [35] et réfute l'hypothèse selon laquelle le raisonnement humain ne peut être modélisé qu'en utilisant des symboles [38].
Les modèles connexionnistes n'utilisent pas de langage logique pour spécifier leurs concepts internes mais un ensemble de neurones formels connectés entre eux pour former un réseau de neurones [7]. Un neurone formel (ou simplement neurone) est une représentation mathématique d'un neurone biologique sans chercher à en être une copie conforme [23]. Un réseau de neurones est alors formé de neurones organisés en couches successives où les neurones d'une couche *n* sont connectés à ceux de la couche *n+1*. Ces réseaux de neurones ont la capacité, entre autres, d'apprendre à partir d'exemples, d'abstraire des données non structurées et d'être robustes face au bruit et aux aberrations dans les données [4], [21], [47]. Ils peuvent être utilisés en fouille de données pour identifier des patterns ou des corrélations dans un grand ensemble de données [33]. Chaque réseau a une couche d'entrée (recevant les données), une couche de sortie (donnant les résultats du réseau) et une ou plusieurs couches cachées (entre la couche d'entrée et la couche de sortie). S'il y a plus d'une couche cachée, alors le réseau est dit *profond* et est utilisé dans les approches d'apprentissage profond (*deep learning*) pour l'apprentissage de concepts plus complexes [6].
L'entraînement d'un réseau de neurones se fait en modifiant la valeur de ses poids synaptiques (modélisant l'importance des connexions entre les neurones). Il existe alors plusieurs méthodes d'apprentissage [6]. Si les données d'entraînement sont étiquetées avec le concept qu'elles représentent, on parle d'apprentissage supervisé. Si une partie seulement des données est étiquetée, il s'agit d'apprentissage semi-supervisé. Si aucune donnée n'est étiquetée, on parle d'apprentissage non supervisé. Pour finir, l'apprentissage par renforcement [24] permet au réseau d'apprendre grâce à un système de récompense : si le réseau donne un résultat correct alors il reçoit un signal positif sinon, il reçoit un signal négatif indiquant qu'il doive corriger la valeur de ses poids synaptiques.
Ainsi, l'IA connexionniste est pertinente par apport au contexte métier présenté ci-dessus notamment pour établir des modèles proactifs et enrichissant la base de connaissances.

4.3 Limites

L'IA symbolique et l'IA connexionniste, présentées ci-dessus, ont toutes deux des avantages et des caractéristiques intéressantes par rapport au contexte métier présenté en section 2. Cependant, elles présentent des limites non négligeables qui font qu'il n'est pas possible d'utiliser uniquement l'une ou l'autre de ces approches dans le cadre du projet AI4NP.
Tout d'abord, les approches issues de l'IA symbolique ne peuvent pas apprendre de nouvelles connaissances à partir d'exemples, ne peuvent manipuler des données non structurées et sont peu tolérantes aux bruits dans les données [4], [21] et [47]. Par conséquent, une base de connaissances peut difficilement s'adapter à l'évolution du domaine qu'elle représente sans intervention humaine (elle doit être enrichie à la main). Ceci est problématique car le domaine des stupéfiants est amené à changer fréquemment et les autorités doivent s'adapter rapidement aux nouvelles méthodes des trafiquants. De plus, les données issues des bases STUPS et OTARIES sont sujettes au bruit (à cause d'erreurs humaines ou de défauts matériel lors des mesures) et les réseaux de trafic forment des graphes difficilement manipulables avec de l'IA symbolique.
Ensuite, malgré leur capacité d'apprentissage, les réseaux de neurones donnent des résultats qui sont difficilement explicables. On parle de modèle en boîte noire (ou black-box) i.e. la connaissance n'est pas explicitement représentée mais distribuée au sein du réseau. Il est alors presque impossible de comprendre le lien entre les valeurs d'entrée et les résultats fournis. Cette caractéristique peut poser un problème pour l'aide à la prise de décision dans les domaines sensibles car il est essentiel pour un expert, de comprendre tous les résultats fournis. De plus, la connaissance métier n'est pas prise en compte au sein de ces modèles qui peuvent alors fournir des résultats non pertinents pour les experts [4], [21], [47].
L'IA neuro-symbolique combine les deux approches présentées ci-dessus. Elle ne présente pas les limites de l'IA symboliques et de l'IA connexionniste tout en conservant leurs avantages. Un état de l'art de ce domaine est dressé dans la section 5.

5 État des lieux de l'IA neuro-symbolique

L'IA neuro-symbolique est un courant de l'IA cherchant à coupler des approches de l'IA symbolique avec des approches connexionnistes afin de profiter des avantages de ces deux approches sans leurs inconvénients [21]. On notera quelques différences au niveau de la sémantique des termes employés en IA neuro-symbolique par rapport à celle des termes de l'IA symbolique. Aussi dans ce qui suit, une « base de

connaissances » est à comprendre comme un ensemble d'axiomes spécifiés dans un formalisme logique, autre que la logique de description vue en section 4.1 IA Symbolique. Une « règle » est à interpréter en tant que structure SI {ensemble axiomes logiques} ALORS {ensemble axiomes logiques}.

Au fil des années, les modèles de l'IA neuro-symbolique ont été classés selon plusieurs critères. Parmi les premières classifications établies, on peut citer celle d'Hilario [20], qui distingue les modèles selon la façon dont ils sont construits et leur structure. En 2020, dans [4], les auteurs classent les modèles selon leur degré d'explicabilité i.e. la façon dont les détails d'un modèle et ses résultats sont compréhensibles. Enfin, en 2021, dans [47], les auteurs proposent une classification selon la manière dont les composants symboliques (e.g. listes de règles, arbres de décisions, ontologies etc.) et les composants connexionnistes (i.e. réseaux de neurones) sont intégrés les uns par rapport aux autres. Une présentation de ces trois classifications ainsi que des exemples de modèles est faite dans cette section. Ces classifications serviront ensuite à se positionner (cf. section 7) quant au(x) type(s) d'approches pour répondre aux différentes problématiques métiers et scientifiques (voir sections 2 et 3).

Chronologiquement, la première classification, parmi celles citées plus haut, est celle de Hilario [20]. Cette classification est composée de deux catégories principales : les **approches unifiées** et les **approches hybrides** (la figure 2 illustre cette classification).

Les **approches unifiées** utilisent un réseau de neurones pour effectuer l'entièreté du traitement symbolique. Deux sous-catégories sont distinguées. On trouve d'abord le *traitement symbolique neuronal* dont l'objectif est de reproduire les hautes fonctions du cerveau. Proche de la neuroscience, les neurones utilisés pour la construction des réseaux sont des neurones imitant le fonctionnement des neurones biologiques. La construction des réseaux de neurones se fait de manière ascendante : le point de départ est le neurone, à partir duquel un réseau complexe est créé et modifié jusqu'à ce que la fonction souhaitée soit réalisée. Un exemple de ce type d'approche est le *Neural Darwinism* [12].

La seconde sous-approche est le *traitement des symboles connexionnistes*. Des réseaux de neurones basés sur des neurones formels sont utilisés pour former des architectures cognitives capables de traitements symboliques complexes. L'approche de construction des réseaux est descendante : le point de départ est la fonction symbolique à réaliser à partir de laquelle une structure connexionniste est élaborée. La représentation de la connaissance de ce type d'approche peut être locale, distribuée ou combinée (à la fois locale et distribuée). Dans une représentation locale, chaque nœud du réseau est un concept. Dans une représentation distribuée, un concept émerge de l'interaction entre plusieurs nœuds. On peut citer en exemple le modèle BoltzCONS [43].

La seconde catégorie donnée dans [20] sont les **approches hybrides**. Elles reposent sur le fait que des interactions entre les composants symboliques et connexionnistes sont nécessaires pour fournir des résultats satisfaisants. Là aussi, deux sous-catégories sont distinguées. D'abord, les *approches hybrides translationnelles*, qui utilisent un seul réseau de neurones pour le traitement symbolique mais, contrairement aux approches unifiées, ce réseau est construit à partir d'une structure symbolique (par exemple des règles logiques). Le réseau est alors contraint par la logique et fournit des résultats en accord avec la connaissance préalable. Ce type de fonctionnement permet de réduire les données nécessaires à l'entraînement du réseau car celui-ci contient la connaissance décrite par la structure. L'entraînement du réseau permet de mettre à jour sa structure pouvant alors être utilisée pour raffiner la structure symbolique de départ [44]. On retrouve au sein de cette catégorie un certain nombre de modèles comme, les *Knowledge-Based Artificial Neural Networks* (KBANN) [45] où des règles en logique propositionnelle sont encodées dans un réseau de neurones à une couche cachée. Cet encodage définit la structure du réseau et fixe les poids synaptiques initiaux. Comme dit précédemment, cela permet de contraindre les résultats (ils ne peuvent être que les conséquences des règles logiques) et réduit le nombre d'itérations nécessaires à l'entraînement (car le réseau contient la sémantique décrite par les règles). Ce modèle a ensuite été étendu par le *Conectionist Indictuve Learning and Logic Programming* (C-IL²-P) [16] pour améliorer les capacités d'apprentissages. Ceci est fait en modifiant la façon dont les règles sont encodées dans le réseau et en ajoutant des boucles de rétroaction entre les neurones d'entrée et de sortie du réseau. Le *Connectionist Inductive Learning and Logic Programming++* (CILP++) [15] est une extension du C-IL²-P pour traiter des règles en FOL. Parmi les approches plus récentes, nous pouvons citer les *Logic Tensor Networks* (LTN) [3] où une base de connaissances est traduite en *logique réelle* (logique proposée par les auteurs) dont les domaines des variables sont des nombres réels permettant ainsi de gérer des cas d'incertitude. Les réseaux utilisés sont des réseaux de tenseurs [5]. Le modèle LYRICS [30] exploite la logique floue [26] afin de gérer l'incertitude au sein des données. Ce modèle appartient à la famille de la *Differentiable Fuzzy Logic* (DFL) [27] et [28]. Dans ce type de modèle, les règles en FOL sont traduites en règles de la logique floue : les opérateurs logiques du premier ordre sont traduits en opérateurs de la logique floue (t-norm, t-conorm, S-implication et R-implication) et les prédicats sont interprétés grâce à des modèles de *deep learning* pouvant être entraînés.

La seconde sous-catégorie des approches hybrides présentée par Hilario [20], sont les *approches hybrides fonctionnelles*. Dans ce type d'approches, les composants connexionnistes et symboliques fonctionnent conjointement. Les composants peuvent être faiblement couplés, c'est-à-dire que les interactions entre les composants sont clairement localisées dans le temps et l'espace, initiées par l'un des composants ou un agent extérieur. Les données sont alors transférées d'un composant à un autre par un appel de fonction ou de procédure. Les composants peuvent aussi être fortement couplés c'est-à-dire que les données et les connaissances sont partagées entre les composants via une structure de données commune. Un changement dans la structure par l'un des composants entraîne des changements dans tous les autres composants.

Il existe aussi quatre modes (ou schémas) d'intégration spécifiant la façon dont les composants sont intégrés les uns par rapport aux autres :

- *Chainprocessing* : un des composants est le composant principal (les autres font des tâches en amont ou en aval). On peut citer en exemple ExpressGNN [50].
- *Subprocessing* : un des composants est intégré et

subordonné à l'autre composant.
- *Metaprocessing* : un composant est le composant de base chargé de résoudre le problème, l'autre agit au niveau des métadonnées (par exemple, pour contrôler les résultats).
- *Coprocessing* : tous les composant sont au même niveau.

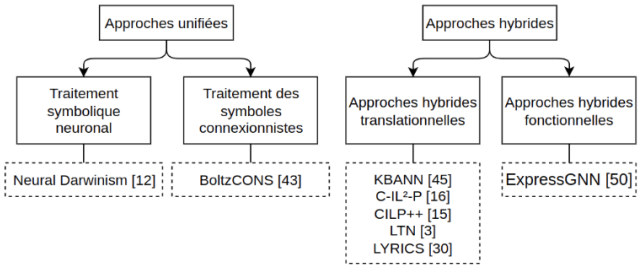


Figure 2. Illustration de la classification d'Hilario, adaptée de [20].

La seconde classification présentée ici est celle proposée en 2020 par Calegari et al. [4]. Les différents modèles sont triés selon leur type d'interprétabilité. Cette classification est illustrée par la figure 3.

Comme précédemment, deux catégories principales sont distinguées : les **approches par intégration** et les **approches par composition**. Tout d'abord, les **approches par intégration** regroupent les modèles intégrant les connaissances du composant symbolique au sein du composant connexionniste. Deux sous-catégories sont distinguées. La première sous-catégorie est l'*intégration via des prédicteurs numériques* (comme des réseaux de neurones par exemple). On retrouve ici des modèles tels que les KBANN [45], LYRICS [30], LTN [3] ou encore la *Semantic Loss Function* (SLF) [49] où l'idée est de manipuler, grâce à de la logique, la fonction de perte utilisée dans l'entraînement d'un réseau de neurones.

La seconde sous-catégorie est l'*intégration via des approches probabilistes ou statistiques*. On y retrouve les CILP++ [15], DeepProbLog [34] ou encore la programmation logique inductive différentiable (∂ILP pour *Differentiable Inductive Logic Programming*) [13].

La seconde catégorie, les **approches par composition**, regroupe les modèles combinant des modèles symboliques et connexionnistes. On distingue deux types de composition : l'*extraction* et l'*injection*. Premièrement, l'*extraction* regroupe les modèles qui extraient des connaissances symboliques des composants connexionnistes. Les auteurs abordent les cas de l'extraction de listes de règles (de la forme *Si … Alors … Sinon …*) et de l'extraction d'arbres de décision. L'extraction de règles peut être faite selon des approches *pédagogiques* comme dans les modèles ALPA [14] et RxREN [1]. Selon ces approches, le composant connexionniste (considéré comme un oracle) est interrogé avec toutes les possibilités de valeurs pour les variables d'entrée. Chaque résultat donné par le composant est une conséquence des entrées et permet de dresser la liste des règles. L'autre type d'approches d'extraction sont les approches *décompositionnelles*. De manière générale, dans ce type d'approches, les liens entre les neurones et les fonctions d'activation des neurones sont étudiés pour former un ensemble de règles. On y trouve par exemple le modèle RX [39].

L'extraction d'arbre de décision peut être fait de plusieurs manières, entre autres, en interrogeant le composant connexionniste comme un oracle, en utilisant une interprétation bayésienne ou en encore en analysant la structure interne du réseau. Par exemple, Schetinin et al. [40] proposent une méthode d'extraction de connaissances sous la forme de forêt aléatoire.

Deuxièmement, l'*injection* concerne les modèles où la connaissance symbolique est injectée dans le composant symbolique. Les auteurs présentent en particulier les approches injectant des graphes de connaissances comme le modèle OSCAR [19]. Ce modèle traduit les graphes dans un espace vectoriel pouvant être traité par un réseau de neurones. La prise en compte de ces graphes de connaissances permet, d'une part, d'améliorer les performances du composant connexionniste et, d'autre part, l'explicabilité par la conception car la connaissance est utilisée comme une contrainte sur le composant connexionniste.

Concernant le type d'explicabilité des catégories de cette classification, les approches par intégrations et par injection de connaissances sont explicables par la conception. Cela signifie que les modèles sont soit transparents (et donc explicables) soit des boîtes noires contraintes par la logique et de ce fait explicable. Les approches par composition ont une explication *post-hoc* ce qui signifie que les modèles doivent être manipulés pour extraire les connaissances qu'ils contiennent.

Dans [4], selon les auteurs, les modèles explicables par la conception sont les plus pertinents dans un contexte où les prises de décision sont critiques car ils mettent l'accent sur le compréhension des modèles et de leurs résultats au détriment des performances. De l'autre, les méthodes *post-hoc* permettent d'utiliser le plein potentiel des composants connexionnistes au détriment de l'explicabilité.

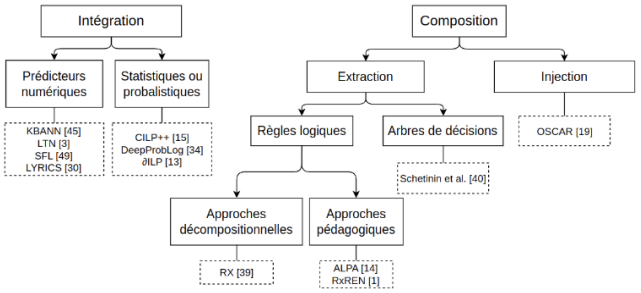


Figure 3. Illustration de la classification de Calegari et al., adaptée de [4].

Pour finir, Yu et al. [47], classent les modèles selon trois modes d'intégration des composants symboliques et connexionnistes. Cette classification est illustrée par la figure 4.

Le premier mode est l'**apprentissage pour le raisonnement**. Ce mode peut être vu comme un type de *chainprocessing* [20] où le premier composant est un réseau de neurones et le second est un composant symbolique. L'utilisation d'un réseau de neurones permet de prétraiter les structures de données complexes (image, texte etc.) afin de les abstraire pour qu'elles puissent être utilisée dans le raisonnement fait par un composant symbolique. Cela permet aussi de réduire l'espace de recherche améliorant ainsi les performances du composant symbolique. Il est possible de citer en exemple des modèles tels que pLogicNet [36] ou ExpressGNN [50] qui effectuent un prétraitement des graphes afin de simplifier les inférences faites sur ceux-ci.

Le second mode est le **raisonnement pour l'apprentissage**.

L'apprentissage fait par le réseau est guidé en prenant en compte la connaissance symbolique (comme dans les approches par intégration ou injection de [4]). En cas de manque de données pour l'entraînement, la connaissance peut être intégrée au réseau de neurones afin de transférer la connaissance sémantique (le temps d'apprentissage est alors réduit). Un exemple de raisonnement pour l'apprentissage est le modèle *Iterative Rule Distillation* (ou *Harnessing Deep Neural Networks with logic*) [22]. Dans ce modèle, un réseau de neurones professeur encode la connaissance et intervient dans le processus d'apprentissage d'un réseau élève qui tente de prédire le label des données d'entraînement et le résultat du réseau professeur. Il existe des modèles plus récents comme PROLONETS [41], basé sur l'apprentissage par renforcement ou *Context-Aware Zero-Shot Recognition* [29], utilisant un ensemble de données pour la reconnaissance d'objets dans un contexte de *zero-shot learning* (ZSL) [46] (i.e. prédiction de la classe d'un échantillon sans que cette classe n'ait été observée lors de la phase d'entraînement).

Le dernier mode présenté dans [47] est l'**apprentissage-raisonnement** où les composants travaillent ensemble, au même niveau. Ce mode d'intégration a pour objectif de combiner les deux modes présentés ci-dessus. D'une part, le composant symbolique contraint le réseau de neurones dans son apprentissage et, d'autre part, le composant connexionniste abstrait la représentation des données pour faciliter le raisonnement fait par le composant symbolique et peut même enrichir la connaissance symbolique. Selon la classification de Hilario [20], ce mode d'intégration serait un type de *coprocessing*. On peut donner en exemple le modèle DeepProbLog [34], extension de ProbLog [37]. Les prédicats des formules de la logique du premier ordre sont implémentés par des réseaux de neurones entraînables par descente de gradient. Les perceptions de bas niveau sont alors réalisées par les réseaux de neurones tandis que le raisonnement est fait au niveau de la logique. L'apprentissage par abduction (ABL pour *ABductive Learning*) [48], parfois appelé retro-production, est un autre exemple d'apprentissage-raisonnement. Des faits et des hypothèses sont inférés à partir d'une base de connaissances pour expliquer des observations.

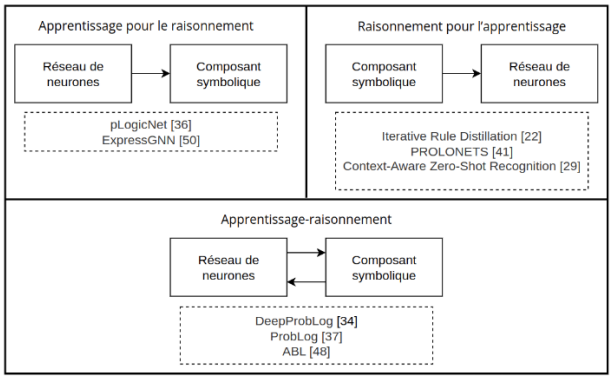


Figure 4. Illustration de la classification de Yu et al., adaptée de [47] .

6 Défis applicatifs

Cette section présente les défis applicatifs découlant des contraintes métiers et de l'état actuel des bases de données.

Le premier défi concerne le traitement de données hétérogènes et multivariées. Pour les bases STUPS© et OTARIES©, l'hétérogénéité est due aux nombres d'acteurs remplissant la base ainsi qu'à leur longévité (notamment pour STUPS©). En effet, différents acteurs n'auront pas toujours la même rigueur lors de l'insertion de données dans les bases et peuvent avoir des avis différents quant à la pertinence d'une donnée. De plus, tous les laboratoires n'analysent pas exactement les mêmes caractéristiques d'un échantillon de drogue (cela tend à s'uniformiser) pouvant conduire à des données manquantes. La longévité des bases peut aussi être une source d'hétérogénéité car les modifications du schéma d'une base au cours des années peuvent conduire à avoir des données manquantes (par exemple, les nouveaux champs de tables déjà existantes ne peuvent pas toujours être remplis pour les anciennes données). Cette longévité donne aussi lieu à une inconsistance dans le temps avec des périodes où le nombre de saisies dans les bases est plus ou moins important que pour d'autres périodes.

Le second défi est lié au fait que les bases de données ne contiennent pas de connaissances métier des experts de la SNPS. Sans ces connaissances, les approches neuro-symboliques ne peuvent pas être exploitées pleinement. Il faudra alors échanger avec les experts pour construire une base de connaissances. Le fait de construire la base en échangeant avec plusieurs experts pose un problème. En effet, dans des domaines où l'analyse est sujette à interprétation, si des experts différents utilisent des règles d'analyse différentes, il faut être capable de prioriser les règles rajoutant ainsi de la complexité dans l'acquisition des connaissances. Ces deux défis applicatifs sont directement reliés à la problématique métier PM1.

Le troisième défi applicatif est lié à PM4 et concerne la modélisation de plusieurs axes sous forme granulaire. En effet, il faudra d'abord identifier quels axes et niveaux de granularité sont utiles aux experts. De plus, une fois ces axes et leurs niveaux identifiés, il faudra étudier leur modélisation. En effet, faudra-t-il modéliser chaque niveau de chaque axe indépendamment ou alors des niveaux d'axes différents ensembles ?

Pour finir, ce travail d'analyse doit s'inscrire dans la durée. L'approche proposée doit pouvoir être adaptée afin que de nouvelles connaissances ou de nouveaux modèles y soient ajoutées car le domaine des stupéfiants est en constante évolution donc, un système figé ne peut pas être utilisé.

7 Premiers travaux et approches envisagées

En tenant compte des différentes éléments présentés dans les sections précédentes, de premiers travaux ont été conduits et des pistes d'approches ont été envisagées.

Afin de répondre à la problématique PS1, nous avons conçu une ontologie pour modéliser les connaissances métier des experts de la SNPS. La TBox ce cette ontologie décrit un ensemble de concepts ainsi que les conditions nécessaires et suffisantes pour que les instances peuplant la ABox appartiennent à ces concepts. La conception de l'ontologie est faite itérativement et la première version modélise les concepts de base nécessaires pour débuter nos travaux. Le concept de saisine a été modélisé en premier car il est à la base de toute analyse faite par les laboratoires de la police scientifique. Il s'agit d'un terme juridique utilisé en matière de procédure qui

désigne l'action qu'accomplit un requérant (on parle de service requérant) lorsqu'il demande l'analyse d'une saisie de drogue. Une saisine contient un ensemble de scellés, qui eux-mêmes contiennent un ou plusieurs prélèvements à partir desquels sont extraits des échantillons. Une saisine est faite dans une ville, par un service capteur, et la drogue de la saisine est à destination d'un pays pour la revente. La définition de ce concept est exprimée en DL comme suit :

$$Saisine \equiv= 1\ aServiceCapteur.ServiceCapteur \cap \exists aScelle.Scelle \cap = 1 estFaitA.Ville \cap = 1\ aPaysDestination.Pays \cap= 1\ aServiceRequerant.ServiceRequerant$$

Dans la première version de l'ontologie nous avons défini au total 43 concepts et 46 relations entre concepts. La présentation de tous ces éléments sort du cadre de cet article.

L'un des intérêts d'une ontologie est le partage et la réutilisation des connaissances qu'elle modélise. Cela évite de redéfinir des connaissances ayant déjà été modélisées dans d'autres travaux. Ainsi, dans une future version, nous allons nous appuyer sur d'autres ontologies, comme l'ontologie des unités de mesures (https://www.ebi.ac.uk/ols/ontologies/om). Celle-ci sera utilisée lorsqu'il sera nécessaire de préciser l'unité d'une mesure d'une caractéristique macroscopique d'un échantillon (par exemple, la masse ou les dimensions).

Au moment de l'écriture de cet article, le travail en cours porte sur l'utilisation de règles d'analyse pour faciliter le processus d'appariement d'échantillons. La métrique de ressemblance présentée dans la section 2 ne suffit pas à elle seule à rapprocher deux échantillons. En effet, les experts doivent procéder à un examen manuel des échantillons potentiellement proches afin d'éviter les faux positifs. Ainsi, pour valider ou réfuter un appariement, un ensemble de règles d'analyse sont utilisées. Par exemple, un expert peut se pencher sur les dates des saisines des échantillons et peut conclure à un faux positif si elles sont trop éloignées. L'ensemble de ces règles d'analyses ont été formalisées en utilisant la logique du premier ordre. Il reste alors à ajouter ces règles à l'ontologie actuelle afin que l'ensemble des connaissances (description du domaine métier et règles d'analyse) soit intégré au sein de la même structure symbolique. Il s'agit d'un choix que nous avons fait afin de limiter le nombre de structures à manipuler. La manière dont est faite cette intégration est en cours d'investigation. Une fois l'ontologie enrichie avec les règles d'analyse, il sera possible de construire des modèles issus de l'IA neuro-symbolique afin de répondre à la seconde problématique scientifique (PS2) - englobant les problématiques métiers PM2 et PM3.

Concernant PM2, le modèle construit servira à faciliter les analyses faites par les experts et à raffiner les règles d'analyse. Cela permettra de simplifier (voire améliorer) le processus d'appariement. Même si le choix du modèle n'est pas encore fait, il est d'ores et déjà possible de se positionner par rapport aux différentes classifications présentées dans la section 5. S'inscrivant dans un contexte d'aide à la prise de décision, le modèle retenu devra fournir des résultats compréhensibles par un humain (i.e. comprendre les causes amenant à ce résultats). De plus, la connaissance décrite au sein de l'ontologie devra être utilisée au sein du modèle pour avoir des résultats respectant la connaissance des experts.

Selon Calegari et al. [4], la connaissance symbolique peut être intégrée ou injectée dans le composant connexionniste. La prise en compte du graphe de connaissances décrit par l'ontologie semble ne pas être nécessaire pour l'appariement d'échantillons. En effet, seules les règles d'analyse pourraient être utilisées. Ainsi, nous pourrions limiter notre étude aux modèles intégrant la connaissance exprimée sous la forme de règles logique (par exemple le modèle CILP++ [15]). En revanche, si cette piste est fausse et que la prise en compte du graphe de connaissances s'avère être nécessaire alors des modèles injectant la connaissance seront à étudier. En plus de la prise en compte de la connaissance symbolique, il faut que le composant connexionniste enrichisse cette connaissance. Ce type de fonctionnement correspond à l'*apprentissage-raisonnement* présenté par Yu et al. [47]. En effet, le composant symbolique contraint le composant connexionniste qui, à son tour, va enrichir la connaissance du composant symbolique. Si l'on croise la classification [4] avec la classification [47], certains modèles peuvent être mis en évidence comme DeepProbLog [34] qui fait à la fois partie de l'*apprentissage-raisonnement* [47] et des modèles intégrant la connaissance symbolique [4]. Pour finir, selon la classification de Hilario [20], le choix le plus pertinent pour traiter PM2 semble être les approches hybrides fonctionnelles où les composants sont intégrés en *coprocessing*. En effet, les composants symboliques et connexionnistes doivent travailler ensemble au même niveau. Un couplage faible entre les composants pourra être suffisant dans la mesure où la vérification des règles par le composant symbolique sera clairement identifiée dans le temps (après l'entraînement du composant connexionniste).

Concernant la problématique métier PM3, le modèle prédictif conçu doit caractériser les produits stupéfiants en France et enrichir la base de connaissances si de nouveaux concepts pertinents sont identifiés dans les données. Comme pour la problématique métier PM2, le choix du modèle n'est pas encore fait mais il est possible de se positionner.

L'utilisation de la connaissance métier sera nécessaire afin de fournir des prédictions de qualité utiles aux experts. De fait, le type de fonctionnement entre les différents composants serait du type *apprentissage-raisonnement* [4]. En effet, le composant symbolique fournirait le graphe de connaissances de l'ontologie au composant connexionniste qui l'utiliserait pour effectuer des prédictions et potentiellement l'enrichir. Ces prédictions et l'enrichissement de la connaissance seraient alors vérifiées par le composant symbolique. Comme précédemment, il semble que les approches hybrides fonctionnelles [20] soient les plus pertinentes car les composants travailleraient ensemble au même niveau. Un couplage faible semble pouvoir être suffisant car la communication entre composants ne se ferait qu'une fois les prédictions faites. De plus, il pourrait être possible d'envisager l'affinement des prédictions grâce à l'apprentissage par renforcement. En effet, si l'on suppose le composant symbolique capable de valider ou non une prédiction (moyennant l'intervention humaine si nécessaire) alors il pourrait influencer le composant connexionniste lors de son entraînement. Par exemple, une prédiction non conforme aux connaissances métier enverrait un signal négatif au composant connexionniste qui serait en mesure de se corriger en prenant

en compte ce signal. Pour finir, l'utilisation du graphe de connaissances par le composant connexionniste pourrait être fait par une méthode d'injection de connaissances comme celles présentées dans [4]. En effet, les règles d'analyse ne seront pas nécessaires pour traiter cette problématique car elles ne sont utilisées que pour l'appariement d'échantillons.
Les catégories de modèles mises en avant pouvant répondre à la problématique PS2, ne représentent pas des conditions nécessaires pour qu'un modèle soit accepté. En d'autres termes, un modèle peut appartenir seulement à une partie des catégories envisagée ci-dessus ou à d'autres catégories.
Les pistes pour répondre à la problématique PS3, à savoir la modélisation de la connaissance sous différents niveaux de granularités, n'ont pas encore été étudiées.

8 Conclusion

Cet article présente les problématiques scientifiques traitées dans le cadre d'un projet de thèse financée par le MESRI. Ces problématiques scientifiques découlent des objectifs visés et des problématiques métiers du projet AI4NP.
Un état de l'art de l'intelligence artificielle neuro-symbolique est dressé à travers trois classifications de modèles. Elles classent les modèles selon leur structure, les interactions entre les différents éléments qui les composent et leur explicabilité. Les premiers travaux, menés en parallèle de l'écriture de cet article, sont présentés ainsi que des pistes d'approches pour les travaux futures. Ces futurs travaux concerneront la conception d'un modèle servant à faciliter l'appariement d'échantillons ainsi que d'un modèle prédictif caractérisant les produits stupéfiants circulant en France.
La représentation de la connaissance sous différents niveaux de granularité sera étudiée afin de choisir des niveaux pertinents pour les experts.
Actuellement, seules des approches combinant réseaux de neurones et IA symbolique ont été étudiées afin d'identifier des éléments de réponse. À mesure que nos analyses avancent, il sera peut-être nécessaire d'investiguer d'autres modèles, par exemple, ceux basés sur des kernel machines, comme c'est le cas dans [8].

Remerciements

Les travaux présentés se déroulent dans le cadre d'un Contrat Doctoral financé par le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation. Une convention de coopération tripartite réunissant l'Université de Bourgogne, l'Institut Agro Dijon et le Service National de la Police Scientifique (SNPS) accompagne ce contrat doctoral. Nous tenons à remercier le SNPS pour l'accès à ses bases de données et particulièrement à Mme Céline Charvoz, cheffe de la section Stupéfiants du laboratoire de Police Scientifique de Lyon et M. Fabrice Besacier, sous-directeur adjoint de la Sous-direction de la Stratégie, de l'Innovation et du Pilotage du SNPS pour les renseignements fournis.

9 Bibliographie

[1] M.G. Augasta, T. Kathirvalavakumar, Reverse engineering the neural networks for rule extraction in classification problems, *Neural processing letters*, Vol. 35(2), pp. 131–150, 2012.

[2] F. Baader, W. Nutt, Basic description logics. In: *The description logic handbook: theory, implementation, and applications*. pp. 43-95, 2003.

[3] S. Badreddine, A. Garcez, L. Serafini, M. Spranger, Logic tensor networks, *Artificial Intelligence*, Vol. 303, DOI : 10.1016/j.artint.2021.103649, 2022.

[4] R. Calegari, G. Ciatto, A. Omicini, On the integration of symbolic and sub-symbolic techniques for XAI: A survey, *Intelligenza Artificiale*, Vol.14(1), pp.7-32, 2020.

[5] R. Socher, D. Chen, C. D. Manning, A. Y. Ng, Reasoning With Neural Tensor Networks For Knowledge Base Completion, *Advances in Neural Information Processing Systems*, Vol. 26, 2013.

[6] A. Cornuéjols, L. Miclet, V. Barra, Apprentissage artificiel : Deep learning, concepts et algorithmes, Eyrolles, 2018.

[7] R. Dastres, M. Soori, Artificial Neural Network Systems. *International Journal of Imaging and Robotics (IJIR)*, 21 (2), pp.13-25, 2021.

[8] M. Diligenti, M. Gori, C. Sacca, Semantic-based regularization for learning and inference, *Artificial Intelligence*, Vol. 244, pp. 143-165, 2017.

[9] V. Dufey, L. Dujourdy, F. Besacier, H. Chaudron, A quick and automated method for profiling heroin samples for tactical intelligence purposes, *Forensic Science International*, Vol. 169(2), pp. 108-117. doi: 10.1016/j.forsciint.2006.08.003, 2007.

[10] L. Dujourdy, F. Besacier**,** A study of cannabis potency in France over a 25 years period (1992–2016*), Forensic Science International*, Vol. 272, pp. 72–80. doi: 10.1016/j.forsciint.2017.01.007, 2017.

[11] L. Dujourdy, F. Besacier, Headspace profiling of cocaine samples for intelligence purposes, *Forensic Science International*, Vol. 179(2-3), pp. 111-22. doi: 10.1016/j.forsciint.2008.04.024, 2008.

[12] G. M. Edelman, Bright air, brilliant fire: On the matter of the mind, BasicBooks, 1992.

[13] R. Evans, E. Grefenstette, Learning Explanatory Rules from Noisy Data, *Journal of Artificial Intelligence Research*, Vol. 61, pp. 1–64, DOI:10.1613/jair.5714, 2018.

[14] E.J. de Fortuny and D. Martens, Active Learning-Based Pedagogical Rule Extraction, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26(11), pp. 2664–2677, 2015.

[15] M. V. França, G. Zaverucha, A. S. Garcez, Fast relational learning using bottom clause propositionalization with artificial neural networks, *Machine learning*, Vol. 94, pp. 81-104, 2014.

[16] A. Garcez, G. Zaverucha, The connectionist inductive learning and logic programming system, *Applied Intelligence*, Vol. 11, pp. 59-77, 1999.

[17] S. Grimm, A. Abecker, J. Völker, R. Studer, Ontologies and the semantic web. *Handbook of Semantic Web Technologies*, pp.507–579, DOI: 10.1007/978-3-540-92913-0_13, Springer, 2011.

[18] M. Flasiński, Symbolic Artificial Intelligence. In: *Introduction to Artificial Intelligence*. Springer, Cham.

pp. 15-22, DOI : 10.1007/978-3-319-40022-8_2, 2016.
[19] T.R. Goodwin, D. Demner-Fushman, Bridging the Knowledge Gap: Enhancing Question Answering with World and Domain Knowledge, arXiv:1910.07429, 2019.
[20] M. Hilario, An overview of strategies for neurosymbolic integration, *Connectionist-Symbolic Integration: From Unified to Hybrid Approaches*, Psychology Press, 1997.
[21] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, Lu Zhou, Neuro-symbolic approaches in artificial intelligence, *National Science Review*, Volume 9(6), 2022.
[22] Z. Hu, X. Ma, Z. Liu, E. Hovy, E. Xing, *Harnessing deep neural networks with logic rules*, arXiv:1603.06318v6, 2020.
[23] A.K. Jain, J. Mao, K.M. Mohiuddin, Artificial neural networks: A tutorial, *Computer*, Vol. 29(3), pp. 31-44, 1996.
[24] L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: A survey, *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237-285, 1996.
[25] C. M. Keet, An introduction to ontology engineering. V1.5, 2020. [Disponible en ligne] https://people.cs.uct.ac.za/~mkeet/files/OEbook.pdf [Consulté le 25/02/2023]
[26] G. Klir, B. Yuan, *Fuzzy sets and fuzzy logic*, DOI:10.5860/choice.33-2786, Prentice Hall, 1995.
[27] E. van Krieken, E. Acar, F. van Harmelen, *Analyzing differentiable fuzzy implications*, arXiv preprint arXiv:2006.0347, 2020.
[28] E. van Krieken, E. Acar, F. van Harmelen, Analyzing differentiable fuzzy logic operators, *Artificial Intelligence*, Volume 302, 2022.
[29] R. Luo, N. Zhang, B. Han, L. Yang, *Context-Aware Zero-Shot Recognition*, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34(07), DOI: 10.1609/aaai.v34i07.6841, 2020.
[30] G. Marra, F. Gianni, M. Diligenti, M. Gori, Lyrics: A general interface layer to integrate logic inference and deep learning, *Proceedings Of ECML PKDD 2019*, Springer International Publishing, 2020.
[31] Plan national de lutte contre les stupéfiants, Dossier de presse, 17 septembre 2019. [Disponible en ligne] https://www.interieur.gouv.fr/Archives/Archives-des-dossiers/Plan-national-de-lutte-contre-les-stupefiants [Consulté le 25/02/2023]
[32] P.D. Magnus, *forall x: An Introduction to Formal Logic*, 2017. [Disponible en ligne] https://www.fecundity.com/codex/forallx.pdf [Consulté le 25/02/2023]
[33] R. Mikut, M. Reischl. Data mining tools, *Wiley interdisciplinary reviews: data mining and knowledge discovery*, Vol. 1(5), pp. 431-443, 2011.
[34] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L. De Raedt, *DeepProbLog: Neural probabilistic logic programming*, Advances in Neural Information Processing Systems, Vol. 31, 2018.
[35] A. Newell, H.A. Simon, Computer Science as Empirical Inquiry: Symbols and Search, *Communications of the ACM*, Vol. 19 (3), pp. 113–126, doi:10.1145/360018.360022, 1976
[36] Qu Meng et Jian Tang. Probabilistic logic neural networks for reasoning, *Advances in neural information processing systems*, Vol. 32, pp.7710-7720, 2019.
[37] L. De Raedt, A. Kimmig, H. Toivonen, *ProbLog: A probabilistic Prolog and its application in link discovery*, International Joint Conference on Artificial Intelligence (IJCAI), 2007.
[38] S. Russell, P. Norvig, *Intelligence Artificielle – Une approche moderne*, Traduit par L. Miclet, F. Popineau, C. Cadet, 4e éd. Pearson, 2021.
[39] R. Setiono, Extracting Rules from Neural Networks by Pruning and Hidden-Unit Splitting, *Neural Computation*, Vol. 9(1), pp. 205–225, 1997.
[40] V. Schetinin, J.E. Fieldsend, D. Partridge, T.J. Coats, W.J. Krzanowski, R.M. Everson, T.C. Bailey, A. Hernandez, Confident interpretation of Bayesian decision tree ensembles for clinical applications, *IEEE Transactions on Information Technology in Biomedicine*, Vol. 11(3), 2007.
[41] A. Silva, M. Gombolay, Encoding human domain knowledge to warm start reinforcement learning, *Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 35(6), pp. 5042–5050, arXiv:1902.06007v4, 2021.
[42] R. Studer, V.R. Benjamins, D. Fensel, Knowledge engineering: principles and methods, *Data & knowledge engineering*, Vol. 25(1), Pp.161–197, 1998
[43] D. S. Touretzky, BoltzCONS : Dynamic symbol structures in a connectionist network, *Artificial intelligence*, Vol. 46(1-2), pp. 5-46, 1990.
[44] G. G. Towell, Symbolic knowledge and neural networks: Insertion, refinement and extraction, 1992.
[45] G. G. Towell, J. W. Shavlik, Knowledge-based artificial neural networks, *Artificial intelligence*, Vol. 70(1-2), pp. 119-165, 1994.
[46] W. Wang, V. W. Zheng, H. Yu, C. Miao, *A survey of zero-shot learning: Settings, methods, and applications*, ACM Transactions on Intelligent Systems and Technology (TIST), Vol. 10(2), pp. 1-37, DOI:10.1145/3293318, 2019.
[47] D. Yu, B. Yang, D. Liu, H. Wang, S. Pan, *Recent Advances in Neural-symbolic Systems: A Survey*, arXiv:2111.08164, 2021.
[48] Z.-H. Zhou, *Abductive learning: towards bridging machine learning and logical reasoning*, Science China Information Sciences, Vol. 62(7), pp. 1–3, 2019.
[49] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. Broeck, *A Semantic Loss Function for Deep Learning with Symbolic Knowledge*, 35th Conference on Machine Learning (ICML 2018), Vol. 80, pp. 5502–5511, 2018.
[50] Y. Zhang, C. Xinshi, Y. Yuan, A. Ramamurthy, B. Li, Y. Qi et L. Song, *Efficient probabilistic logic reasoning with graph neural networks*, arXiv preprint arXiv:2001.11850 (2020).

RL-Net: Apprentissage de Règles Interprétables avec des Réseaux de Neurones

Lucile Dierckx[1,2], Rosana Veroneze[2,3], Siegfried Nijssen[1,2]

[1] Institut TRAIL, Louvain-la-Neuve, Belgique
[2] UCLouvain, ICTEAM/INGI, Louvain-la-Neuve, Belgique
[3] Unicamp, FEEC/DCA, Campinas-SP, Brésil

{prénom.nom}@uclouvain.be

Résumé

Les modèles symboliques et interprétables, tels que ceux à règles, sont étudiés depuis longtemps car de nombreux domaines ont besoin d'interpétabilité. De récents travaux ont étudié l'apprentissage de règles utilisant des approches pour apprendre des réseaux neuronaux plutôt que des heuristiques. Ces travaux se limitent aux règles non-ordonnées. Nous proposons RL-Net, qui apprend des règles ordonnées avec un réseau neuronal. Notre modèle performe de manière similaire aux algorithmes de pointe pour la classification binaire et multiclasse, et peut être adapté aux tâches à étiquettes multiples.

Mots-clés

Interprétabilité, extraction d'ensembles de motifs, apprentissage de règles, réseaux neuronaux binaires.

Abstract

Symbolic, interpretable models, such as rule-based models, have been studied for years due to the need for interpretable classification models in many domains. Recent studies have explored gradient-based rule learning using neural networks instead of heuristics. However, these works focus on unordered rule sets only. We propose RL-Net, an approach for learning ordered rule lists based on neural networks. Our model performs similarly to state-of-the-art algorithms for rule learning in binary and multi-class classification settings, and can be adapted to multi-label tasks.

Keywords

Interpretability, pattern set mining, rule learning, binary neural networks.

1 Introduction

Cet article a été accepté à PAKDD 2023 [6]

De nombreux domaines d'application requièrent des modèles de classification non seulement performants mais aussi interprétables. C'est pourquoi, la recherche sur la classification symbolique et interprétable est un domaine fort présent dans la littérature. Une catégorie importante de ces classificateurs interprétables est celle des modèles basés sur des règles de logique [4, 5]. Ces classifieurs à base de règles utilisent des heuristiques pour apprendre les règles interprétables et sont conçus pour des tâches de classification spécifiques. De récentes études se sont penchées sur l'utilisation d'approches basées sur le gradient des réseaux neuronaux afin d'apprendre ce type de classificateur. Leur objectif est de combiner l'apprentissage neuronal et l'apprentissage symbolique pour pouvoir ensuite exploiter la littérature sur les réseaux neuronaux dans le cadre de l'apprentissage de règles. Les articles existants sur cette combinaison neuro-symbolique se concentrent sur les classificateurs utilisant des règles non-ordonnées [2, 3, 7, 9, 10, 11]. Cependant, une grande partie de la littérature portant sur les modèles à règles utilise des règles ordonnées [5], qui ont l'avantage de rester entièrement interprétables, même pour des tâches de classification avec plus de deux classes. La classification avec des listes de règles ordonnées n'a pas encore été étudiée dans un cadre neuro-symbolique entièrement interprétable.

Dans ce travail, nous nous concentrons sur cette classification en étendant le réseau de règles de décision (DR-Net) de Qiao et al [9]. Nous avons implémenté une hiérarchie entre les règles, permettant ainsi d'apprendre des classificateurs basés sur des listes ordonnées de règles au lieu d'ensembles non-ordonnés de règles. De plus, notre modèle permet de résoudre des problèmes de classification multiclasses au lieu d'être limité aux problèmes binaires. Enfin, notre proposition peut facilement être adaptée pour résoudre les problèmes de classification à étiquettes multiples.

2 Approche

Notre réseau de neurones pour l'apprentissage de règles (RL-Net) a été conçu pour apprendre des listes de règles interprétables qui peuvent faire de la classification multiclasse. RL-Net utilise la structure d'un réseau neuronal ainsi que son apprentissage par optimisation du gradient.

Le réseau neuronal que nous avons imaginé reproduit le modèle `SI ... ALORS ... ; SINON SI ... ALORS ... ; (...) ; ENFIN ...`. Pour ce faire, il est composé de quatre couches, tel qu'illustré Figure 1. La première couche est la couche d'entrée qui reçoit les carac-

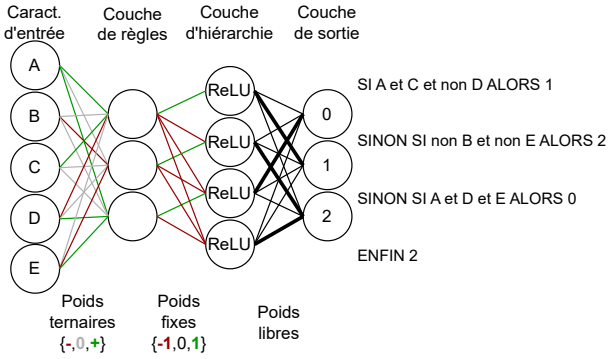


FIGURE 1 – Architecture de RL-Net

téristiques binarisées du jeu de données. Elle est connectée à la couche de règles, où les conditions composant les différentes règles sont apprises. La couche suivante exprime la hiérarchie [1] entre les règles, ce qui est nécessaire pour apprendre une liste de règles ordonnée au lieu d'un ensemble de règles non-ordonné. Enfin, la couche de sortie attribue une étiquette de classe spécifique à chaque règle. Pour rester entièrement interprétable, les activations et les poids des noeuds du modèle sont tous binarisés ou ternarisés [8] de façon à imiter le comportement de l'opération logique `ET` et la hiérarchie entre les règles. Grâce à son architecture neuronale, l'apprentissage de notre modèle de règles ordonnées peut être fait à l'aide des techniques classiques d'apprentissage des réseaux de neurones (telles que l'optimiseur Adam, la minimisation de l'entropie, ...). Une description plus détaillée de chaque couche du modèle est faite dans l'article original. Le code source se trouve sur GitHub [1].

3 Résultats

Le modèle que nous avons développé peut être utilisé pour des tâches de classification binaire et multiclasse, et atteint des performances similaires aux algorithmes fréquemment utilisés, RIPPER [5] et CART [4], en particulier lorsque le nombre maximal de règles défini par l'utilisateur est peu élevé. Nous avons également travaillé sur l'adaptation de notre modèle au contexte à étiquettes multiples. Dans cette configuration, nous observons que le modèle possède un bon potentiel d'apprentissage, mais ne peut pas encore rivaliser avec les algorithmes bien connus.
La description des datasets et expériences ainsi que les résultats détaillés sont disponibles dans l'article original.

4 Conclusion

Nous avons proposé RL-Net, une approche par réseau de neurones pour l'apprentissage de listes de règles ordonnées. Notre modèle a donné des résultats similaires à ceux des algorithmes de pointe pour l'apprentissage de règles dans des contextes de classification binaire et multiclasse, et a du potentiel pour être adapté aux tâches d'apprentissage à étiquettes multiples. Les pistes d'amélioration pour le modèle seraient de travailler plus en profondeur sur l'adaptation à la classification à étiquettes multiples ainsi que d'arriver à prendre plus avantage du nombre maximal de règles que l'utilisateur autorise le modèle à utiliser.

1. `https://github.com/luciledierckx/RLNet`

Remerciements

Ce travail a été soutenu par le Service Public de Wallonie Recherche sous la subvention n°2010235 - ARIAC by DIGITALWALLONIA4.AI et n°2110107 - SERENITY2 by WIN2WAL. R. Veroneze tient aussi à remercier FAPESP, Brésil (Subventions n°2017/21174-8 et 2020/00123-9).

Références

[1] John OR Aoga, Siegfried Nijssen, and Pierre Schaus. Modeling pattern set mining using boolean circuits. In *International Conference on Principles and Practice of Constraint Programming*. Springer, 2019.

[2] Florian Beck and Johannes Fürnkranz. An empirical investigation into deep and shallow rule learning. *Frontiers in Artificial Intelligence*, 4, 2021.

[3] Florian Beck and Johannes Fürnkranz. An investigation into mini-batch rule learning. *arXiv preprint arXiv :2106.10202*, 2021.

[4] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[5] William W Cohen. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123. Elsevier, 1995.

[6] Lucile Dierckx, Rosana Veroneze, and Siegfried Nijssen. Rl-net : Interpretable rule learning with neural networks. In *PAKDD 2023 : Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, 2023.

[7] Jonas Fischer and Jilles Vreeken. Differentiable pattern set mining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 383–392, 2021.

[8] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L0 regularization. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[9] Litao Qiao, Weijia Wang, and Bill Lin. Learning accurate and interpretable decision rule sets from neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4303–4311, 2021.

[10] Shaoyun Shi, Yuexiang Xie, Zhen Wang, Bolin Ding, Yaliang Li, and Min Zhang. Explainable Neural Rule Learning. In *WWW 2022 - Proceedings of the ACM Web Conference 2022*, pages 3031–3041. Association for Computing Machinery, Inc, 2022.

[11] Yongxin Yang, Irene Garcia Morillo, and Timothy M Hospedales. Deep neural decision trees. *ICML Workshop on Human Interpretability in Machine Learning. arXiv preprint arXiv :1806.06988*, 2018.

Comment gagner : expliquer les bonnes trajectoires

M. Kazi Aoual[1,2], C. Rouveirol[2], H. Soldano[1 2,3], V. Ventos[1]

[1] Nukkai,France
[2] UMR CNRS 7030 Institut Galilée – Université Sorbonne Paris Nord, LIPN
[3] UMR CNRS 7205 Museum National d'Histoire Naturelle, ISYEB

mkazi[at]nukk.ai / rouveirol[at]lipn.univ-paris13.fr / henry.soldano[at]mnhn.fr / vventos[at]nukk.ai

Résumé

Dans un projet de construction d'un joueur artificiel pour un jeu de bridge très simplifié, nous nous intéressons ici à la question de la manière de jouer pour gagner. Cela nous amène à faire de l'apprentissage relationnel et à proposer une technique de constructions d'explications de trajectoires, ce qui implique en particulier de construire les clauses minimales couvrant un ensemble d'exemples. Nous présentons une étude de cas et discutons des questions encore ouvertes dans ce projet.

Mots-clés

Explications, Programmation Logique Inductive, Abduction

Abstract

To address how to build an artificial player for a simplified bridge game, we are interested in the question of how to play to win. This leads us to do relational learning and to propose a technique for constructing trajectory explanations, which involves in particular constructing minimal clauses covering a set of examples. We present a case study and discuss the open questions in this project.

Keywords

Explanation, Inductive Logic Programming, Abduction

1 Introduction

Nous considérons les explications du classement d'une observation par un classifieur logique lorsque cette observation est un arbre des actions possibles au cours du temps et des états associés. Chaque branche représente alors une trajectoire possible. Notre motivation est un scénario où un joueur artificiel joue à un jeu de bridge simplifié, dans la position du déclarant contre un programme jouant le défenseur, et doit répondre à tout moment du jeu à des requêtes émises par un interlocuteur humain et de la forme "comment l'action choisie conduit-elle à un nombre de plis optimal ?". On optimise ici la récompense totale maximale que l'on peut obtenir en suivant une trajectoire à partir d'un état donné, au sens d'un processus de décision de Markov (MDP)[23].

Les explications recherchées pour le choix d'une action a dans un état donné s portent alors sur un certain nombre de trajectoires optimales possibles. Comme ce nombre peut-être élevé nous allons regrouper ces trajectoires optimales de manière à ce que l'optimalité des trajectoires de chaque groupe puisse être expliquée de la même manière. Cela nous amène à définir d'abord ce qu'est une *explication commune* pour l'optimalité d'un groupe de trajectoires. La première étape du processus explicatif est alors de former des groupes de trajectoires optimales ayant chacun au moins une explication commune de l'optimalité de ses trajectoires. Pour une paire (état, action) donnée nous proposons d'obtenir ces groupes en construisant un classifieur logique à base de règles. Les trajectoires *couvertes* par une règle, c'est-à-dire celles satisfaisant sa prémisse, auront alors au moins une explication commune par définition.

La notion d'explication explorée ici provient de divers travaux sur les explications abductives du label attribué par un classifieur logique à une observation décrite en logique propositionnelle ou par une liste de paires attribut-valeur [10, 2, 3, 6]. Nous avons adapté et étendu ces définitions dans plusieurs directions. Nos observations sont ici des trajectoires : nous associons à une paire (état, action) donnée (s, a) l'ensemble des trajectoires *possibles* que nous appelons *l'univers* associé à (s, a). Nous supposons également avoir un classifieur logique D qui étiquette les trajectoires comme optimales ou non-optimales. Les explications sont alors adaptées comme suit :

- Une trajectoire est décrite en logique d'ordre un comme un ensemble de faits, c'est-à-dire d'atomes instanciés. L'explication de l'optimalité d'une trajectoire p est alors définie comme un sous-ensemble des faits décrivant la trajectoire.
- L'explication de l'optimalité de p dépend du classifieur D mais aussi de l'univers des trajectoires possibles. Dans la définition U sera considéré comme une formule dont les modèles sont ses trajectoires.
- Une *explication commune* pour un ensemble de trajectoires optimales est une conjonction existentiellement quantifiée sans variable libre, que nous appelons par la suite un *motif relationnel*, c'est-à-dire la forme générale proposée pour une explication abductive en ordre un par P. Marquis [12].
  Nous représenterons ce qui est commun à un groupe de trajectoires optimales O par le motif relationnel maximalement spécifique au sens de la θ-

subsomption satisfait par les trajectoires du groupe, et noté $lgg(O)$ [1]. Nous définirons les explications communes comme des sous-ensembles de la lgg du groupe.

Dans la mise en oeuvre l'univers est un ensemble de trajectoires partitionné en U_{opt} (optimales) et U_{notOpt} (non-optimales) à partir desquels nous construirons par Programmation Logique Inductive (PLI) [15] un classifieur D. Une règle r de D est alors une clause de la forme $opt \leftarrow q$ qui couvre une trajectoire quand q est satisfaite. Par construction elle couvre un sous-ensemble U_r de U_{opt} et aucune trajectoire de U_{notOpt}. U_r définit donc un groupe de trajectoires optimales pour lequel on cherchera des explications communes.

Nous réduisons le problème de l'énumération des explications communes minimales de l'optimalité d'un tel groupe de trajectoires à celui de l'énumération des sous-ensembles minimaux m de la lgg du groupe tels que $opt \leftarrow m$ ne couvre aucune trajectoire de U_{notOpt}. Techniquement cela amène à résoudre deux problèmes :

1. La construction de la lgg d'un ensemble d'observations sous θ-subsomption.
2. La recherche des sous-ensembles minimaux de littéraux de la lgg qui ne θ-subsument pas un sous-ensemble d'observations donné.

Ces problèmes ont été étudiés dans différents contextes. La construction de la lgg, unique sous θ-subsomption, a été utilisée en particulier dans les méthodes de PLI ascendantes depuis l'article de Plotkin [17]. Cependant pour limiter la taille de la lgg, ce qui est impératif pour notre propos, nous construirons une approximation de celle-ci en précisant un certain nombre de contraintes déclaratives. La recherche des sous-ensembles minimaux satisfaisant une contrainte de couverture a été étudiée en fouille de données [22], mais peu abordée pour des données relationnelles pour lesquelles différentes propriétés algorithmiques essentielles ne sont pas satisfaites [9]. Les algorithmes proposés pour ces deux problèmes, en particulier le second, font partie des contributions de cet article, et sont décrits section 5.

Nous montrons en section 6 sur un cas d'étude simple, le type d'explications obtenues et le sens ou l'intérêt qu'elles ont relativement aux buts poursuivis.

2 Scenario

2.1 Le jeu

Nous considérons un jeu de cartes de type bridge simplifié avec une seule couleur dont les 13 cartes sont réparties entre les joueurs en une *donne*. Le nombre de cartes de chaque joueur dans la donne n'est pas fixé. On considère les enchère terminées par un contrat demandé par Sud et qui oppose donc le déclarant, c'est-à-dire la paire North-South (NS) au défenseur, la paire West-East (WE). Nous traitons ici du jeu de la carte commençant quand West vient de jouer la première carte du premier pli. La main de North est dévoilée sur la table, et c'est au déclarant North de jouer.

1. pour Least General Generalisation

2.2 Le scenario

Dans le scénario ci-dessous notre joueur artificiel *Noo* répond à des requêtes d'explications de ses décisions émises par son interlocuteur *M. X*. Chaque requête et ses réponses peuvent initier un ensemble d'interactions. Dans le scénario général, les mains *WE* sont inconnues du déclarant *NS*, mais dans cet article nous considérons le scénario simplifié dans lequel les mains *WE* sont connues de *Noo*. Le jeu commence quand West joue, et que les cartes de North sont dévoilées. *Noo* résout le MDP de la donne *NSEW*. On suppose que *Noo* connait le modèle, ici déterministe, du défenseur : s'il connait l'état courant du jeu, *Noo* sait la carte que jouera le défenseur *WE*. Une trajectoire partant d'un état s est optimale lorsqu'elle maximise le nombre de plis fait par le déclarant à la fin de la partie. Lorsque le MDP est résolu, dans tout état s du jeu et pour toute action possible a du déclarant en s, la q-valeur $q(s, a)$ est connue. $q(s, a)$ est ici le nombre de plis que le déclarant fera en jouant a en s et en suivant à partir de l'état s' résultant de a une trajectoire optimale. L'action a est optimale en s si elle est de q-valeur maximale. Le scénario simplifié où l'état initial est donc connu est le suivant :

1. Les cartes *WE* sont retournées.
2. *Noo* choisit une action optimale a dans l'état courant s
3. *M. X* demande s'il le souhaite "Comment l'action a conduit-elle à un nombre de plis maximal" ?"
4. *Noo* construit alors un modèle D permettant de distinguer dans l'ensemble U des trajectoires possibles commençant par (s, a), les trajectoires optimales de celles qui ne le sont pas. Il utilise D pour constituer des groupes de trajectoires optimales et propose pour chaque groupe une ou plusieurs explications communes aux trajectoires du groupe.
5. S'ensuivent des interactions entre *Noo* et *M. X* conduisant à des explications alternatives.
6. *Noo* joue a conduisant à un nouvel état courant et le scénario reprend en 2.

Dans cet article nous abordons les étapes 3 et 4 de ce scénario simplifié.

3 Notations dans les représentations relationnelles

On traite ici des langages Datalog (i.e. Logique des prédicats sans autre symbole de fonctions que des constantes). Les seuls termes sont des *constantes* et *variables*. Les constantes sont soit des nombres soit des atomes commençant par une minuscule. Par exemple, on peut représenter des cartes par des entiers de l'intervalle $[2..14]$ et les quatre joueurs par les constantes $west$, $north$, $east$, $south$. Les autres termes sont des variables, identifiées par des symboles commençant par une majuscule (X, Y, $Card$, ...). Un *littéral* est un symbole de prédicat appliqué à des termes. Un *fait* est un littéral complètement instancié (sans variables). Par exemple, le littéral $small\_card(C)$ énonce

que la variable C est une petite carte (i.e. entre 2 and 10), alors que $honor(12)$ est un fait qui déclare que 12 (représentant le valet) est un honneur. Par la suite, nous utiliserons la notation classique p/N où p est un symbole de prédicat et N est un entier indiquant le nombre d'arguments du prédicat (i.e. $small\_card/1$ indique un prédicat à un seul argument).
Dans le domaine de la PLI qui est notre cadre, on travaille traditionnellement avec deux types de formules : des *clauses définies* et des *conjonctions quantifiées existentiellement*. Une *clause* est une disjonction de littéraux quantifiées universellement $\forall[h_1 \vee \ldots \vee h_m \vee \neg b_1 \vee \ldots \vee \neg b_n]$ ou de manière équivalente $\forall[(h_1 \vee \ldots \vee h_m) \leftarrow (b_1 \wedge \ldots \wedge b_n)]$ où $h_1, \ldots, h_m$ représente une disjonction de littéraux positifs (la *tête* de la clause) et $b_1, \ldots, b_n$ est la conjonction de littéraux formant le *corps* de la clause.
Une clause *définie* est une clause avec exactement un littéral positif. Si la tête est un littéral sans argument, le corps de la clause, appelé dans la suite *motif relationnel* est la conjonction quantifiée existentiellement qui correspond au corps d'une clause. De la même manière qu'on omet classiquement le quantificateur $\forall$ lorsqu'il est clair que l'on manipule des clauses, on omettra dans la suite de de l'article le quantificateur $\exists$ lorsqu'il est clair qu'on parle d'un motif relationnel. Nous allons adapter dans la suite toutes les définitions introduites initialement pour des clauses à des motifs relationnels.
Le vocabulaire $\mathcal{V}$ d'un programme Datalog P est l'ensemble de ses constantes et symboles de prédicats. L'*univers de Herbrand* d'un programme Datalog P est l'ensemble des terms clos construit sur $\mathcal{V}$, i.e., les constantes de $\mathcal{V}$, la *base de Herbrand* est l'ensemble des faits construits sur l'univers de Herbrand et les symboles de prédicats de $\mathcal{V}$. Une *interprétation de Herbrand* de P est un sous-ensemble de la base de Herbrand de P.
La relation de généralité entre deux clauses en PLI [15] est le relation de θ-subsomption [17] que nous étendons ici à des motifs relationnels.

Définition 1. *Un motif relationnel G θ-subsume un motif relationnel S (noté $G \preceq_\theta S$) si et seulement si (ssi) il existe une substitution θ telle que $G.\theta \subseteq S$.*

Plotkin a également introduit la notion de *généralisation maximalement spécifique* ou *lgg* de deux clauses que nous reformulons ci-dessous pour des motifs relationnels.

Définition 2. *Un motif relationnel S est le* généralisé le plus spécifique *d'un ensemble de motifs relationnels O (noté par $S = lgg(O)$) si et ssi $S \preceq_\theta o_i$ pour tout $o_i \in O$ et pour tout motif G tel que $G \preceq_\theta o_i$ pour tout $o_i \in O$, alors $G \preceq_\theta S$. La* lgg *de deux motifs relationnels C et D est unique et calculée en temps $\mathcal{O}(|C||D|)$ [16].*

On appelle dans la suite *observation* une interprétation de Herbrand dans le vocabulaire $\mathcal{V}$ (voir le cadre de l'apprentissage à partir d'interprétations [4] ou d'inteprétations [9]). La relation de couverture entre un motif relationnel C et une observation o est définie par $couvre(C, o)$ ssi il existe une substitution θ telle que $C \preceq_\theta conj(o)$ où $conj(o)$ est le motif relationnel complètement instancié correspondant à o (la conjonction de tous les faits de o) et O est un ensemble d'interprétations de Herbrand étant donné $\mathcal{V}$. Par abus de langage et afin d'alléger les notations, nous identifierons l'observation o et la conjonctions associée à l'observation $conj(o)$. La lgg d'un ensemble d'observations O est définie comme $lgg(\{o_i | o_i \in O\})$ et donc en particulier $lgg(\{o\}) = o$.

Exemple 1. *Soit $\mathcal{V}$ un vocabulaire avec quatre constantes $\{1, 2, 3, 4\}$ et trois symboles de prédicats $\{p/2, r/1, q/1\}$ et deux observations o_1 et o_2, $o_1 = \{p(1, 2), r(2), p(2, 3), q(3)\}$ et $o_2 = \{p(1, 3), q(3), p(2, 4), r(4)\}$. La lgg des deux observations o_1 et o_2 est : $\exists\, p(1, X), p(X', Y'), r(Y'), p(X'', 3), q(3), p(2, Y'')$ avec les substitutions $\theta_1 = \{X/2, X'/1, Y'/2, X''/2, Y''/3\}$ et $\theta_2 = \{X/3, X'/2, Y'/4, X''/1, Y''/4\}$. Notons que la lgg de o_1 et o_2 est plus longue (en nombre de littéraux) que o_1 et o_2.*

4 Explications décisionnelles

4.1 Trajectoires et leur représentation

1. Une trajectoire est une séquence $s_0 a_0 \ldots s_t a_t \ldots s_n a_n$ où s_t est l'état observé au temps t et a_t l'action effectuée en t et conduisant en l'état s_{t+1}.
2. Nous considérerons par la suite l'arbre représentant le sous-ensemble des trajectoires possibles partant d'un état s_0 suivi d'une action a_0, s_0 étant un des états de l'arbre complet des trajectoires de la donne *NSEW* étudiée.
3. Chaque trajectoire est décrite par l'ensemble des faits vrais dans la trajectoire. La plupart des prédicats utilisés sont dits temporels : ils sont vrais ou faux selon l'instant t considéré dans la trajectoire.

Nous avons introduit des arguments de type intervalle dans une partie des prédicats décrivant les états du jeu le long d'une trajectoire. Ces intervalles représentent un segment temporel dans lequel le prédicat est vrai entre ses bornes et faux à l'extérieur. Nous nous plaçons ainsi dans la lignée de travaux classiques de représentation des intervalles [1] et de fouille de données séquentielles [14].
Dans ce cas, pour un atome a donné, instancié sauf pour sa partie temporelle, nous divisons le domaine temporel $t_0, \ldots t_n$ en intervalles durant lesquels a est vrai. L'atome $a([b, e])$ est alors vrai lorsque a est vrai pour tout $t \in [b, e]$ et faux aux temps $b - 1$ (si $b \neq 0$) et $e + 1$ (si $e \neq n$).

Exemple 2. *Soit une trajectoire où a est vrai aux temps $1, 2, 3, 5, 6$ et b est vrai aux temps $2, 3, 4, 6, 8, 9$. La trajectoire s'écrit $a([1, 3]), a([5, 6]), b([2, 4]), b([6, 6]), b([8, 9])$.*

De tels prédicats sont utilisés dans l'étude de cas en section 6.

4.2 Explications

Les *explications abductives* ont récemment été définies pour rendre compte de l'étiquette attribué à une observation o par un arbre de décision [10, 2, 3, 6]. Dans ces travaux l'observation est représentée par un ensemble ou une conjonction de paires attribut-valeur. Une explication abductive minimale du classement de o par l'arbre de décision D est définie comme une conjonction incluse dans o et suffisante pour classer o avec l'étiquette c. Dans notre contexte nous devons changer et étendre cette définition de plusieurs manières. En particulier chaque observation o est une interprétation de Herbrand d'un langage Datalog, représentée par une conjonction de littéraux instanciés (voir la section 3). Le classifieur D est alors une formule de ce langage. L'explication d'un classement peut alors être représentée comme une clause instanciée [19]. Cependant pour notre définition des explications communes, nous reprenons plutôt ici la définition générale d'une explication abductive en ordre un proposée par P. Marquis [12], sous la forme d'un motif relationnel et étudiée par la suite d'un point de vue opérationnel [11, 8].
Nous considérons de plus que l'ensemble des observations possibles dans le problème en cours ne forme qu'une partie, appelée l'*univers*, des interprétations syntaxiquement correctes. Dans les définitions qui suivent l'univers est représenté par les modèles d'une formule U. Notre connaissance U, D du problème courant est alors divisé en une partie U restreignant les observations o autorisées et une partie D permettant d'inférer le label d'une observation. Ceci nous amène à la définition suivante ajoutant U à la définition usuelle d'une explication abductive de l'inférence d'une étiquette pour une observation :

Définition 3 (Explication minimale de l'étiquette d'une observation)**.** *Un explication de l'étiquette c d'une observation o relativement au classifieur D et à l'univers représenté par U est une conjonction $e \subseteq o$ telle que $e, U, D \models c$.*
Si pour tout $e' \subset e$ nous avons $e', U, D \not\models c$ alors e est une explication minimale de l'étiquette c de o relativement à D et U.

Dans l'exemple ci-dessous nous illustrons les explications minimales et montrons que, comme attendu, on obtient des explications minimales différentes selon que l'on utilise ou non l'univers des observations possibles.

Exemple 3. *Soient les étiquettes $+$ et $-$. D est une ensemble de clauses définies et nous considérons que si $o, D, U \not\models +$ alors D attribue l'étiquette $-$ à o.*
Soient p et r deux prédicats unaires et le domaine de constantes $\{1, 2\}$ formant la base de Herbrand $\{p(1), p(2), r(1), r(2)\}$. Une observation est notée comme un sous-ensemble de cette base. Soient le classifieur D et l'observation o_1 définis comme suit :
- *$o_1 = \{p(1), p(2), r(1)\}$*
- *$D = \{+ \leftarrow p(X), r(X); + \leftarrow p(2)\}$*

Considérons d'abord que toute observation est dans l'univers, c'est-à-dire $U = Vrai$. Les explications minimales de l'étiquette $+$ attribuée à o_1 sont alors $p(1), r(1)$ et $p(2)$. Ceci résulte de ce que nous avons $p(X), r(X).\{X/1\} = p(1), r(1) \subseteq o_1$, et $p(2) \subseteq o_1$.
Considérons maintenant l'univers représenté par $U = \{p(X) \leftarrow r(X); p(1) \vee p(2)\}$. Comme toute observation doit satisfaire U, de $r(1)$ et U nous déduisons $p(1)$ et de $p(1), r(1), D$ nous déduisons $+$. En conséquence $p(1), r(1)$ n'est plus une explication minimale du classement de o_1 car car $r(1)$ explique également l'étiquette.

Nous définissons maintenant une *explication commune* de l'étiquette c commune à un ensemble d'observations O comme un motif relationnel. Pour cela nous cherchons d'abord ce qui est commun à ces observations, sous la forme du généralisé le plus spécifique $lgg(O)$, puis considérons les explications communes comme des motifs relationnels inclus dans $lgg(O)$.

Définition 4. *Une explication commune du classement c des observations de O est un motif relationnel e inclus dans $lgg(O)$ et tel que $e, U, D \models c$.*
Si pour tout $e' \subset e$ on a $e', U, D \not\models c$, e est appelée une explication commune minimale de o.

Notons que par définition, pour tout o et toute explication commune e, e est plus générale que o c'est-à-dire $e \preceq_\theta o$.

Exemple 4. *Ajoutons à l'exemple 3 une deuxième observation $o_2 = \{p(1), p(2), r(2)\}$. On obtient $lgg(\{o_1, o_2\}) = p(1), p(2), r(X)$ avec $lgg(O).\{X/1\} \subseteq o_1$ et $lgg(O).\{X/2\} \subseteq o_2$. On observe que de $r(X), U, D$ on peut dériver l'étiquette. En effet, de U et $r(X)$ on déduit $p(X)$ et de $r(X), p(X)$ et D on déduit $+$. Par ailleurs, comme vu précédemment de $p(2)$ et D on déduit $+$. Les explications communes minimales de $\{o_1, o_2\}$ sont $r(X)$ et $p(2)$.*

On a également la propriété suivante :

Proposition 1. *Si e est une explication commune minimale du classement de O, alors pour tout $o \in O$ et pour toute substitution θ telle que $e.\theta \subseteq o$ alors $e.\theta$ est une explication, non nécessairement minimale, du classement de o.*

Dans ce qui suit nous considérons que l'univers est connu à travers l'ensemble de ses modèles $M(U)$ qui est partitionné selon les étiquettes attribuées par D en $\{U_c \mid c \in C, U_c \subseteq M(U)\}$. Dans ce cas, une explication commune e de l'ensemble d'observations $O \subseteq U_c$ est telle que e ne couvre aucune observation de label différent, c'est-à-dire appartenant à $U_{\neg c} = \bigcup_{d \in C \setminus c} U_d$. On peut alors directement construire les explications communes minimales sans utiliser le classifieur :

Proposition 2. *$e \subseteq lgg(O)$ est une explication commune minimale du classement des observations de O par D en c si et seulement si $\forall u \in U_{\neg c}$ e ne couvre pas u et $\forall e' \subset e, \exists u \in U_{\neg c}$ t.q. e' couvre u.*

Exemple 5. *Nous poursuivons l'exemple 4. L'univers contient 8 observations modèles de U parmi lesquelles*

seule $o_- = \{p(1)\}$ est dans U_-. Nous obtenons alors les explications communes minimales de $\{o_1, o_2\}$, c'est-à-dire $r(X)$ et $p(2)$, comme sous-ensembles minimaux de $p(1), p(2), r(X)$ ne couvrant pas o_-.

Remarquons qu'une explication commune d'un ensemble d'observations quelconque O ayant même label n'existe pas nécessairement. Cependant, lorsque le classifieur D est un ensemble de clauses définies, la lgg de la couverture O d'une clause $c \leftarrow b$, c'est-à-dire l'ensemble des observations $O \subseteq U_c$ couvertes par la clause, est moins générale que b. En conséquence $lgg(O)$ ne couvre aucune observation de $U_{\neg c}$ et est donc une explication commune de O. Les corps des clauses de D formant un recouvrement de U, on en déduit le résultat suivant :

Proposition 3. *Pour toute observation o de U il existe au moins une explication minimale commune à la couverture d'une clause de D couvrant o.*

Nous exploiterons dans les suivantes ces résultats et définitions en faisant les remarques suivantes :

1. Si on ne dispose pas d'un classifieur , mais que l'on connait les labels des observations de U on peut en construire un par PLI dont on supposera qu'il ne commet pas d'erreurs sur $M(U)$.
2. Les explications communes minimales d'un sous-ensemble d'observations O dépendent de la clause de D dont la couverture est O et de $U_{\neg c}$.
3. Par univers on entend ici l'ensemble des observations possibles, mais cet univers est contextuel : Dans le cas des trajectoires U représente le sous-ensemble des trajectoires possibles à un moment du jeu, donc un sous-arbre de l'arbre de jeu.

5 Construction d'explications

5.1 Approximation de la généralisation maximalement spécifique d'un sous-ensemble de trajectoires

La complexité du calcul de la *lgg* exacte pour un ensemble d'observations O est prohibitive, dans le pire des cas en $\mathcal{O}(C)^n$ où C est la taille (en nombre de littéraux) de la plus grande observation de O et $n = |O|$. Sans détailler l'algorithme de Plotkin, pour un symbole de prédicat donné p et étant donné deux observations o_1 et $o_2 \in O$, l'algorithme naïf de calcul de la *lgg* calcule pour chaque symbole de prédicat p et pour toutes les paires de littéraux $\in selection(p, o_1) \times selection(p, o_2)$ un littéral maximalement spécifique qui θ-subsume chacune de ces paires. Considérant que dans notre étude de cas, les observations ont un nombre moyen de littéraux autour de 260, certains symboles de prédicats ayant un nombre d'instances moyen par observation autour de 10, il n'y a pas moyen d'obtenir une *lgg* en temps raisonnable.
Nous avons donc adapté pour ce faire un algorithme descendant de type *générer et tester* qui permet de calculer un sous-ensemble de $lgg(O)$ appelé $Bottom$ dans la suite de l'article (voir Algorithme 1). Cet algorithme prend en compte des contraintes pertinentes pour notre problème et de nature à borner la taille de $Bottom$.
Nous définissons dans un premier temps et de façon classique en PLI un biais de langage $\mathcal{B}$ comme une liste de symboles de prédicats associés aux types de leurs arguments. Ce biais de langage permet de construire un motif relationnel $Bottom$ (et des explications associées) pour un sous-ensemble des prédicats de $\mathcal{V}$.
Dans les sections suivantes, et comme dans [9], on représente les motifs relationnels comme des listes d'atomes, i.e. $[l_1, ..., l_n]$. Si C et un motif relationnel et l un littéral, on représente par $[C, l]$ le motif relationnel obtenu en ajoutant l après le dernier littéral de C. Etant donné une graine $g \in O$ (choisie aléatoirement dans notre cas), et pour chaque symbole de prédicat $p \in \mathcal{B}$, l'algorithme 1 calcule pour tous les littéraux l_i instances de p dans g les littéraux généralisés candidats de façon générer et tester (noté par $\rho(Bottom, l_i)$),

Algorithm 1 Calcule $Bottom$, approximation de *lgg(O)*

Require: O un ensemble d'observations, $\mathcal{B}$: un biais de langage , une fonction de score $Score$
Ensure: Un motif relationnel $Bottom$ qui θ-subsume toutes les observations de *O*, et de taille $\leq k * |g|$

```
g ← random_choice(O) ; Bottom ← ∅
for each p ∈ B do
    for each l_i instance de p dans g (l_i ∈ g) do
        Cands ← k-meilleurs de ρ(Bottom, l_i) pour Score
        for each lg_i ∈ Cands do
            if [Bottom, lg_i] couvre toutes les observations de O then
                Bottom ← [Bottom, lg_i]
            end if
        end for
    end for
end for
Bottom ← reduce(Bottom)
return Bottom
```

Cet opérateur ρ prend en argument la généralisation courante $Bottom$ et le littéral courant l_i pour engendrer des littéraux généralisés candidats à l_i. Par exemple, ρ peut contrôler pour les littéraux généralisés de l_i le nombre de variables, le nombre maximum de nouvelles variables (n'apparaissant pas déjà dans $Bottom$), etc. A cette étape, certains littéraux de $\rho(Bottom, l_i)$ peuvent être en relation de généralité (voir exemple). Le cardinal de $\rho(Bottom, l_i)$ pouvant être élevé, l'algorithme sélectionne les k meilleurs candidats pour la fonction *Score*. Un littéral $lg \in \rho(Bottom, l_i)$ intégre $Cands$ si $[Bottom, lg]$ couvre toutes les observations de O. La fonction de score définit les critères de tri permettant de sélectionner les k meilleurs candidats. Elle prend en compte différents éléments syntaxiques (par exemple, le nombre d'occurrences des variables dans $[Bottom, lg]$). On peut également si nécessaire dès cette étape sélectionner en priorité les littéraux candi-

dats permettant de rejeter des observations de $U_{\neg c}$. Une fois les k-meilleurs littéraux généralisés candidats pour l_i identifiés, ils sont ajoutés de façon gloutonne à $Bottom$ si $[Bottom, lg]$ couvre toutes les observations de O.
$Bottom$ est finalement réduite [17] pour ne garder *in fine* que les littéraux maximalement spécifiques.

Exemple 6. *Reprenons l'exemple 1 et calculons le motif relationnel $Bottom$ associé aux deux observations o_1 et o_2. On suppose que ρ n'engendre que les littéraux généralisés qui contiennent au plus une variable (nouvelle ou apparaissant déjà dans $Bottom$). Commençons par le symbole de prédicat p le plus fréquent dans les deux observations et supposons que o_1 soit la graine. Soit la première instance de p dans o_1, $p(1,2)$. $\rho(Bottom, p(1,2)) = \{p(1,2), p(X,2), p(1,Y)\}$. $p(1,2)$ ne couvre pas o_2, $p(X,2)$ et $p(1,Y)$ couvrent tous les deux o_1 et o_2 et sont ajoutés à $Bottom$. La substitution θ pour o_1 devient $\{X/1, Y/2\}$. $\rho(Bottom, p(2,3)) = \{p(2,3), p(Y,3), p(Y',3), p(2,Z)\}$. p(2,3) ne couvre pas o_2, $p(1,Y), p(Y,3)$ ne couvre pas o_2, on ajoute donc à Bottom $p(Y',3), p(2,Z)$ et $\{Y'/2, Z/3\}$ à θ. De la même manière, on ajoutera à Bottom $r(Y)$ et $q(3)$. On obtiendra finalement $Bottom = \{p(X,2), p(1,Y), p(Y',3), p(2,Z), r(Y), q(3)\}$, qui n'a pas besoin d'être réduite et qui est la lgg exacte de o_1 et o_2, avec la substitution $\theta = \{X/1, Y/2, Y'/2, Z/3\}$ pour la graine o_1.*

Le motif relationnel $Bottom$ est la borne inférieure de l'espace de recherche pour les explications communes minimales de O, que nous décrivons dans la section suivante.

5.2 Construction d'explications dynamiques

L'algorithme 1 construit un motif relationnel $Bottom$ qui est une approximation de la $lgg(O)$ (i.e., $Bottom \preceq_\theta lgg(O)$). Nous construisons maintenant des explications communes à O comme des sous-ensembles (au sens de $\subseteq$) corrects de $Bottom$ (voir def. 4). Le premier système de fouille de motifs relationnels est Warmr [7], qui construisait des motifs relationnels fréquents appartement à un biais de langage. Depuis, d'autres travaux se sont également intéressés à la fouille de motifs relationnels, notamment clos [9]. On souhaite ici résoudre un problème un peu différent, celui d'extraire de $Bottom$ – motif relationnel – l'ensemble de ses sous-ensembles minimaux (sous $\subseteq$) et corrects. Par correct, nous entendons que si toutes les observations de O ont la classe c (dérivée grâce au classifieur D), alors les motifs relationnels recherchés ne devront couvrir aucune observation de l'ensemble $U_{\neg c}$ des observations associées à une classe autre que c. Si un motif relationnel m couvre de telles observations, nous nommerons cet ensemble d'observations *l'ensemble critique* de m. Dans la suite, nous appellerons ces observations *critiques*. En d'autres termes, nous cherchons à identifier la borne G d'un espace des versions dont la borne inférieure est $Bottom$ [13], le tout dans un langage relationnel.
Cette tâche a été étudiée dans le cadre le la fouille de données booléennes [21]. Dans cet article, les auteurs proposent un algorithme descendant dans lequel un motif mg est spécialisé en ajoutant un attribut booléen l si $[mg, l]$ rejette au moins une observation critique de mg. L'ajout de l à mg est effectué si aucun des sous-ensembles de $[mg, l]$ qui contient l ne rejette les mêmes observations critiques.
Nous adaptons dans l'article cet algorithme à la recherche de motifs relationnels sous-ensembles (pour $\subseteq$) de $Bottom$ et corrects. Cette adaptation n'est pas triviale pour les raisons suivantes. Lorsque l'on spécialise un motif relationnel mg afin de rejeter de nouvelles observations critiques, il peut être nécessaire d'ajouter pour ce faire des littéraux qui ne permettent pas dans un premier temps d'écarter de nouvelles observations critiques mais qui introduisent de nouvelles variables (objets) portant une information discriminante (voir [18] pour une des premières discussions à ce sujet). De multiples stratégies de *lookahead* ont été proposées dans ce cadre toutes fondées sur des biais de recherche *ad-hoc*. Nous proposons ici une stratégie originale qui s'appuie sur la structuration de $Bottom$ en *locales*.

Définition 1. *(d'après [5]). Soit $VarBottom$ l'ensemble des variables de $Bottom$, soient X_1 et X_2 deux variables de $VarBottom$. X_1* touche *X_2 si les deux variables apparaissent dans le même littéral, et X_1* influence *X_2 si elle touche X_2 ou si elle touche au moins une variable X_i qui influence X_2. La* locale *d'une variable X est l'ensemble des littéraux qui contiennent X ou au moins une variable influencée par X.*

Intuitivement, une locale est un ensemble maximal de littéraux qui partagent des variables. Les instanciations possibles d'une variable sont donc contraintes uniquement par des variables de sa locale. Un fait est une locale de taille un.

Exemple 7. *La lgg de l'exemple 1 contient 5 locales $\{p(1,X)\}$, $\{p(X',Y'), r(Y')\}$, $\{p(X'',3)\}$, $\{q(3)\}$, $\{p(2,Y'')\}$*

On note dans la suite par $couverture(m, O)$ où m est un motif relationnel et O un ensemble d'observations l'ensemble des observations $o_i \in O$ telles que $couvre(m, o_i)$. Les algorithmes 2, 3 et 4 étendent l'algorithme décrit dans [21] au calcul de motifs relationnels corrects. L'algorithme 4 recherche dans une locale un sous-ensemble minimal de littéraux qui permet de rejeter au moins une observation de l'ensemble critique de mg. Un tel sous-ensemble de littéraux peut-être géré comme un attribut booléen dans l'algorithme de [21]. L'algorithme 3 permet d'identifier quand la recherche doit s'arrêter : soit mg est correct (l'ensemble critique de mg, NCE est vide), soit il n'existe plus de solution si la formule maximale atteignable n'est pas correcte (ligne 2 de l'algorithme 3).
Par définition de $Bottom$, mg couvre toutes les observations de O, et donc de tout sous-ensemble de O. Comme mg et LK ne partagent aucune variable (par définition d'une locale), si LK_i et LK_j sont deux locales de $Bottom$, $i \neq j$, $couverture([LK_i, LK_j], NCE) = couverture(LK_i, NCE) \cap couverture(LK_j, NCE)$.

Exemple 8. *Toujours dans le cadre de l'exemple 1, supposons que o_1 et o_2 soient de classe c et qu'il n'existe qu'un*

Algorithm 2 Common_Explanations($Bottom$,$U_{\neg c}$)

Ensure: MG : ensemble de sous-ensembles minimaux et corrects de $\subseteq$ $Bottom$ (explications communes pour tout $o_i \in O$)

1: $LLK \leftarrow compute\_locales(Bottom)$
2: **return** $mings(\emptyset, LLK, U_{\neg c})$

Algorithm 3 $mings(mg, LLK, NCE)$

Require: mg : motif minimal courant, NCE : ensemble critique de mg, LLK : locales à explorer

Ensure: MGF : ensembles des motifs minimaux et corrects de $Bottom$

1: $MGF \leftarrow \emptyset$
2: **if** $mg \cup LLK$ est correct **then**
3: **if** mg est correct ($NCE = \emptyset$) **then return** mg
4: **end if**
5: $LK \leftarrow head(LLK)$
6: $nLLK \leftarrow tail(LLK)$
7: **for** tous les $M = maxGen(LK, NCE)$ **do**
8: ▷ voir alg.4
9: $emg \leftarrow [mg, M]$
10: $rNCE \leftarrow couverture(emg, NCE)$
11: $MGLK \leftarrow mings(emg, nLLK, rNCE)$
12: $MGF \leftarrow MGF \cup MGLK$
13: **end for**
14: $MGWLK \leftarrow mings(mg, nLLK, NCE)$
15: $MGF \leftarrow check\_min(MGF \cup MGWLK)$
16: **end if**
17: **return** MGF

Algorithm 4 $maxGen(LK, NCE)$

Require: LK : une locale de $Bottom$, NCE : observations critiques de mg

Ensure: MG ensembles des minimaux de LK qui rejettent au moins une obs. critique de NCE

1: **if** LK ne rejette aucun exemple de NCE **then**
2: **return** $\emptyset$
3: **end if**
4: $MG \leftarrow \emptyset$
5: **for** pour tous les sous-ensembles mg_i de LK **do**
6: $rNCE \leftarrow couverture(mg_i, NCE)$
7: **if** $|rNCE| < |NCE|$ et mg_i est minimal **then**
8: $MG \leftarrow MG \cup mg_i$
9: **end if**
10: **end for**
11: **return** MG

exemple o_3 de classe $\neg c$: $\{p(2,4), r(2), p(2,3), q(3)\}$. Si on ordonne les locales par rejet des observations de classe $\neg c$, $\{q(3)\}$, $\{p(2,Y'')\}$ et $\{p(X'',3)\}$ sont immédiatement rejetées. $\{p(1,X)\}$ et $\{p(X',Y'), r(Y')\}$ sont en revanche chacune des minimaux corrects ($p(X',Y')$ et $r(Y')$ couvrent o_3 alors que $p(X',Y'), r(Y')$ rejette o_3 et est donc un minimal correct).

Proposition 4. *L'algorithme 2 est correct et complet.*

La preuve est fournie en annexe [2].

6 Exemple développé

Dans cette section nous considérons une donne du jeu simplifié et deux ensembles de trajectoires. Dans chaque cas nous calculons

— un ensemble de clauses D couvrant les trajectoires optimales et aucune trajectoire non-optimale [3].
— pour chaque clause l'ensemble des explications communes minimales des trajectoires couvertes par la clause. Pour certaines explications nous donnons une preuve informelle de ce qu'elle implique l'optimalité des trajectoires.

La donne est la suivante

```
W 8 9 11      N 3 4 5 12   E 6 7      S 2 10 13 14
```

Le problème de décision du déclarant posé au joueur artificiel Noo est considéré comme un MDP déterministe. Les actions sont les cartes jouées par le déclarant en North ou South aux différents instants du jeu. La transition d'un état du jeu au suivant est déterministe et connue de Noo ce qui signifie en particulier que Noo a un modèle de la défense : il sait à tout moment du jeu quelle carte jouera le défenseur East ou West. On décrira une trajectoire par les cartes jouées par les déclarants, c'est-à-dire les actions, et celles jouées par les défenseurs permettant les transitions d'un état au suivant. La récompense obtenue $r(s,a)$ quand le déclarant joue une carte a à l'instant t dans l'état s est 1 si le pli courant est gagné entre t et $t+1$. Chaque instant t marque le moment où le déclarant (North ou South) doit jouer.

Exemple 9. *Nous considérons le début d'une trajectoire commençant en l'état initial s_1 après que West ait joué le 8 au début du premier pli :*
W8 (t1) N12 E6 (t2) S2 (t3) S13 W9
Quand North joue le 12 en t1 la récompense r(s1,12) est nulle mais on a r(s2,2) =1 car après que South ait joué le 2 le déclarant a fait le premier pli.

On peut associer à chaque trajectoire la récompense totale qui lui est associée, c'est à dire le nombre de plis fait par le déclarant à la fin du jeu.

2. L'annexe est consultable à l'adresse suivante : `https://github.com/Malickick/PFIA23-trajectoires/blob/main/annexe_PFIA_23_trajectoires.pdf`
3. en utilisant cLear, un programme d'apprentissage relationnel développé par NukkAI

Notons que si dans l'état courant s le déclarant dont c'est le tour de jouer a plusieurs cartes *consécutives* dans sa main, jouer l'une ou l'autre ne change pas le déroulement de la trajectoire. Formellement deux cartes a et a' dans la main d'un joueur à l'état s sont s-consécutives si il n'existe pas en s de carte a'' dans la main d'un autre joueur telle que $a < a'' < a'$ ou $a' < a'' < a$ Par exemple que North joue 5 ou 3 en s_1 n'affecte pas la suite du jeu. Par la suite dans l'illustration d'un sous-arbre de trajectoires de racine s on considère équivalentes les cartes s-consécutives pour North ou South. L'arbre représenté est alors un arbre de trajectoires dites *abstraites* dans lequel on a un seul arc pour plusieurs actions équivalentes. Nous ne représentons que la partie *utile* de la trajectoire se terminant en t_7, temps où le défenseur n'ayant plus de cartes, l'optimalité ou non de la trajectoire st acquise.

6.1 North joue son honneur en t_1

Ci-dessous nous considérons le sous-arbre des trajectoires commençant en l'état s_2 après que North ait joué 12 (action N12) et East ait répondu 6. L'action N12 conduit a une q-valeur optimale de 3 plis et nous cherchons ici à répondre à la question "comment atteindre cette q-valeur optimale à partir de l'état s_2 ?". L'ensemble des trajectoires optimales se représente par deux trajectoires abstraites qui ne diffèrent que par la carte 2 ou 10 jouée par South au temps t_2 et que nous représentons de manière compacte ci-dessous :

W8	N12	E6	S(2 ou 10)
N3-4-5	E7	S13-14	W9
S13-14	W11	N3-4-5	E0

On remarque que dans ces deux trajectoires abstraites optimales South joue une petite carte (≤ 10) au premier pli remporté par North et ses deux honneurs 13 (Roi) et 14 (As) aux plis 2 et 3. Ici l'univers est l'ensemble des chemins du sous-arbre de racine s_2, composé des 24 trajectoires optimales de U_{opt}, résumées en les deux trajectoires abstraites ci-dessus et des 120 non optimales de U_{notOpt}. On construit à l'aide de *cLear* un modèle pour la cible *optimale* à partir de U et on obtient une clause unique :

$$opt \quad \leftarrow \quad nbSmallCards(1, B, [3, 7]).$$

Tous les prédicats utilisés sont définis dans l'annexe. Cette clause permet de dériver *opt* pour toute trajectoire commençant en t_2 après W8 N12 et dit qu'il existe un joueur B dont le nombre de cartes de valeur inférieure ou égale à 10 atteint 1 en t_3 puis reste constant jusqu'à la fin de la trajectoire utile au temps t_7.
On construit alors $lgg(U_{opt})$ (Algorithm 1) et on cherche les explications communes minimales de U_{opt} (Algorithm 2). Parmi celles-ci on trouve les deux explications communes minimales suivantes dont nous montrons l'optimalité.

1. $x_1 = nbSmallCards(1, south, [3, 7])$
   Le corps x_1 d'unique clause du modèle est complètement instancié et est une explication commune minimale des 24 trajectoires couvertes par la clause. Pour montrer qu'elle conduit à l'optimalité on raisonne sur la partie en cours. x_1 implique que South a joué une petite carte en t_2, puisque en t_3 son nombre de petites cartes a diminué, et c'est donc North qui fait le premier pli avec le 12. Puis, ce nombre ne diminuant plus jusqu'au 4ème pli, South joue ses deux honneurs 13 et 14 aux plis 2 et 3, qu'il remporte (East et West n'ont pas de cartes plus fortes). Le déclarant remporte donc les 3 plis et la trajectoire est optimale.

2. $x_2 = \exists B\ action(B, 5), honor(B)$ qui dit que le déclarant joue un honneur en t_5 début du 3ème pli.
   L'explication commune alternative x_2 conduit l'interlocuteur à un raisonnement plus élaboré pour conclure à l'optimalité. North n'ayant plus d'honneur en t_5 c'est forcément South qui joue l'honneur B (donc 13 ou 14) en t_5 et fait le pli 3. South jouant en t_5 cela implique que South ou East a gagné le second pli : en effet si North ou West gagnait le pli 2, South jouerait au pli 3 en t_6 et non en t_5. Comme au début du pli 2 l'unique carte de East (le 7) est plus petite que celles de West (9 ou 11) East ne peut pas faire ce pli et c'est donc South qui a gagné le pli 2. Finalement, North ayant fait le premier pli (avec le 12) et South les deux suivants, la trajectoire est optimale.

6.2 North joue une petite carte en t_1

Si au temps t_2 North joue une petite carte, ici le 3 (mais 3 4 et 5 sont équivalentes car s_1-consécutives) il y a plusieurs trajectoires abstraites optimales représentées Figure 1 dans un arbre dont les feuilles, numérotées de 1 à 10 à partir de la gauche, représentent les trajectoires abstraites optimales commençant par W8 N3. Les actions non optimales, possibles mais conduisant à des trajectoires non optimales, sont marquées par un arc conduisant à une feuille vide.

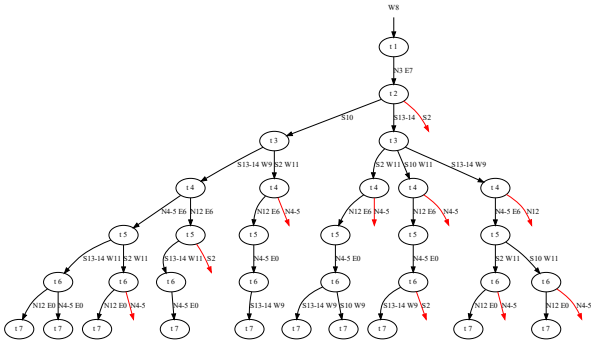


FIGURE 1 – Arbre de trajectoires abstraites optimales

On obtient cette fois quatre clauses pour le label *opt* par apprentissage sur les 40 trajectoires optimales et 104 non-

optimales :

$$opt \leftarrow playSmallestCard(C, south, 3), \\ willTakeTrickWithDominant(12, north, F). \quad (1)$$

$$opt \leftarrow action(A, 12, 6), nbSmallCards(A, 1, B, [1, 3]). \quad (2)$$

$$opt \leftarrow nbHonors(A, 1, B, [4, 5]). \quad (3)$$

$$opt \leftarrow nbThreats\_h(A, 2, B, 0, [7, 7]). \quad (4)$$

Les trajectoires abstraites couvertes par les clauses sont dans l'arbre respectivement aux feuilles 5,6,7 pour la clause 1, aux feuilles 1,3,9,10 pour la clause 2, aux feuilles 1,2,4 pour la clause 3, et à la feuille 8 pour la clause 4. Avant de détailler quelques explications associées à la clause 3, notons qu'on trouve comme explication associée à la clause (1) une instanciation du corps de la clause avec $C = 2$ et $F = 4$: les clauses apprises sont maximalement générales, mais ce n'est pas le cas des explications qui sont des sous-ensemble de la lgg. Même chose pour la clause (2) où le joueur ayant un honneur entre t_4 et t_5 est $B = West$. Parmi les explications communes minimales de la clause 3 on trouve :

1. $nbHonors(1, south, [4, 5])$. Il s'agit du corps de la clause (3). Il dit qu'il y a 1 honneur à South exactement entre t_4 et t_5. On sait donc a) l'un des honneurs (13-14) est donc joué en t_3 début du 2ème pli et b) l'autre en t_5 début du 3ème pli. De a) on déduit que South a fait le premier pli (East n'a pu gagner ce pli) et le second (13 et 14 sont dominants dès le début). De b) on déduit qu'il a fait le 3ème.

2. $action(10, 2), action(13, B), action(14, C)$. dit que South joue le 10 au premier pli, puis le 13 et le 14 aux deux autres plis de la trajectoire, la trajectoire est donc optimale. Ce qui donne un plan de jeu très directif au déclarant et est d'une interprétation bien plus simple que la clause (3) elle-même.

3. $maxCardHand(A, 2, south, [6, 7])$, $nextDominant\_h(A, 12, north, [B, C]), geq(C, 4)$. Le premier atome dit qu'à partir de t_6 la plus grande carte de South est le 2 et donc que South a déjà joué le 10, le 13 et le 14 et a donc fait en particulier le pli 1. L'un de ces cartes est jouée en t_5 donc South a fait le pli précédent, le pli 2. Le deuxième et troisième atomes disent que 12 est la seconde carte dominante de North dans un intervalle se finissant en un $t \geq t_4$. On en déduit qu'en t_4 South n'a pas encore joué son deuxième honneur (13 ou 14), il jouera donc cet honneur en t_5 et fera le pli 3. Le déclarant ayant remporté les 3 plis la trajectoire est optimale.

La valeur de ces explications dépend de leur propos. S'il s'agit de voir la suite d'actions résumant ces trajectoires favorables, l'explication 2, la plus directe convient. Si le but est pédagogique, on peut voir les explications 1 et 3 comme des exercices poussant un apprenti-joueur à raisonner sur les actions de jeu et leurs conséquences.

7 Conclusion

Nous avons proposé dans cet article une formalisation et une proposition fondée sur des outils de Programmation Logique Inductive de la notion d'explication de l'étiquette partagée par un groupe d'exemples. Nous utilisons, de même que [20], la notion de voisinage de l'observation à expliquer (défini comme l'ensemble des observations couvertes par une même règle du classifieur D) en l'adaptant au cadre relationnel. Contrairement à [10], nous nous appuyons fortement sur les exemples dont nous disposons pour construire les explications, et pour ce faire, nous avons recours largement à des opérations d'apprentissage. Nous avons proposé une formalisation logique de la notion d'explications en ordre un d'un groupe d'observations et nous avons conçu et implémenté des algorithmes de calcul d'une borne inférieure de l'espace des explications qui s'approche d'un moindre généralisé du groupe à expliquer. Nous avons également proposé un algorithme de calcul de motifs minimaux relationnels à partir de cette borne inférieure. Enfin, nous avons pu mener de premiers tests de la méthode sur des problèmes d'analyse de trajectoires d'un jeu de Bridge simplifié.
Ces premières expérimentations nous ont montré que les explications calculées ici définissent a minima la notion intuitive d'explication : on suppose l'interlocuteur omniscient, et connaissant donc la sémantique du langage utilisé, et l'univers U des observations possibles et leurs labels (via la classifieur D). Dans la mesure où on lui fournit un ensemble de faits extraits d'une observation o suffisants, sachant U, pour en déduire le label, on suppose qu'il est en mesure de faire un travail d'interprétation/validation de cette explication, qui lui fournit des informations minimale nécessaire pour dériver le label de cette trajectoire.
Une première perspective de ce travail est la construction d'une théorie reliant les prédicats du langage, incluant une version intensionnelle de l'univers U. On pourra alors faire des déductions à partir des explications. Nous avons en particulier montré informellement dans notre étude de cas comment on conclut à partir d'une explication à l'optimalité d'un groupe de trajectoires. Par ailleurs les explications fournies peuvent être nombreuses et partiellement redondantes selon le langage et les contraintes de score utilisées pour le calcul de $Bottom$. Nous pensons utiliser une partie de cette théorie dans le calcul d'une distance entre explications et garantir ainsi une certaine diversité des explications.
Une perspective majeure est le passage au cas où le joueur artificiel ne connait pas les mains de l'adversaire. Du point de vue de l'optimalité on passe de la résolution d'un MDP à celle d'un MDP partiellement observé, donc de trajectoires de récompense maximale à des trajectoires dont l'espérance de la récompense totale est maximale. La question de ce qu'est une explication dans ce cadre est ouverte.

Remerciements

Nous tenons à exprimer notre gratitude envers NukkAI et l'ANRT pour le soutien et la confiance qu'ils ont placés en notre recherche. Egalement, nous tenons à remercier Domi-

nique Bouthinon pour son élaboration du programme cLear, qui a permis de réaliser les expérimentations présentées dans cet article, et Junkang Li qui a élaboré le système que nous avons utilisé afin de résoudre le MDP. Nous adressons enfin nos remerciements aux relecteurs de CNIA pour leurs précieuses suggestions et commentaires qui ont contribué à améliorer cet article.

Références

[1] Allen, J.F. : Maintaining knowledge about temporal intervals. Commun. ACM **26**(11), 832–843 (1983)

[2] Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J., Marquis, P. : On preferred abductive explanations for decision trees and random forests. In : Raedt, L.D. (ed.) Proceedings of IJCAI'22. pp. 643–650 (2022)

[3] Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J., Marquis, P. : Sur le pouvoir explicatif des arbres de décision. In : EGC. vol. E-38, pp. 147–158. Editions RNTI (2022)

[4] Blockeel, H., Raedt, L.D., Jacobs, N., Demoen, B. : Scaling up inductive logic programming by learning from interpretations. DMKD'99 **3**, 59–93 (1999)

[5] Cohen, W.W., Jr., C.D.P. : Polynomial learnability and inductive logic programming : Methods and results. New Gener. Comput. **13**(3&4), 369–409 (1995)

[6] Darwiche, A., Hirth, A. : On the reasons behind decisions. In : L (ed.) ECAI'20. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 712–720 (2020)

[7] Dehaspe, L. : Frequent pattern discovery in first-order logic. AI Commun. **12**(1-2), 115–117 (1999)

[8] Echenim, M., Peltier, N. : A calculus for generating ground explanations. In : IJCAR'12. vol. 7364, pp. 194–209. Springer (2012)

[9] Garriga, G.C., Khardon, R., Raedt, L.D. : Mining closed patterns in relational, graph and network data. Ann. Math. Artif. Intell. **69**(4), 315–342 (2013)

[10] Huang, X., Izza, Y., Ignatiev, A., Marques-Silva, J. : On efficiently explaining graph-based classifiers. In : Proceedings of KR'21. pp. 356–367 (2021)

[11] Inoue, K. : Consequence-finding based on ordered linear resolution. In : IJCAI'91. pp. 158–164. Morgan Kaufmann (1991)

[12] Marquis, P. : Extending abduction from propositional to first-order logic. In : FAIR. vol. 535, pp. 141–155. Springer (1991)

[13] Mitchell, T.M. : Generalization as search. Artif. Intell. **18**(2), 203–226 (1982)

[14] Mörchen, F., Fradkin, D. : Robust mining of time intervals with semi-interval partial order patterns. In : SDM. pp. 315–326. SIAM (2010)

[15] Muggleton, S., Raedt, L.D. : Inductive logic programming : Theory and methods. J. Log. Program. **19/20**, 629–679 (1994)

[16] Nienhuys-Cheng, S.H., Wolf, R.D., de Wolf, R. : Foundations of Inductive Logic Programming. Springer-Verlag New York, Inc. (1997)

[17] Plotkin, G.D. : A note on inductive generalization. Machine Intelligence **5**, 153–163 (1970)

[18] Quinlan, J.R., Cameron-Jones, R.M. : FOIL : A midterm report. In : ECML'93. Lecture Notes in Computer Science, vol. 667, pp. 3–20. Springer (1993)

[19] Rabold, J., Siebers, M., Schmid, U. : Generating contrastive explanations for inductive logic programming based on a near miss approach. Mach. Learn. **111**(5), 1799–1820 (2022)

[20] Ribeiro, M.T., Singh, S., Guestrin, C. : "Why Should I Trust You ?" : Explaining the predictions of any classifier. In : Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)

[21] Soulet, A., Rioult, F. : Exact and approximate minimal pattern mining. In : Advances in Knowledge Discovery and Management - Volume 6 [Best of EGC 2014-2015]. Studies in Computational Intelligence, vol. 665, pp. 61–81 (2015)

[22] Soulet, A., Rioult, F. : Exact and Approximate Minimal Pattern Mining, pp. 61–81. Springer International Publishing (2017)

[23] Sutton, R.S., Barto, A.G. : Reinforcement Learning : An Introduction. MIT Press (1998)

Session 2 : Apprentissage

L'apprentissage algorithmique, une nouvelle étape pour l'IA. Une application aux opérations arithmétiques

Frédéric Armetta[1], Anthony Baccuet[1], Mathieu Lefort[1]

[1] Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France

{frederic.armetta, mathieu.lefort}@univ-lyon1.fr

Résumé

L'apprentissage profond fournit les meilleures performances de l'état de l'art dans de nombreux domaines (reconnaissance d'images, traitement naturel du langage, ...). Une des prochaines étapes consiste en l'apprentissage d'algorithmes, afin de proposer de nouvelles formes de généralisation pour l'IA. Ce problème est actuellement très difficile car il nécessite des récurrences algorithmiques, la gestion d'une mémoire et la combinaison de sous-tâches, ce qui engendre des problèmes d'apprentissage et de faibles performances. Nous présentons une méthode d'apprentissage originale (UAT : Unrolling Algorithmic Training) pour aborder ces spécificités, appliquée à la multiplication multi-digit.

Mots-clés

Apprentissage automatique d'algorithmes, Apprentissage profond, Apprentissage Actif

Abstract

Deep learning achieved state of the art performances in multiple domains (image recognition, natural language processing, ...). One of the next steps is to be able to learn algorithms, as a way to provide some new forms of generalization for AI systems. This is currently a hard and challenging problem as it involves algorithmic recurrence, memory management and sub-tasks to combine, which leads to trainability problems and poor performance. We present an original training method to address these specificities for the multi-digit multiplication learning, called Unrolling Algorithmic Training (UAT).

Keywords

Algorithmic Machine Learning, Deep Learning, Active learning

1 Introduction

Depuis les premiers succès des réseaux de neurones convolutifs pour la classification d'images [18], les méthodes d'apprentissage profond ont fait progresser l'état de l'art dans de nombreux domaines [12]. Cela se traduit par des tâches de plus en plus complexes comme la traduction automatique [32] ou la résolution de jeux combinatoires [30]. L'une des prochaines étapes correspond à l'apprentissage d'algorithmes, qui peut être utile au calcul d'expressions mathématiques, afin de permettre l'apprentissage de toute tâche pouvant être exprimée dans une machine de Turing et de développer ainsi l'autonomie des intelligences artificielles.

L'exécution d'algorithmes par réseaux de neurones est encore un domaine de recherche émergent. Il s'agit d'un problème difficile car il nécessite de mémoriser des variables pendant de longues périodes, d'apprendre des inférences, de combiner des procédures, d'extrapoler à des domaines inconnus, etc. Plus fondamentalement, il s'agit de concilier la compréhension et la manipulation symboliques avec l'apprentissage statistique. La machine de Turing neuronale [13] a été proposée pour apprendre des procédures algorithmiques de bout en bout (résolution *end-to-end*), telles que le tri de listes. Ce modèle propose de réduire le problème d'apprentissage (*trainability problem*) rencontré par les RNN (*Récurrent Neural Networks*), qui sont Turing complets [28], en ajoutant des mécanismes spécifiques tels qu'une mémoire externe et des possibilités d'accès différentiables. Cependant, ce modèle rencontre malgré tout des problèmes d'apprentissage, de sorte que les performances sont difficiles à reproduire et inconstantes [8]. Ces problèmes sont causés par la profondeur de l'architecture proposée [17] mais aussi par la complexité intrinsèque des opérations à apprendre, en particulier les dépendances à long terme dans les données ou les variables à manipuler.

Pour surmonter ce problème d'apprentissage d'algorithmes, une autre solution consiste à décomposer cet algorithme en étapes successives que le réseau neuronal calculera et apprendra de manière itérative en réintroduisant sa sortie précédente par le biais d'une récurrence externe au modèle. Cette procédure a été appliquée à une architecture inspirée d'une architecture transformeur pour l'apprentissage de certaines procédures algorithmiques [34] et à un MLP (*multilayer perceptron*) pour les opérations arithmétiques [21, 34].

Dans cet article, nous avons choisi comme application les opérations arithmétiques pour illustrer notre proposition. En particulier, la multiplication à plusieurs chiffres qui peut être résolue par un algorithme. Ce calcul implique de nombreuses opérations et variables dans le flux de calcul (calculer plusieurs multiplications à un chiffre, additionner correctement les résultats intermédiaires afin d'obtenir le résultat final). La complexité d'une telle opération est également liée à la propagation des retenues [7]. L'apprentissage direct de bout en bout des multiplications conduit à des performances

médiocres [16]. Pour résoudre ce problème d'apprentissage, nous proposons avec la méthode UAT (*Unrolling Algorithmic Training*) d'introduire conjointement pendant l'apprentissage différentes sous-tâches intermédiaires utiles à la résolution et la multiplication complète de bout en bout, en pondérant les différentes tâches par une stratégie d'apprentissage actif. Cette approche est en quelque sorte similaire à un apprentissage multitâche, en utilisant des tâches complémentaires pertinentes pour aider le réseau à apprendre la multiplication complète. Cependant, pour l'approche que nous proposons, le réseau apprend toutes ces tâches sans aucune couche dédiée à chacune d'entre elles. Ainsi, le réseau devra s'accommoder des opérations intermédiaires afin d'apprendre la multiplication complète. Une autre différence est que cette adaptation doit suivre le cours de l'algorithme, en connectant séquentiellement les tâches et en propageant les valeurs intermédiaires utiles à la résolution. Nous montrons que cette procédure d'apprentissage peut également être utilisée pour pré-entraîner le réseau.

Nous présentons les travaux connexes à ce travail dans la section 2. Nous montrons dans cette section que les modèles de langage larges ne fournissent pas de bons résultats concernant les opérations arithmétiques telles que la multiplication et connaissent des difficultés à généraliser. Nous détaillons le cas d'utilisation des multiplications à plusieurs chiffres dans la section 3. Cette tâche est l'une des plus difficiles des quatre opérations arithmétiques car lorsque résolue algorithmiquement elle s'appuie sur un grand nombre d'étapes intermédiaires. Le modèle sur lequel notre procédure d'apprentissage est appliquée est ensuite détaillé dans la section 4. Le protocole et les résultats sont présentés dans la section 5. Nous conclurons et discuterons des perspectives de ce travail dans la section 6.

2 État de l'art

Au cours des années, de nombreux modèles d'apprentissage profond ont été proposés pour résoudre différents types de problèmes [12]. Les réseaux neuronaux convolutifs [18] permettent de classer des images avec des performances dépassant parfois celles des humains dans des scénarios spécifiques. Les réseaux neuronaux récurrents (*RNN*) permettent quant à eux de manipuler des données temporelles, certains modèles permettent d'atténuer les phénomènes d'oubli des données manipulées [15] [6].

Les modèles d'apprentissage profond sont des approximateurs de fonctions universels [9], mais ils sont souvent difficiles à entraîner. En théorie, une architecture neuronale même peu profonde est suffisante pour apprendre toute fonction. En pratique, la capacité des réseaux profonds à apprendre des représentations pertinentes à partir des données est bien meilleure [1]. Les réseaux de neurones récurrents sont Turing-complets [28], mais ils sont également confrontés à des difficultés d'apprentissage [13]. Cette limitation tend à être plus prononcée pour les tâches complexes, en particulier lorsque des inférences à long terme sont nécessaires. Par exemple, le taux d'erreur pour une opération arithmétique est lié au nombre de retenues pour un perceptron multicouche [7].

Certaines stratégies globales ont été proposées pour rendre l'apprentissage des réseaux plus efficace. Parmi celles-ci, les stratégies d'apprentissage actif s'intéressent à sélectionner les données pour l'apprentissage qui permettent de maximiser la progression de l'apprentissage [27, 4]. Une façon courante de le faire est l'apprentissage par curriculum pour lequel les tâches soumises sont ordonnées par complexité croissante [2]. Une autre stratégie repose sur de l'apprentissage multitâche qui consiste à combiner différentes tâches avec des objectifs similaires afin que le réseau puisse mieux généraliser et apprendre la structure sous-jacente des données [35]. Dans ce cas, certaines couches en sortie de réseau sont spécifiques aux différentes tâches apprises. Même si notre proposition peut sembler s'apparenter à de l'apprentissage multitâche, dans notre modèle les tâches sont apprises sur le même réseau et sont de nature algorithmique.

La machine de Turing neuronale (*NTM*) [13] a été spécifiquement conçue pour apprendre des tâches algorithmiques. Afin de limiter le problème de dépendance temporelle rencontré par les réseaux récurrents, elle imite certains des mécanismes d'une machine de Turing en introduisant une mémoire et des accès différenciables en lecture/écriture. L'ordinateur neuronal différentiable [14] améliore ces accès pour apprendre des tâches plus complexes comme des problèmes d'inférence ou de raisonnement en langage naturel. Malgré tout, ces modèles sont difficiles à entraîner avec une très forte sensibilité aux conditions d'initialisation. [8, 25].

L'architecture neuronale GPU [17] partage des idées similaires avec la machine de Turing neuronale mais utilise des réseaux réccurents GRU (*Gated Recurrent Unit*) convolutifs afin d'obtenir un calcul parallèle. Elle est capable d'apprendre certaines tâches algorithmiques telles que l'addition et la multiplication binaires, l'inversion de séquence, ... Sa principale réussite est sa capacité à généraliser l'apprentissage à des entrées de taille plus importante dans les tests. C'est le cas pour la multiplication binaire à 20 chiffres dans l'apprentissage, et testée jusqu'à 2000 chiffres dans les tests sans aucune erreur. Cependant, ce modèle ne parvient pas à apprendre les multiplications décimales [10]. Cette limitation peut être liée à la propagation de retenue qui est difficile à entraîner et apparaît plus fréquemment avec le codage décimal des nombres et peut être partiellement surmontée en utilisant l'apprentissage par curriculum (de binaire à quaternaire puis vers le décimal) [23].

De nombreux articles ont étudié l'apprentissage du traitement des opérations mathématiques, souvent en utilisant des modèles issus du traitement du langage. Dans [11], différents types d'architectures d'encodage-décodage sont comparés sur des tâches simples d'évaluation d'expressions numériques. Cependant, l'article s'intéresse principalement au mélange de formats d'écriture textuelle des nombres et des opérations pour la formulation des opérations arithmétiques. Il est donc difficile d'analyser finement les performances des opérations arithmétiques. Dans [25], plusieurs algorithmes (réseaux réccurents LSTM avec ou sans mécanisme d'attention, architecture neuronale de type *transformeur*) ont été testés sur différentes tâches d'inférence mathématique.

L'objectif principal était de proposer un jeu de données et de comparer les modèles, en particulier sur l'aspect attentionnel, de sorte qu'une fois de plus, les performances détaillées manquent. L'une des conclusions est que les dépendances à long terme sont les plus difficiles à apprendre, ce qui peut être compromettant lorsque les expressions se complexifient. Un transformeur, qui lit l'opération arithmétique caractère par caractère, obtient de bonnes performances sauf pour la soustraction et la multiplication [33]. [19] résout des équations mathématiques, y compris des équations différentielles, avec des modèles Seq2Seq. L'originalité réside dans le fait que l'équation est codée par une représentation arborescente. Une autre approche propose de résoudre des expressions arithmétiques en exploitant des modules feuilles pour la résolution de sous-tâches à un chiffre (multiplication ou addition), ou des modules noeuds exploitant d'autres modules noeuds ou feuilles. Les modules feuilles sont préentrainés. Les modules noeuds apprennent par renforcement avec curriculum les modules à exploiter séquentiellement, l'emplacement mémoire des données à leur communiquer et l'emplacement ou reporter leurs résultats [5]. L'expression à calculer est communiquée au système sous la forme des deux opérandes séparés par le signe de l'opération arithmétique. Les résultats indiquent la résolution de multiplications pour des expressions allant jusqu'à la taille 10, cependant la difficulté de résolution, sous forme de décomposition hiérarchique dans le cadre de cette approche, est liée à la taille minimale des deux opérandes, celle-ci ne peut être déduite de la simple longueur de l'expression. Contrairement à ce qui est indiqué les données utilisées ne sont pas précisées en annexe. Les auteurs soulignent que l'apprentissage par curriculum est nécessaire pour atteindre les valeurs de performance reportées. Les unités logiques arithmétiques neurales [31] sont des cellules capables d'effectuer des opérations arithmétiques en introduisant des modules de calcul spécifiques tels que log, exp, ... L'objectif ici n'est pas d'apprendre l'arithmétique mais de fournir des cellules dédiées capables d'extrapoler l'apprentissage avec des calculs qui s'étendent à des domaines inconnus. Une extension directe de ce travail traitant également des entrées négatives a été proposée dans [26]. D'autres modèles modèles dérivés peuvent calculer des opérations arithmétiques sur des nombres réels [20]. Comme nous l'avons vu, une grande variété de tâches a été explorée dans la littérature, avec des objectifs variés qui ne peuvent être comparées facilement.

Le raisonnement algorithmique fait également partie du matériel qui peut être utile pour le traitement du langage naturel. Les modèles de langage préentrainés, qui se sont beaucoup développés ces dernières années, sont sollicités pour leurs capacité à raisonner. Ils peuvent par exemple apprendre à effectuer un raisonnement numérique à partir de peu d'exemples (en utilisant une requête telle que "Q : Combien font 24 fois 18 ?", en prenant par exemple GPT-J-6B comme base). Il a été démontré que la performance est alors fortement corrélée à la fréquence des termes dans le corpus d'entraînement[24]. Lorsque la cooccurrence des termes est faible, la précision est également faible. Ces observations soulignent le problème de la capacité d'entraînement des algorithmes et les problèmes de généralisation sous-jacents. En conséquence, les performances diminuent à mesure que la taille des opérandes augmente [3]. En outre, alors que l'addition semble moins affectée par la taille des opérandes, les performances s'effondrent pour la multiplication, qui nécessite davantage de raisonnement ou de traitement algorithmique. Les modèles comme GPT-3 sont très larges, GPT-3 exploite environ 500 milliards de tokens pour son apprentissage [22]. Cependant, les résultats montrent que la taille du corpus ne semble pas suffisante pour permettre une généralisation efficace sur des données inconnues. ChatGPT qui repose sur des modèles basés sur GPT hérite des mêmes limitations.

Ces observations suggèrent que les compétences en matière de raisonnement pour de tels modèles ne doivent pas être surestimées. Elles dépendent fortement de la taille du corpus disponible et des inférences statistiques. Le corpus ne contiendra jamais toutes les combinaisons de termes ou les paramètres des algorithmes envisageables. Une meilleure approche consiste donc à améliorer la capacité d'apprentissage des algorithmes, ce qui permettrait de faire de meilleures prédictions pour des configurations inconnues.

Dans cet article, nous avons choisi de nous concentrer sur la multiplication à plusieurs chiffres de deux nombres décimaux. Cette opération est suffisamment simple pour ne pas mélanger différents problèmes, mais elle constitue néanmoins un défi, car de nombreux modèles sont incapables de l'apprendre correctement. Cette difficulté provient de la dépendance à long terme due à la propagation de la retenue, mais aussi et plus généralement de la complexité inhérente aux algorithmes qui mettent en jeu de nombreuses opérations et variables dans le flux de calcul.

De plus, certains articles ont mesuré précisément les performances de cette tâche. [16] propose un MLP pour apprendre l'addition et la multiplication soit à partir d'entrées visuelles, soit à partir d'un encodage numérique. Dans les deux cas, la précision obtenue pour la multiplication est faible.

3 Énoncé du problème

Dans cet article, nous considérons la multiplication à plusieurs chiffres comme tâche d'apprentissage par un réseau de neurones. Sachant n le nombre de chiffres des opérandes considérés. La multiplication des deux opérandes conduit à $n+1$ sous-tâches : n multiplications à un chiffre et 1 addition finale du résultat des multiplications partielles (voir figure 1). Selon la représentation des données choisie, chaque opération correspondra à deux lignes de calcul : une pour les retenues et une pour le résultat. La taille maximale de toute opération intermédiaire est $N = 2n$ (cette longueur maximale sera obtenue pour l'addition finale avec une retenue générée à la position du chiffre le plus significatif). Tous les nombres sont complétés par des chiffres de valeur nulle pour correspondre à une taille de N, mais il y aura exactement $n+1$ opérations intermédiaires même si le premier opérande a moins de n chiffres.

0023	(1)
×0048	(2)
———	
0012	(3)
0184	(4)
0010	(5)
+0920	(6)
———	
0110	(7)
1104	(8)

FIGURE 1 – Exemple de représentation d'une multiplication de deux opérandes à 2 chiffres. Notez que les signes (+ et ×) et les lignes horizontales ne sont représentées que pour plus de clarté. Les lignes (1) et (2) correspondent aux deux opérandes. Les lignes (3) et (4) (respectivement (5) et (6)) représentent les retenues et le résultat de la multiplication à 1 chiffre de 8 (respectivement 4) par 23. Les lignes (7) et (8) correspondent aux retenues et au résultat de l'addition des lignes (4) et (6), qui est aussi le résultat final de la multiplication globale de 48 par 23, soit 1104.

4 Modèle

Notre modèle adopte un fonctionnement très similaire à celui des réseaux récursifs utilisés par certains modèles pour le traitement du langage naturel. La récurrence ainsi représentée permettant de potentiellement dérouler étape par étape l'algorithme. Nous n'utilisons pas d'encodage spécifique tel que le binaire, afin de maintenir une manipulation symbolique plus agnostique et générale.

4.1 Représentation des données

Chaque chiffre c est représenté par un vecteur à 10 dimensions selon un codage *one-hot*, soit $(\delta_{ci})_{i\in\{0,\dots,9\}}$. Ce vecteur peut également prendre deux autres valeurs. Tout d'abord, un vecteur nul (composé uniquement de bits à 0) est utilisé pour coder d'éventuelles lignes vides (voir la liste des tâches sur le tableau 1). Notez qu'il sera également utilisé comme caractère de départ pour le décodeur détaillé ci-dessous. D'autre part, un vecteur 1 (composé uniquement de bits à 1) correspond à la fin de la ligne lors de la lecture ou de l'écriture des données.

4.2 Architecture du modèle

Nous utilisons un modèle *Seq2Seq* tel que proposé dans [29] (voir figure 2) . Celui-ci est composé d'un encodeur qui va lire séquentiellement les données et les intègre par récurrence dans l'espace latent du réseau. À partir de l'espace latent résultant, un décodeur produira itérativement une séquence de valeurs, en commençant par un code prédéterminé.

L'encodeur et le décodeur sont tous deux implémentés par des réseaux récurrents LSTM (*Long-Short-Term-Memory*).

Pendant l'apprentissage, nous avons utilisé un apprentissage forcé (*teacher forcing*), c'est-à-dire que les caractères qui alimentent récursivement le décodeur sont issus des valeurs cibles (et ne correspondent pas aux valeurs inférées par le réseau, tel qu'utilisé pour la phase de test). La sortie du décodeur est alors comparée pour chacune des valeurs à celles attendues, l'erreur est rétropropagée à travers le décodeur et l'encodeur.

Un bandeau permet de lire et d'écrire deux lignes successives à la fois, chiffre par chiffre. Les chiffres, encodés dans des vecteurs *one-hot*, sont ensuite lus de droite à gauche, avec le vecteur de fin de ligne <eol>(*end of line*) positionné à la fin de chaque ligne. De la même manière, les vecteurs *one-hot* sont écrits sur les sorties. L'entrée du codeur et la sortie du décodeur ont ainsi une taille de 2×10. Chacune des différentes tâches (détaillées ci-dessous) utilise ainsi le même codage.

4.3 Définition des tâches

L'algorithme de multiplication étudié comporte des sous-tâches que l'on soumet également au modèle. L'encodeur enregistre chaque tâche selon sa formulation et le décodeur produit la valeur cible associée (voir le tableau 1). Tel que présenté dans la section 3, l'opération de multiplication de deux nombres à n chiffres peut être décomposée en n multiplications à un chiffre et 1 addition finale. Ces n+1 opérations et la multiplication globale (c'est-à-dire le calcul du résultat final et des retenues) constituent les n+2 tâches d'apprentissage pour le réseau. Pour chaque tâche, l'encodeur lit les entrées correspondantes et le décodeur produit la valeur cible de l'opération correspondante, comme illustré dans le tableau 1. On remarque que la tâche de multiplication globale de bout en bout (*end-to-end*) est soumise sans distinction, avec une phase d'encodage de lignes vides complémentaires dont on pourra par la suite faire varier le nombre afin d'en étudier l'influence. Ces lignes complémentaires permettent au réseau de différencier dans l'encodage la tâche globale de la première multiplication à un chiffre (st1).

Tâche	Entrée Encodeur	Sortie Décodeur
st1	(1//2)	(3//4)
st2	(1//2) (3//4)	(5//6)
st3	(1//2) (3//4) (5//6)	(7//8)
tglobale	(1//2) (ligne double vide $*$ 2)	(7//8)

TABLE 1 – Sous-tâches et tâche globale associées à une multiplication à 2 chiffres (voir les lignes numérotées sur la figure 1, les lignes sont lues et écrites deux par deux ("//"))

4.4 Mécanisme d'apprentissage actif

Nous avons décrit jusqu'à présent plusieurs tâches. On peut choisir de sélectionner un ensemble de tâches à entraîner simultanément sur le modèle. Dans ce qui suit, différentes configurations sont utilisées (sous-tâches uniquement, toutes les tâches, tâche globale uniquement) pour la phase d'apprentissage.

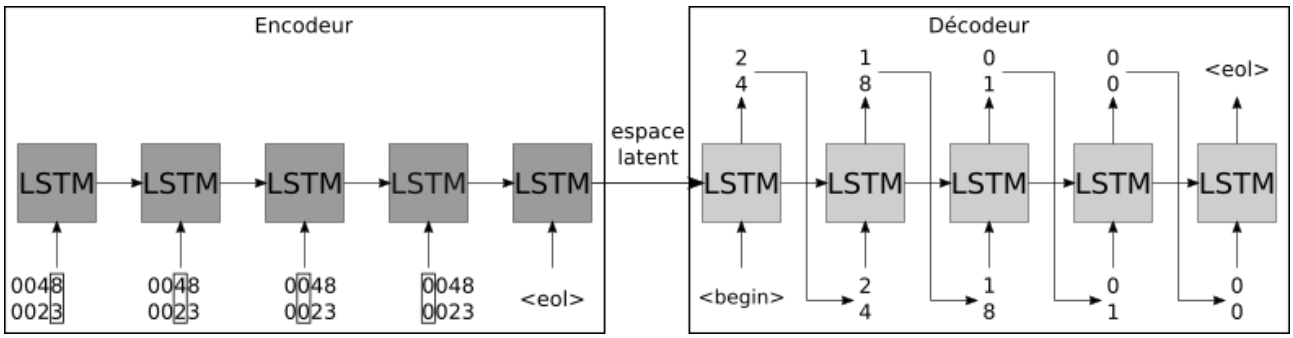


FIGURE 2 – Architecture *Seq2Seq*. L'encodeur reçoit successivement les deux chiffres (encadrés) de deux lignes consécutives (les opérandes 23 et 48) de droite à gauche. À partir de l'encodage, le décodeur produit de façon récurrente les chiffres, de droite à gauche, des deux lignes suivantes (ici la retenue et le résultat de la première multiplication intermédiaire 8×23). Dans la pratique, chaque chiffre est codé sous la forme d'un vecteur *one-hot*.

Afin d'équilibrer l'effort d'apprentissage entre les différentes tâches, nous utilisons le même mécanisme d'apprentissage actif que celui proposé dans [21]. Il consiste d'une part à mesurer le taux d'erreur, noté $err_{tâche}$, de toute tâche pour chaque période d'apprentissage. Pour chaque période (*epoch*), les instances de tâches à apprendre sont sélectionnées aléatoirement parmi l'ensemble des données d'apprentissage préalablement générées. Les instances de tâches sont sélectionnées en respectant la proportion :

$$F_{tâche} = \lambda \frac{err_{tâche}}{\sum_{ta \in listeTâche} err_{ta}} + (1-\lambda)\frac{1}{card(listeTâche)}$$

où λ est un hyperparamètre et *listeTâche* est la liste de toutes les tâches concernées. L'idée générale étant que plus une tâche est difficile à apprendre (premier terme), plus elle est présente pour la période d'apprentissage, avec une limite inférieure dépendant de la valeur lambda.

5 Expérimentations

5.1 Protocole

Pour l'apprentissage, 100 000 couples uniques d'opérandes sont générés. Pour chaque période d'apprentissage, les entrées et sorties du réseau sont générées pour chacun de ces couples, tout en respectant la proportion de chacune des tâches calculée par le mécanisme d'apprentissage actif présenté dans la section 4.4.

Pour les ensembles de données de validation et de test, nous utilisons respectivement 1000 et 10000 couples uniques supplémentaires d'opérandes. La taille de l'espace latent du codeur est fixée à 500 et le paramètre d'apprentissage actif λ à 0.5. Le modèle est entraîné à l'aide de l'optimiseur ADAM avec un taux d'apprentissage de 10^{-4} et un *batch* de taille 10. Tous les résultats présentés dans les sections suivantes sont moyennés sur 4 exécutions de 500 périodes d'apprentissage (*epochs*).

5.2 Résultats

Dans cette section, nous quantifions l'effet de notre proposition d'apprendre simultanément les différentes sous-tâches avec la multiplication globale de bout en bout.

Afin de nous comparer, nous reproduisons le format des multiplications présentées dans [16] et [21], avec des opérandes à 4 chiffres sélectionnés de manière à ce que les résultats soient restreints à des nombres de 7 chiffres.

5.2.1 Sous-tâches et tâche globale

L'objectif principal de cet article est d'aborder le problème d'apprentissage rencontré pour la tâche de multiplication de bout en bout que nous avons choisie comme une première étape vers l'apprentissage algorithmique. Pour le groupe de tâches considérées, les sous-tâches et la tâche globale sont entraînés simultanément sur le même réseau en respectant les proportions de l'apprentissage actif décrit. Nous présentons dans le tableau 2 les performances pour la tâche de multiplication globale, et les comparons avec [16] qui utilise un perceptron multicouches (*MLP*) alors que nous utilisons un modèle *Seq2Seq*. Pour les expérimentations reportées, nous faisons également varier l'échelle et la complexité des problèmes afin de souligner les améliorations apportées par notre proposition.

Les résultats démontrent clairement la contribution des sous-tâches pour l'apprentissage de la tâche globale, en réduisant le taux d'erreur (de 35,87% à 4,51% pour les multiplications restreintes 4×4).

Cela suggère que le modèle est capable de s'auto-organiser et de mettre à profit les sous-tâches complémentaires apprises, validant ainsi le fait que la procédure d'apprentissage proposée est l'élément clé de notre modèle. Bien que les résultats soient similaires sans l'introduction de ses sous-tâches, notre approche est nettement plus performante que [16]. La réduction de l'écart-type lorsque l'apprentissage comprend les sous-tâches montre que la stabilité de l'apprentissage est également améliorée (à l'exception de UAT (train = tglobale) à 8 chiffres de sortie qui peut être exclu de la comparaison par ses faibles résultats).

Nous présentons ci-dessous d'autres améliorations significatives pour le problème le plus difficile présenté (le problème 4×4 (8 chiffres de sortie)) suite a un apprentissage de finition (*fine tuning*) et à l'introduction de récurrences supplémentaires.

	test = tglobale		
	3×3 (6 chiffres en sortie)	4×4 restreinte (7 chiffres en sortie)	4×4 (8 chiffres en sortie)
Hoshen et al.[16]	n.c	37.6 %	n.c.
UAT (train = tglobale)	4.05% (± 1.72%)	35.87% (± 28.10%)	92.42% (± 3.64%)
UAT (train = sous-tâches et tglobale)	3.33% (± 1.32%)	**4.51% (± 1.21%)**	23.34% (± 14.69%)

TABLE 2 – Taux d'erreurs pour la multiplication (tglobale)

4x4 (8 chiffres en sortie)	taux d'erreurs
initial UAT train = sous-tâches + tglobale	23.34% (± 14.69%)
UAT fine tuning (pre-training = sous-tâches + tglobale)	**6.68% (± 4.31%)**
UAT fine tuning (pre-training = sous-tâches)	73.31% (± 11.10%)

TABLE 3 – Influence de l'apprentissage de finition (*fine tuning*)

ligne double vide	taux d'erreurs
0	88.67% (± 5.03%)
1	50.19% (± 18.16%)
2	14.39% (± 2.65%)
3	4.44% (± 1.65%)
4	6.68% (± 4.31%)
5	5.53% (± 2.86%)
6	3.84% (± 1.61%)
7	**3.83% (± 1.29%)**

TABLE 4 – Influence de la quantité de récurrences disponible lors de la phase d'apprentissage de finition (pre-training = sous-tâches + tglobale, 4 lignes doubles vides)

5.2.2 Apprentissage de finition de la tâche globale

Afin d'améliorer les performances de notre modèle, nous proposons de poursuivre l'apprentissage pour la tâche globale seule, durant 500 périodes supplémentaires, pour le problème le plus difficile de taille 4×4.
Comme nous pouvons le voir sur le tableau 3, en conservant la même récurrence pour le réseau, cet apprentissage de finition conduit à une amélioration significative des performances puisque l'erreur passe de 23,34% à 6,68% pour le problème à 8 chiffres de sortie.
Ce résultat montre également qu'une fois que la tâche globale a pu être amorcée (grâce aux sous-tâches complémentaires), elle peut être maintenue seule en apprentissage afin de finaliser celui-ci. Au contraire, le tableau 3 montre que le fait de ne pas inclure la tâche globale pour le pré-entraînement entraîne des problèmes également lors de la phase de finalisation de l'apprentissage. Cela confirme que le modèle doit apprendre conjointement les opérations intermédiaires et la tâche globale afin de permettre l'apprentissage plus efficace de la tâche globale cible.

5.2.3 Flexibilité

Dans notre modèle, l'entrée de la tâche globale (multiplication) est composée de quelques lignes vides (voir le tableau 1). Ces lignes vides constituent des étapes de récurrence pour le réseau lors de l'encodage. Afin de mesurer l'influence de ces lignes encodées, nous en avons fait varier la quantité lors de l'apprentissage de finition uniquement (l'entraînement initial est effectué avec 4 lignes doubles tel que précédemment).
Nous pouvons observer sur le tableau 4 que les taux d'erreur tendent à diminuer de façon monotone avec le nombre de récurrences additionnelles. Cela peut sembler logique puisque l'augmentation du nombre de récurrences augmente également la puissance de calcul du modèle.

5.2.4 Dynamique de l'apprentissage actif

Dans la section 4.4, nous avons présenté le mécanisme d'apprentissage actif que nous avons utilisé pour équilibrer l'effort d'apprentissage entre les différentes tâches soumises. La figure 3 représente l'évolution des erreurs au cours de l'apprentissage pour les différentes tâches. Nous pouvons observer qu'elle est relativement constante après un certain temps et que la tâche la plus difficile est bien la tâche globale. Cependant, on peut remarquer que la deuxième tâche la plus difficile est la procédure d'addition et non l'une des multiplications intermédiaires, dont le taux d'erreur moyen est même proche de 0. Ceci est surprenant car dans la littérature, l'addition semble être une tâche simple à apprendre. Cela peut cependant s'expliquer par les nombreux opérandes impliqués pour cette addition.

5.2.5 Sous-tâches uniquement

Nous n'entraînons le réseau que sur les sous-tâches (c'est-à-dire en excluant l'opération de bout en bout). Pour obtenir le résultat global, nous déroulons l'algorithme par une récurrence manuelle (les différentes sous-tâches "s'alimentent" consécutivement). Seule la sortie de l'addition finale comme résultat global est évaluée. Cette méthode a également été utilisée dans [34], avec un simple perception multi-couches. Cette tâche est plus facile à traiter (car l'opération globale est décomposée en ses étapes successives pour son exécution).
En particulier, nous cherchons à estimer l'effet de la récurrence disponible dans notre réseau sachant que [21] utilise

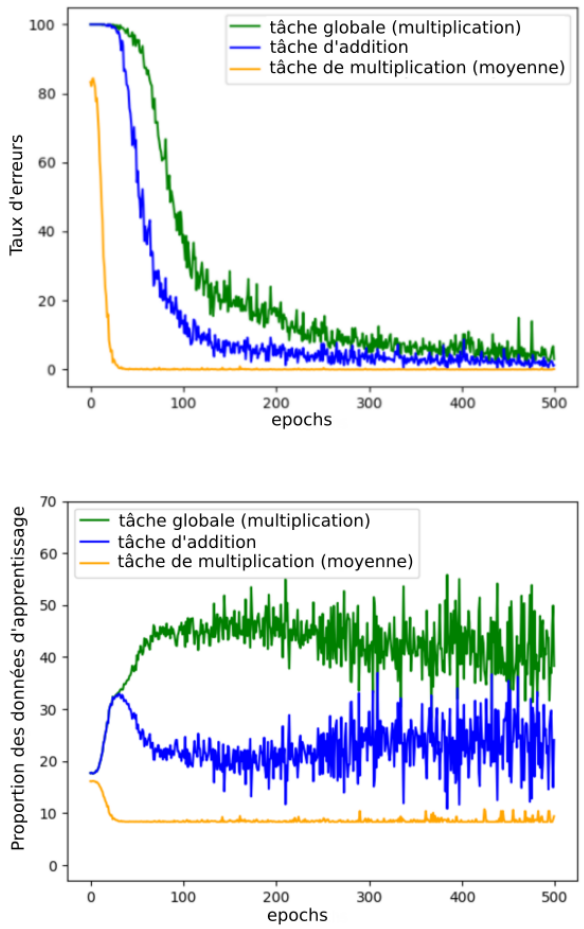


FIGURE 3 – (En haut) Évolution du taux d'erreur pour chaque tâche (estimé à partir de l'ensemble des données d'apprentissage) utilisé pour l'apprentissage actif. (En bas) Évolution associée de la proportion des différentes tâches dans l'ensemble des données d'apprentissage

un réseau sans récurrence.

Le tableau 5 montre que notre modèle réduit le taux d'erreur, dans ce contexte d'exécution. Cela montre que le modèle récurrent et la représentation des données que nous utilisons parviennent à apprendre toutes les sous-tâches.

	test = récurrence manuelle
	4×4 restreinte (7 chiffres en sortie)
[21]	2 %
UAT	**0.31% (± 0.11%)**

TABLE 5 – Comparaison du taux d'erreur entre [Nollet et al., 2020] et UAST, en combinant des tâches secondaires uniquement grâce à une récurrence manuelle

6 Conclusion and perspectives

Malgré ses multiples succès, les réseaux profonds rencontrent des difficultés à apprendre les algorithmes, comme certaines opérations arithmétiques. Cela s'explique par la complexité de la tâche, en particulier les dépendances à long terme, mais correspond aussi à un problème d'entraînement rencontré par de multiples modèles.

Cette limitation peut également être observé pour les modèles de langue, dont les performances proviennent du corpus disponible, et ne s'applique pas bien aux inférences algorithmiques lorsque la variabilité des valeurs des paramètres est importante et que le processus de calcul de l'algorithme implique de nombreuses étapes, comme pour la multiplication arithmétique.

Dans cet article, nous proposons une méthode originale pour apprendre la multiplication à plusieurs chiffres de bout en bout de deux opérandes (décimaux), en guidant le modèle par l'introduction de sous-tâches (ou sous-routines) en même temps que la tâche cible, tout en appliquant une stratégie d'apprentissage actif entre tâches. Nous montrons à travers l'analyse de nos expériences que l'amélioration des performances mesurée est directement causée par la méthode d'apprentissage que nous introduisons. L'apprentissage de la tâche globale suite à celui des sous-tâches peut améliorer les résultats. Cependant, la meilleure façon de procéder est sans équivoque de soumettre l'ensemble des tâches simultanément à l'apprentissage pour permettre d'amorcer la tâche d'apprentissage globale de bout en bout (la multiplication complète). Une fois amorcée, la tâche globale peut poursuivre son apprentissage seule de manière efficace.

En finalisant l'apprentissage sur la tâche globale, nous montrons une propriété supplémentaire intéressante de notre méthode d'apprentissage. Une fois que la multiplication globale est apprise, le réseau peut s'adapter à un nombre de récurrences différent de celui pour lequel il a initialement été paramétré. Cela semble indiquer que le modèle est capable non seulement de combiner les sous-tâches pour résoudre la tâche globale de bout en bout, mais aussi d'extraire et d'adapter de manière autonome une sorte de connaissance de haut niveau. La restriction des récurrences fournies pour le calcul et le maintien de la précision ressemble à une parallélisation contrainte de la tâche algorithmique. Bien que l'approche décrite permette d'apprendre plus efficacement la multiplication, elle met en oeuvre de nombreuses récurrences qui limitent les capacités de résolution pour les opérations de plus grande taille. Le problème du passage a l'échelle a également été observé sur les autres approches pour la multiplication décimale. Tout en conservant le mixage des différentes sous-tâches lors de l'entraînement algorithmique, qui est central à l'approche proposée, certaines optimisations pourraient être étudiées pour de futurs travaux.

La principale motivation de ce travail n'est pas seulement la résolution de la multiplication, mais également de développer un procédé permettant d'atténuer le problème d'entraînement difficile qui a été observé pour les algorithmes. Il serait intéressant d'étudier comment permettre d'apprendre des algorithmes afin de réaliser de meilleures inférences lorsque

une approche statistique ne permet pas de généraliser, tel qu'observé pour les multiplications sur les grands modèles de langage.

Ce travail soulève également de nombreuses questions de recherche concernant la dynamique d'apprentissage. Il serait intéressant d'étudier plus précisément comment le transfert des étapes intermédiaires vers la tâche globale est réalisé par le réseau. Pour surmonter le problème d'apprentissage rencontré, nous fournissons toutes les étapes intermédiaires au réseau pendant l'apprentissage. Un cas spécifique que nous souhaiterions étudier serait de ne fournir que certaines des tâches de soutien et d'observer la capacité du réseau à compléter les tâches inconnues de lui-même. Pour cela, il serait intéressant d'étudier le mécanisme de transfert des étapes intermédiaires vers la tâche globale. Cette compréhension pourrait nous donner la possibilité de contrôler le flux d'interaction entre les étapes de soutien et la dynamique générale du transfert algorithmique, et donc de peut-être mieux appréhender la façon dont une IA pourrait exploiter et adapter ses connaissances algorithmiques afin d'étendre ses capacités.

Remerciements

Ce travail a été réalisé en utilisant les ressources HPC de GENCI-IDRIS et un GPU offert par @NVIDIA Corporation. par @NVIDIA Corporation. Nous remercions chaleureusement ce soutien.

Références

[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep ? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33 :1877–1901, 2020.

[4] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv :1808.04355*, 2018.

[5] Kaiyu Chen, Yihan Dong, Xipeng Qiu, and Zitian Chen. Neural arithmetic expression calculator. *CoRR*, abs/1809.08590, 2018.

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[7] Sungjae Cho, Jaeseo Lim, Chris Hickey, and Byoung-Tak Zhang. Problem difficulty in arithmetic cognition : Humans and connectionist models. 2019.

[8] Mark Collier and Joeran Beel. Implementing neural turing machines. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 94–104, Cham, 2018. Springer International Publishing.

[9] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4) :303–314, 1989.

[10] Karlis Freivalds and Renars Liepins. Improving the neural gpu architecture for algorithm learning. *arXiv preprint arXiv :1702.08727*, 2017.

[11] Isha Ganguli, Rajat Subhra Bhowmick, and Jaya Sil. Study of recurrent neural network models used for evaluating numerical expressions. In *2019 IEEE Region 10 Symposium (TENSYMP)*, pages 403–408. IEEE, 2019.

[12] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[13] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

[14] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626) :471–476, 2016.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8) :1735–1780, 1997.

[16] Yedid Hoshen and Shmuel Peleg. Visual learning of arithmetic operations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3733–3739. AAAI Press, 2016.

[17] Łukasz Kaiser and Ilya Sutskever. Neural gpus learn algorithms. *arXiv preprint arXiv :1511.08228*, 2015.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. *arXiv preprint arXiv :1912.01412*, 2019.

[20] Andreas Madsen and Alexander Rosenberg Johansen. Neural arithmetic units. In *International Conference on Learning Representations*, 2020.

[21] Bastien Nollet, Mathieu Lefort, and Frédéric Armetta. Learning Arithmetic Operations With A Multistep Deep Learning. In *The International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom, July 2020.

[22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang,

Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv :2203.02155*, 2022.

[23] Eric Price, Wojciech Zaremba, and Ilya Sutskever. Extensions and limitations of the neural gpu. *arXiv preprint arXiv :1611.00736*, 2016.

[24] Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics : EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[25] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv :1904.01557*, 2019.

[26] Daniel Schlör, Markus Ring, and Andreas Hotho. inalu : Improved neural arithmetic logic unit. *arXiv preprint arXiv :2003.07629*, 2020.

[27] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[28] H.T. Siegelmann and E.D. Sontag. On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1) :132 – 150, 1995.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[30] Yuandong Tian, Jerry Ma, Qucheng Gong, Shubho Sengupta, Zhuoyuan Chen, James Pinkerton, and Larry Zitnick. Elf opengo : An analysis and open reimplementation of alphazero. In *International conference on machine learning*, pages 6244–6253. PMLR, 2019.

[31] Andrew Trask, Felix Hill, Scott E Reed, Jack Rae, Chris Dyer, and Phil Blunsom. Neural arithmetic logic units. In *Advances in Neural Information Processing Systems 31*, pages 8035–8044. Curran Associates, Inc., 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[33] Artit Wangperawong. Attending to mathematical language with transformers. *arXiv preprint arXiv :1812.02825*, 2018.

[34] Yujun Yan, Kevin Swersky, Danai Koutra, Parthasarathy Ranganathan, and Milad Hashemi. Neural execution engines : Learning to execute subroutines. *CoRR*, abs/2006.08084, 2020.

[35] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv :1707.08114*, 2017.

Champ neuronal et apprentissage profond de topologies pour la fusion multimodale

S. Forest[1,2], J.-C. Quinton[1], M. Lefort[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, UMR 5224, F-38000, Grenoble, France
[2] Univ. Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR 5205, F-69622, Villeurbanne, France

{simon.forest, quintonj}@univ-grenoble-alpes.fr, mathieu.lefort@univ-lyon1.fr

Résumé

Des agents artificiels tels que des robots ont souvent intérêt à fusionner des données issues de différentes modalités. Pour cela, il peut être pertinent de prendre en compte les variations dans la structure et la résolution des topologies sous-jacentes aux espaces sensoriels. Nous proposons d'utiliser un champ neuronal dynamique pour sélectionner des stimuli dans un contexte multimodal. Nous avions adapté le modèle à des topologies apprises (par des gaz neuronaux croissants notamment) pour la fusion, nous l'étendons maintenant en insérant un auto-encodeur pour réduire la dimensionnalité des données d'entrée.

Mots-clés

Fusion multimodale, champ neuronal dynamique, apprentissage de variétés, auto-encodeur, gaz neuronal croissant

Abstract

Artificial agents such as robots can often benefit from merging data from different modalities. For this purpose, it may be relevant to take into account the variations in the structure and resolution of the topologies underlying the sensory spaces. We propose to use a dynamic neural field to select stimuli in a multimodal context. We had adapted the model to learned topologies (with growing neural gas in particular) for fusion, we now extend it by inserting an auto-encoder to reduce the dimensionality of the input data.

Keywords

Multimodal fusion, dynamic neural field, manifold learning, auto-encoder, growing neural gas

1 Introduction

Quand on parle de traitement de l'information et de prise de décision comportementale, la façon dont on fusionne les données issues de différentes sources n'est pas à négliger. Prenons un exemple : un robot a la tâche de trouver et atteindre un réveil lorsqu'il se met à sonner. Au début, le robot peut faire face à plusieurs objets ressemblant à un réveil, qu'il ne devrait avoir aucune difficulté à distinguer. Lorsqu'une sonnerie se fait entendre, le robot devrait être capable de localiser son origine, mais avec généralement une faible précision. Avant d'entreprendre une action, le robot doit sélectionner un objet. Dans ce cas, il s'agit de l'objet, parmi ceux qui ressemblent à un réveil, qui coïncide le plus avec la localisation de la source sonore. Mais la manière dont les modalités visuelle et auditive doivent être pondérées dépend non seulement de la tâche (une horloge visible de face est moins importante qu'un son provenant d'un côté), mais aussi de la fiabilité des capteurs (la réverbération de la pièce peut rendre l'orientation du son moins décisive).

La tâche dans cet exemple est confrontée à de multiples défis, notamment la fusion de modalités sensorielles de disponibilité et de fiabilité différentes, et la sélection de (et l'attention vers) la cible. Pour résoudre ces problèmes, beaucoup de modèles actuels se basent exclusivement sur l'apprentissage profond. Dans cet article, nous prenons l'apprentissage profond comme un outil de pré-traitement de données et de réduction de dimension. Pour la fusion, nous nous reposons sur une autre approche à partir d'un champ neuronal dynamique (*dynamic neural field*, DNF), un modèle bio-inspiré d'activité neuronale [2]. Il s'agit d'un réseau récurrent en temps continu placé dans une topologie, où les poids sont connus et dépendent de la distance entre les neurones. Avec un mélange d'excitation à courte portée et d'inhibition à longue portée, les stimuli d'entrée sont mis en compétition jusqu'à ce qu'une bulle d'activité émerge, qui peut être interprétée comme une décision, décentralisée et dynamique, de sélection d'une cible et/ou d'action. De plus, la dynamique donne un lissage temporel à la bulle d'activité, malgré les fluctuations des entrées et les distracteurs potentiels. Le DNF a connu diverses applications, notamment en robotique [14, 34, 39]. En particulier, les propriétés d'interaction du DNF le rendent très adapté à la fusion multimodale [11, 35].

Les premières implémentations de DNF trouvent une limite dans la nature de la variété (*manifold* en anglais) sur laquelle elles évoluent. La plupart des applications de la littérature supposent l'existence d'une topologie régulière sous-jacente, le plus souvent 1D ou 2D [36]. Mais elle n'est guère représentative des disparités de l'espace sensoriel, disparités qui deviennent cruciales lors de la fusion multimodale. En effet, intéressons-nous à la forme des stimuli perçus dans l'environnement. La quantité d'informations disponibles est énorme, et les données qu'un agent reçoit

de ses capteurs n'en sont qu'une projection dans quelques dimensions données. Équipé d'une caméra standard, un robot recevra une projection en 2D de la partie de l'environnement à laquelle il fait face. Avec un seul microphone, il peut détecter des sons provenant de n'importe où autour de lui, mais il peut difficilement les localiser. Deux microphones peuvent permettre une certaine localisation sonore 1D le long de l'axe sur lequel ils sont alignés, généralement azimutal (grâce à la différence de temps ou d'intensité interaurale), et même un peu de 2D ou 3D en exploitant la forme des pavillons des oreilles, avec une fonction de transfert liée à la tête (*head-related transfer function*, HRTF) [4]. Nous devons d'abord tenir compte des spécificités de chaque modalité sensorielle avant de créer des comportements qui l'exploitent au mieux. De plus, nous devons trouver un moyen de faire correspondre des informations complémentaires provenant de différentes modalités, ce qui revient généralement à projeter des stimuli sur une variété commune.

Dans [10], nous avons proposé une manière d'adapter le DNF à des variétés de dimensionnalité et structure arbitraires. À l'aide de gaz neuronaux croissants (*growing neural gas*, GNG), une topologie multimodale est créée à l'intérieur de laquelle le DNF peut fusionner et sélectionner des stimuli. Les applications testées dans ce précédent article étaient limitées par la faible capacité d'apprentissage du GNG : s'il permet de faire ressortir un espace sous-jacent dans des données de plus grande dimension, il ne permet pas de réaliser des tâches plus complexes dans des espaces de très grande dimension. Localiser un stimulus visuel à partir d'une photographie, ou un stimulus auditif à partir d'un enregistrement audio brut, n'est pas possible avec un GNG seul. Il est nécessaire au préalable de réduire la dimensionnalité des entrées en apprenant des projections vers un espace plus facile à exploiter. Cette solution peut être apportée par des réseaux de neurones. Dans cet article, nous étendons la précédente contribution en ajoutant une méthode d'apprentissage profond de variétés, à savoir un auto-encodeur de Wasserstein en coupes (*sliced Wasserstein auto-encoder*, SWAE), en amont du modèle. Notre objectif est de vérifier si les propriétés du modèle précédent se maintiennent lorsqu'on ajoute une opération visant à réduire les dimensions de l'espace d'entrée, au risque de dégrader la réelle topologie sous-jacente pendant l'apprentissage de l'encodeur.

Cet article est structuré comme suit. Dans la section 2, nous présentons des travaux existants sur l'apprentissage de variétés et le DNF, et en particulier leurs applications à la fusion multimodale. Puis nous décrivons notre modèle complet dans la section 3, et montrons sa robustesse, ses performances et ses propriétés dans la section 4. Nous concluons et ajoutons des perspectives additionnelles dans la section 5.

2 Travaux existants

2.1 Apprentissage de variétés

Les capteurs fournissent des échantillons de haute dimension de l'environnement, mais les espaces sensoriels correspondent souvent à des variétés de dimension inférieure. L'apprentissage profond est particulièrement adapté à la génération de telles variétés (voir [5] pour une revue). Par exemple, il a été démontré que les dernières couches d'un réseau neuronal profond contiennent une dimensionnalité intrinsèque inférieure au nombre de descripteurs dans les données [3]. Des méthodes spécifiques telles que les auto-encodeurs variationnels [16] peuvent apprendre une structure sous-jacente de manière non supervisée. D'autres types d'auto-encodeurs existent, notamment le SWAE [18]. Ce dernier utilise dans l'apprentissage une distance de Wasserstein, qui compare la distribution des données encodées dans l'espace latent à une distribution choisie. On peut ainsi imposer, par exemple, que l'espace latent suive une distribution uniforme en 2D. Il est ensuite possible d'exploiter les propriétés géométriques de la topologie ainsi encodée [19].

Une approche moins contraignante, privilégiée dans notre précédente contribution [10], repose sur les méthodes d'auto-organisation. Dans les cartes auto-organisatrices (*self-organizing maps*, SOM), par exemple le modèle de Kohonen [17], chaque neurone représente une entrée prototypique dans l'espace sensoriel à haute dimension, de sorte que l'espace d'entrée est projeté sur un treillis neuronal de forme et de taille fixes. Dans le cas du gaz neuronal (*neural gas*, NG), les neurones ne sont pas disposés sur un treillis, mais sont connectés selon une règle Hebbienne, de sorte que les neurones de prototypes proches sont reliés entre eux [25]. Le gaz finit par remplir l'espace d'entrée, d'une manière qui imite la distribution des stimuli. Le gaz neuronal croissant (*growing neural gas*, GNG) [12] est un dérivé du NG, dans lequel des neurones sont ajoutés (ou supprimés) au fur et à mesure jusqu'à ce qu'une condition spécifique soit remplie, s'adaptant ainsi à la topologie indéterminée de l'espace d'entrée.

Variétés en fusion multimodale

De multiples articles ont montré des résultats prometteurs en fusion multimodale à l'aide d'apprentissage profond. L'apprentissage profond non supervisé peut être utilisé pour projeter des données multimodales sur une variété de faible dimension pour une utilisation en robotique [8, 21]. Les entrées peuvent être mélangées pendant l'entraînement du réseau de neurones pour exploiter les corrélations entre les modalités [40, 43]. Récemment, des extensions du réseau Transformer ont été proposées, permettant de recevoir des entrées multimodales pondérées par un module d'attention [15, 28]. Cependant, la plupart de ces travaux partent du principe que toutes les données multimodales sont corrélées entre elles. De plus, les architectures profondes sont dédiées à une tâche spécifique et aucun paradigme générique n'émerge [29].

Nous cherchons à créer une nouvelle topologie multimodale sur laquelle de nouvelles propriétés dynamiques pourraient être appliquées. Dans un premier temps, l'auto-organisation offre des solutions pour un coût bien moindre [13,20,22,27,31,41]. Les SOM et leurs dérivés sont utilisés depuis longtemps comme modèles de fusion multimodale,

mais les façons de combiner les modalités peuvent être très diverses. Les architectures composées de SOM peuvent être divisées en deux catégories. Dans la première, une SOM est formée pour chaque modalité, puis toutes les cartes unimodales sont connectées en fonction d'une règle d'apprentissage spéciale [13, 22]. Dans la seconde, les cartes unimodales sont reliées à une nouvelle SOM [20, 27] ou un NG [41] multimodal qui combine toutes les informations. Des couches supplémentaires de SOM peuvent également être envisagées pour créer un flux hiérarchisé d'informations [31]. De plus, les modèles peuvent être rendus plus adaptatifs aux tâches dépendantes du temps à l'aide de cartes « croissantes au besoin » [31], une alternative au GNG conçue pour les distributions d'entrées dynamiques [24]. Certains de ces modèles ont déjà été testés pour des modalités visuelles, auditives et/ou proprioceptives sur des robots [13, 20].

2.2 Champ neuronal dynamique

Après l'apprentissage de cartes multimodales et/ou de cartes unimodales interconnectées, nous avons besoin d'un paradigme pour dicter la manière dont la perception va se produire. La perception multimodale peut être considérée comme une forme de décision prenant en compte des entrées sensorielles de fiabilité et de pertinence variables. Nous suivons le choix d'architecture fait dans [27] et [22], où le champ neuronal dynamique (DNF) est utilisé comme paradigme qui régit la fusion ou la ségrégation des stimuli dans l'espace topologique multimodal. Le DNF a de nombreuses propriétés utiles pour la perception multimodale.

Originaire du domaine des neurosciences, le DNF a diverses applications en robotique [36]. Par exemple, l'attention visuelle peut être cumulée avec un contrôle moteur pour qu'un robot fixe de manière autonome les objets de son environnement et apprenne une carte sensorimotrice [34]. Le DNF repose sur une population d'unités connectées topologiquement, à une échelle mésoscopique, où l'activité apparente (ou potentiel membranaire moyenné sur des groupes de neurones) peut être lue pour interpréter des décisions à un niveau comportemental. L'activité évolue dans le temps en fonction d'une somme de stimulations externes et d'interactions latérales entre les neurones. Les neurones stimulés vont envoyer une forte excitation à leurs voisins les plus proches, et une inhibition modérée à leurs voisins plus éloignés, conduisant à l'émergence d'une bulle d'activité stable. En fonction du paramétrage, cela peut conduire à plusieurs types de comportements [36]. Avec une forte excitation locale, la bulle peut être auto-entretenue, agissant comme une mémoire à long terme [34]. L'inhibition à longue portée créera une compétition entre les stimuli conflictuels, jusqu'à ce que l'un d'entre eux domine les autres, ou qu'ils soient fusionnés en une seule bulle à une position interpolée [11, 38]. Ensuite, la bulle auto-entretenue peut être utilisée pour une attention sélective robuste, capable d'ignorer le bruit et les faibles distracteurs [9]. Enfin, la sortie du DNF peut être directement exploitée pour générer une commande motrice [33, 34].

Les propriétés du DNF peuvent être très utiles à la fusion multimodale. Il fournit des moyens non seulement pour améliorer la robustesse des décisions lorsque les modalités sont congruentes [35], mais aussi pour résoudre les conflits entre modalités [11]. C'est là que le choix de la variété sous-jacente peut être très important, car tout le processus de décision se base dessus, et sa structure peut déterminer entre autres la fiabilité des stimuli, qui peut avoir une influence sur la sélection.

La grande majorité des travaux utilisant le DNF supposent que la dynamique se déroule sur une topologie complètement régulière, par exemple un treillis 2D dans le cas de la vision. Cependant, il n'existe pas de moyen clair de projeter deux modalités ou plus sur le même treillis. Dans [35] et [11], des hypothèses fortes sont faites sur la forme des stimuli dans une modalité afin qu'ils s'intègrent dans la topologie de l'autre. Pour résoudre ce problème, [22] propose d'utiliser des variétés distinctes pour chaque modalité, chacune appris par une SOM, et d'appliquer un DNF sur chacun d'eux. La communication entre les modalités est assurée par un ensemble de connexions topographiques.

Une contribution intéressante de cette dernière référence est l'utilisation d'une variété apprise comme siège de la dynamique neuronale. Par ailleurs, certaines tentatives visant à modifier la projection des entrées dans la variété ont donné des résultats satisfaisants : [38] et [11] ont ainsi reproduit des comportements biologiques après avoir appliqué aux stimuli visuels une transformation logpolaire, qui modélise les variations de résolution de la rétine humaine [30]. Dans le cas de [22], les projections reçues par les neurones sont modifiées, bien qu'elles soient toujours organisées en un réseau rectangulaire. Étant donné que le DNF est fortement dépendant de la topologie et qu'il repose sur un noyau d'interaction symétrique [1], rompre la régularité de la topologie sous-jacente doit être fait avec prudence. Une démonstration de ceci a été faite dans [10], où le DNF est adapté à des variétés de forme et de dimension non contraintes, avec des résultats probants. Cependant, les topologies qui y sont présentées sont relativement simples, et le DNF n'a pas encore été testé sur des espaces réduits par apprentissage profond, qui risquent d'être beaucoup plus irréguliers. C'est l'objet de cet article.

3 Méthodes

Cet article reprend en partie les méthodes développées dans [10], où nous utilisons des GNG pour apprendre des variétés de l'espace sensoriel dans chaque modalité (sous-section 3.3), avant de les assembler en un graphe multimodal (3.4), puis d'utiliser un DNF pour créer des comportements dans cette nouvelle topologie (3.5). Nous ajoutons une étape préalable d'apprentissage d'un auto-encodeur, dont l'encodeur servira à réduire la dimension des données d'entrées avant l'apprentissage des graphes (3.2), et le décodeur servira à l'analyse des sorties du DNF (3.6). Les

1. Il a été proposé de rompre la symétrie du côté du DNF, soit par des noyaux asymétriques [6], soit par des distorsions de la topologie à l'aide de renforcements prédictifs [32], mais les deux requièrent une étape d'apprentissage supplémentaire.

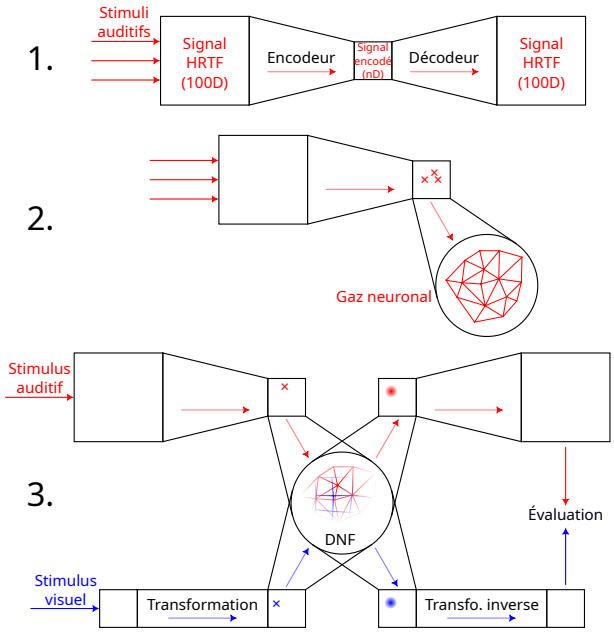


FIGURE 1 – Résumé des principales étapes de cette contribution. **1.** Apprentissage de l'auto-encodeur (sous-section 3.2). **2.** Apprentissage des graphes topologiques (sous-sections 3.3 et 3.4). **3.** Stimulation, émergence d'une activité dans le graphe bimodal et évaluation (sous-sections 3.5 et 3.6).

grandes étapes de cette contribution sont synthétisées dans la figure 1.

3.1 Données

Pour cette contribution, nous nous concentrons sur une architecture bimodale, même si ces travaux pourraient également être applicables à trois modalités ou plus. Nous prenons l'exemple d'un robot devant localiser un signal (e.g., la position d'une personne qui l'appelle) en recevant simultanément des données visuelles (détection de visage) et auditives (spectre de fréquences sonores).

Audition. Une façon de localiser les sources sonores pour les robots consiste à calculer une HRTF, une fonction qui associe des caractéristiques fréquentielles (causées par les interférences de la tête et des pavillons des oreilles sur le signal) à l'orientation de l'origine du signal [4]. Les données procurées par [1] comprennent les réponses enregistrées par un robot équipé de pavillons artificiels, à un son émis depuis différents angles. Étant donnée la position d'un stimulus externe en 2D, nous pouvons interpoler les réponses reçues par les deux oreilles robotiques. Nous calculons ensuite leur transformée de Fourier et faisons la différence entre les oreilles pour obtenir une HRTF. À la fin, chaque entrée audio est à 100 dimensions.

Cette HRTF 100D porte implicitement les informations 2D de la localisation du signal dans le référentiel de la tête du robot : azimut et élévation. C'est le signal 100D que nous donnerons en entrée de l'auto-encodeur.

Vision. La vision artificielle a très généralement une résolution homogène. Cependant, nous pouvons concevoir des cas où la perception visuelle n'est pas parfaitement régulière, par exemple à cause d'une tâche sur l'objectif de la caméra. Nous n'ajoutons pas d'étape d'apprentissage pour la vision ici. Mais, pour tester la robustesse de la fusion dans des modalités de résolution changeante, nous traitons les stimuli visuels comme des points dans un espace 2D, auxquels nous appliquons une transformation bio-inspirée, pour rester dans le même niveau de réalisme que la HRTF. Nous modifions donc l'espace visuel par une transformation logpolaire. Cette transformation a originellement été utilisée pour décrire chez l'humain la façon dont un stimulus capté par la rétine est projeté sur le colliculus supérieur, une région du cerveau impliquée dans la génération de mouvements oculaires [30]. Elle permet notamment de reconstituer la différence de résolution entre le centre de la rétine et sa périphérie. Elle a déjà été appliquée à des systèmes artificiels, par exemple pour améliorer le contrôle du regard chez les robots [23], ou pour renforcer l'apprentissage d'un réseau de neurones sur des données visuelles [7]. Et elle a déjà été couplée avec un DNF [11, 38].

3.2 Encodage

La réduction de données est effectuée par un auto-encodeur. Nous utilisons en particulier un SWAE (*sliced Wasserstein auto-encoder*), car il permet de former un espace latent euclidien muni d'une distance directement exploitable par le GNG. Il est entraîné sur les données HRTF 100D, avec un espace latent 2D, 5D ou 20D, dans lequel les entrées encodées doivent suivre une distribution uniforme. L'encodeur et le décodeur sont faits d'un réseau de neurones entièrement connecté. Les couches cachées sont séparées par un ReLU pour l'encodeur, et un LeakyReLU pour le décodeur. Le nombre de neurones dans chaque couche est listé dans la table 1.

TABLE 1 – Taille des couches de neurones des encodeurs de SWAE. Les mêmes hyper-paramètres sont utilisés pour le décodeur, en sens inverse.

Entrée	Couches cachées			Sortie
100	90	70	50	20
100		50	20	5
100	64	32	16	2

3.3 Topologies unimodales

L'objectif de cette étape est de créer des graphes représentatifs de l'espace latent de chaque modalité. Ces graphes seront ensuite fusionnés pour servir de support à la prise de décision dans un environnement multisensoriel.

Le choix de l'auto-encodeur peut permettre d'imposer la structure de l'espace latent et la distribution des stimuli encodés dans cet espace. Connaissant ceci, il serait aisé d'échantillonner la distribution par un ensemble de nœuds, et de les connecter en fonction de la structure de la topologie sous-jacente. Il n'est pas indispensable d'ajouter une

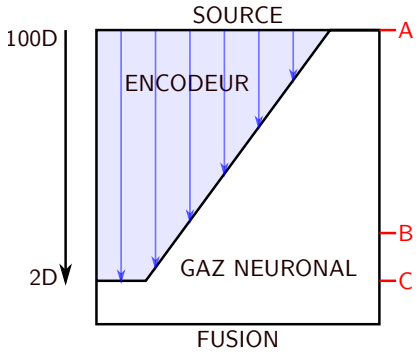


FIGURE 2 – Pré-traitement des données de HRTF avant la fusion. Plusieurs degrés de compression par l'encodeur sont possibles avant la création d'un graphe (ici, par gaz neuronal) utilisé pour la fusion et la prise de décision. Nous testons différents niveaux dans cet article. A : Pas d'encodage. C : Réduction maximale du nombre de dimensions de l'espace d'entrée. B : Réduction intermédiaire, par exemple 20D ou 5D.

étape d'apprentissage. Cependant, nous reprenons le choix de [10] de créer ces graphes à l'aide de GNG. L'algorithme en question n'introduit pas d'hypothèse supplémentaire sur la structure de l'espace latent. Il a l'avantage de fonctionner à la fois sans et avec auto-encodeur, pouvant aussi apprendre une structure sous-jacente de l'espace latent qui diffère de celle imposée dans l'apprentissage du SWAE.

L'algorithme complet du GNG est décrit dans [12]. Pour résumer, le GNG se forme à partir de stimuli tirés aléatoirement. À chaque fois, les deux nœuds dont l'entrée prototypique correspond le mieux au stimulus sont connectés entre eux. Ensuite, l'unité qui correspond le mieux (*best-matching unit*, BMU) et ses voisins topologiques directs voient leur prototype déplacé vers le stimulus. Les connexions qui n'ont pas été mises à jour depuis longtemps sont supprimées, de même que les nœuds isolés. Puis, à intervalles fixes, un nouveau nœud est inséré [2]. Son entrée prototypique va au milieu de la connexion la plus activée.

Cette étape diffère de l'article précédent dans la mesure où le GNG peut recevoir des données plus ou moins compressées par l'encodeur (figure 2). Dans le cas des signaux auditifs, le GNG peut opérer aussi bien sur des HRTF brutes en 100 dimensions que sur un encodage en 2D dans le cas le plus extrême. Pour les signaux visuels, l'encodage est remplacé par un pré-traitement manuel.

3.4 Topologie multimodale

Cette étape est directement reprise de [10] : nous créons un nouveau graphe bimodal qui contient tous les nœuds et arêtes d'une modalité et de l'autre. Pour créer de nouvelles arêtes bimodales, nous connectons des neurones de chaque modalité qui s'activent ensemble, d'une façon inspirée d'un apprentissage Hebbien. Plus précisément, nous tirons une entrée multimodale aléatoire, et si elle se trouve à portée des deux modalités, nous récupérons la BMU de chaque GNG et nous les connectons. Un seul changement est fait par rapport à l'article précédent : afin d'éviter que des nœuds trop épars dans un des GNG se connectent à énormément de nœuds de l'autre modalité, le nombre de connexions intermodales est limité à deux par nœud.

3.5 Sélection d'une activité

L'adaptation du DNF aux graphes créés précédemment est une des contributions centrales de [10]. Nous la récapitulons dans cette sous-section.

Une fois le graphe bimodal créé, les neurones qui lui sont associés peuvent être stimulés par des entrées sensorielles (via leur modalité respective), et nous pouvons utiliser un DNF pour sélectionner un stimulus. Le DNF s'exprime généralement sous la forme d'une équation intégro-différentielle dans un champ continu de neurones, qui est ensuite discrétisée et calculée par la méthode d'Euler (voir eq. 2). L'intégration d'un DNF est comparable à la simulation de réseaux de neurones récurrents en temps continu. Dans le DNF, la distance entre les neurones joue un rôle important, car elle détermine s'ils vont s'exciter ou s'inhiber mutuellement. Notre modèle diffère des autres modèles de la littérature dans la mesure où tous les neurones ne partagent pas un système de coordonnées commun. Nous devons donc adapter l'équation du DNF, afin que les distances soient définies sur le graphe, et seulement cela. Nous nous basons sur la distance standard de la théorie des graphes, c'est-à-dire le nombre d'arêtes sur le chemin le plus court entre deux sommets quelconques [42].

Dans notre modèle, chaque neurone est lié à une modalité spécifique. Ainsi, l'entrée externe reçue individuellement sera spécifique à la modalité (bien que le reste des opérations du DNF ne le soit pas). Pour s'assurer que la quantité totale de stimulation externe soit indépendante de la résolution locale d'une modalité, nous rangeons tous les neurones d'une modalité par ordre de proximité au stimulus (en utilisant la distance euclidienne dans le système de coordonnées de cette modalité), et les excitons en fonction de leur rang par ordre décroissant. Pour chaque neurone indexé k, étant donné un stimulus indexé i, on note $r_{k,i}$ le rang de proximité entre l'entrée prototypique de k et les coordonnées de i. La stimulation externe I_k reçue par k est donnée par :

$$I_k = \lambda_{m,i}\, \mathrm{e}^{\frac{-r_{k,i}^2}{2\sigma_I^2}} \qquad (1)$$

où $\lambda_{m,i}$ est l'intensité du stimulus i par rapport à la modalité m de k. Un neurone ne peut recevoir que des entrées externes provenant de sa propre modalité.

Ensuite, nous calculons l'évolution de l'activité dans le graphe au cours du temps. Ce qui suit est complètement indépendant de la modalité. Le potentiel U_k du neurone k est initialisé à 0 et mis à jour de façon incrémentale par [3] :

2. Dans cet article, le nombre de nœuds est plafonné à 500 dans chaque modalité.

3. Dans cette équation, seul U_k est incrémenté dans le temps, et les entrées I_k sont statiques. Cependant, aucune de nos hypothèses n'empêche les entrées d'évoluer au cours du temps. Nous faisons ce choix car les entrées dynamiques ne sont pas nécessaires pour les résultats présentés dans cet article. Sinon, l'équation (2) pourrait être réécrite en exprimant $U_k(t)$ comme une fonction de $U_*(t - \Delta t)$ et $I_k(t)$.

$$\Delta U_k = \frac{\Delta t}{\tau}\left(-U_k + I_k + \sum_{k'} W(<k,k'>)\, f(U_{k'}) + h\right) \quad (2)$$

où Δt est le temps de simulation entre chaque étape, τ une constante de temps qui détermine la vitesse de mise à jour du DNF, f une fonction d'activation (ReLU), et h un potentiel de repos négatif. $< \cdot, \cdot >$ désigne la distance minimale en nombre d'arêtes entre deux nœuds dans le gaz neuronal bimodal, et W est une fonction de poids exprimée comme suit :

$$W(\delta) = \lambda_+ \, \mathrm{e}^{\frac{-\delta^2}{2\sigma_+^2}} - \lambda_- \, \mathrm{e}^{\frac{-\delta^2}{2\sigma_-^2}} \quad (3)$$

avec des amplitudes $\lambda_+ > \lambda_- > 0$ et des largeurs $\sigma_+ < \sigma_-$. W peut être vu comme un noyau en forme de chapeau mexicain [2].

3.6 Évaluation

Une façon possible d'interpréter la décision est de lire la sortie $f(U)$. Il est courant de prendre un barycentre de la sortie comme estimateur de la position sélectionnée par le modèle. En l'occurence, les coordonnées des nœuds sur lesquels le DNF évolue ne sont pas directement exploitables. Nous devons d'abord décoder ces coordonnées, soit en inversant la transformation logpolaire pour la modalité visuelle, soit en utilisant le décodeur appris précédemment pour la modalité auditive (figure 1, étape 3). Les HRTF décodées sont reliées à des coordonnées 2D par interpolation dans la base de données de signaux audio. La somme de toutes les coordonnées 2D pondérées par l'activation $f(U)$ donne la position perçue du signal.

À des fins d'évaluation, nous comparons la position perçue à la position réelle du stimulus. Ceci nous donne un indicateur de la précision du modèle de fusion, même si cette étape n'est pas indispensable à ce stade. Le modèle pourrait fonctionner sans que nous ne possédions de manière supervisée de replacer, pour chaque modalité, chaque nœud de l'espace latent dans un système de coordonnées 2D intelligible (azimut–élévation). En effet, si l'espace latent contient des représentations internes de l'espace de décision, une décision prise par le DNF pourrait être transformée directement en action sur l'environnement, sans décodage explicite. Le décodage n'est pas non plus nécessaire à la fusion. Nous le faisons donc principalement pour l'évaluation et la visualisation.

4 Résultats

4.1 Topologies apprises

Un graphe est créé à partir de données audio encodées par un SWAE (figure 3). Plusieurs niveaux de compression des dimensions ont été testés (voir figure 2) : pas d'encodage (A), une réduction intermédiaire (B) vers 20 ou 5 dimensions, ou une réduction maximale (C) vers 2 dimensions.

Le premier GNG est le même qu'obtenu dans [10] (figure 3, première ligne). Une fois replacé dans des coordonnées 2D, il paraît assez régulier. L'allongement apparent du graphe le long de la direction azimutale est dû aux conditions matérielles de la perception des sons, qui rendent une discrimination gauche/droite plus facile à mener qu'une discrimination haut/bas. Une analyse plus poussée est proposée dans l'article cité.

En ajoutant un SWAE avec un espace latent imposé en 20 dimensions (figure 3, deuxième ligne), on constate que le GNG produit est très similaire au premier. Les données pertinentes dans notre évaluation (azimut et élévation) n'ont presque pas été dégradées. Cela signifie qu'il n'y a aucune perturbation à anticiper dans les propriétés du modèle de fusion, et ce, même après l'insertion d'un réseau de neurones qui transforme les données sans avoir connaissance des descripteurs les plus pertinents dans cette tâche.

Cependant, la conservation de ces propriétés n'est pas systématique en fonction de l'encodage employé. La forte dégradation des GNG 5D (figure 3, troisième ligne) et 2D (quatrième ligne) montre qu'une compression trop forte peut aboutir à une perte partielle des informations d'azimut et d'élévation. En effet, le SWAE n'a pas de raison de privilégier ces deux caractéristiques. Il sélectionne seulement les descripteurs qui permettent de mieux encoder les données en quelques dimensions et de les reconstruire. Dans les cas 5D et 2D, on voit que le GNG reste assez régulier en périphérie, et est particulièrement dégradé entre -40 et 40 degrés environ. C'est cohérent avec le fait que la localisation des sons est plus facile sur les côtés de la tête. Le SWAE a ainsi appris que la HRTF était mieux encodée par les informations de localisation sur les côtés, mais pas au centre du champ perceptif, où ces informations sont perdues au profit d'autres caractéristiques, moins pertinentes dans notre tâche, mais plus utiles pour la reconstruction des données par le décodeur.

Cette dégradation involontaire ne rend pas le nouveau GNG inexploitable, d'autant plus dans un contexte de fusion de données, où une autre modalité peut apporter des informations complémentaires. Nous le testons avec l'audition encodée en 2D (le cas le plus critique) et la vision. Comme le GNG 5D est qualitativement proche du GNG 2D, nous nous attendons à ce que les performances soient similaires avec ce nombre de dimensions. De même, nous n'attendons pas de différence dans le comportement du modèle entre des données encodées en 20D et des données non encodées, puisque les GNG obtenus sont qualitativement très proches.

4.2 Évaluation en deux dimensions

Pour cette expérience, nous apprenons un second GNG à partir de données visuelles 2D altérées par transformation logpolaire (figure 4). Comme prévu, la résolution est beaucoup plus élevée au centre du champ de vision [4] qu'en périphérie. Ensuite, comme décrit en section 3.4, nous connectons ce GNG visuel au GNG auditif 2D pour former un nouveau graphe bimodal (figure 5).

Sur ce graphe multimodal, nous pouvons envoyer des stimuli audiovisuels et faire opérer un DNF pour sélection-

4. L'absence de nœuds autour d'un azimut de zéro est due à la présence d'un log dans la transformation. Le modèle sur lequel elle s'appuie est originellement prévu pour deux hémi-champs disjoints.

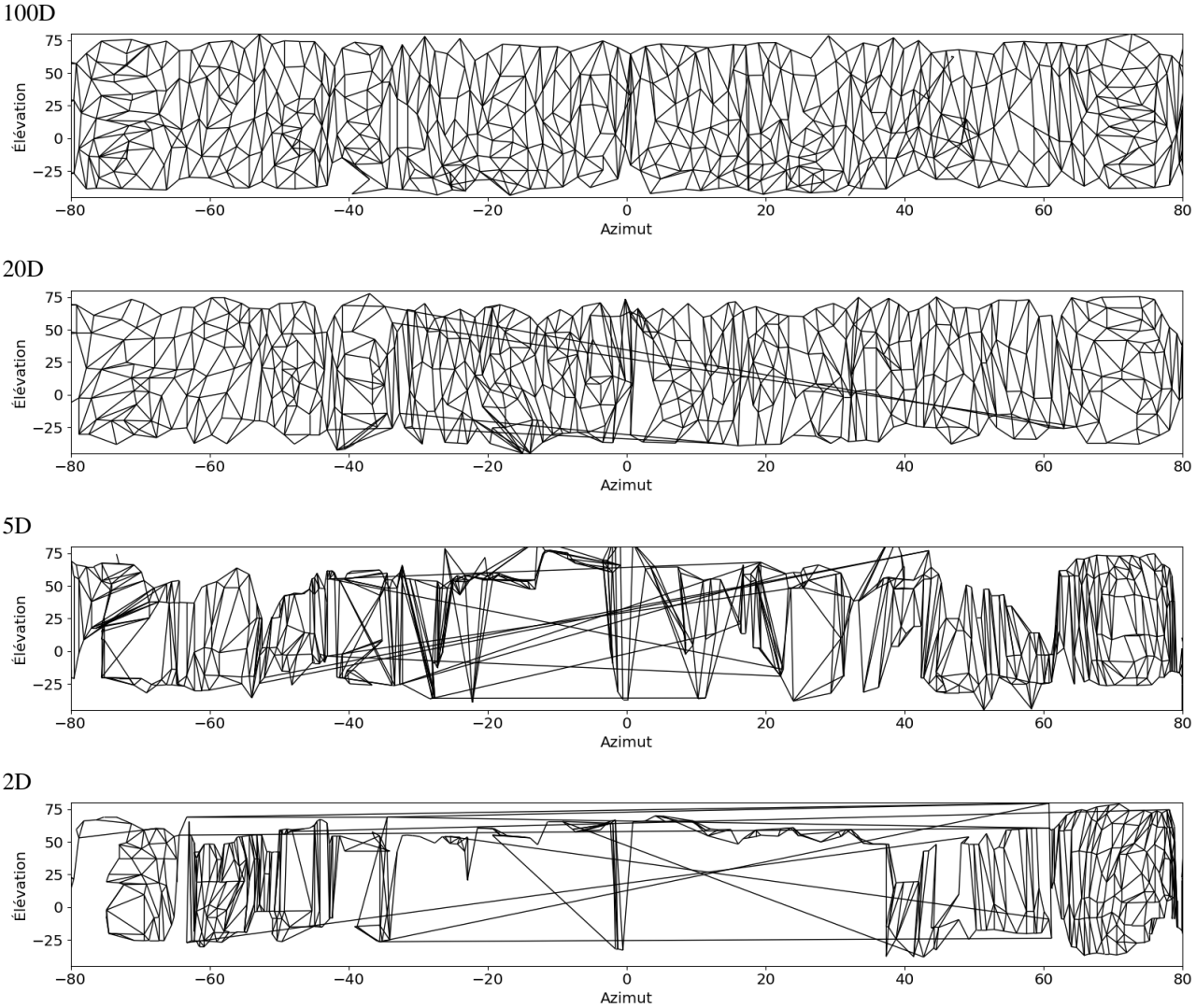


FIGURE 3 – GNG créés à partir de données audio. Le premier est appris à partir des données HRTF 100D sans encodage. Les autres sont appris à partir d'un encodage vers un espace latent 20D, 5D ou 2D. L'affichage des nœuds en 2D est réalisé en décodant leurs entrées prototypiques avec le SWAE le cas échéant, puis en interpolant les azimut et élévation d'après la base de données de HRTF. Notez que dans cet affichage, les deux axes n'ont pas la même échelle.

ner une localisation. Nous évaluons la distance entre la position trouvée et la position réelle de la source pour plusieurs azimuts (figure 6). Afin de mitiger l'effet du choix de l'élévation sur la sélection (la précision peut varier selon que la position du stimulus coïncide par hasard avec l'entrée prototypique d'un des nœuds, ou qu'au contraire elle soit très éloignée de la BMU), nous testons les élévations $[-30, -25, -20, \ldots, 30]$ et gardons la valeur de distance moyenne.

Les performances dans les cas unimodaux confirment nos observations précédentes. Hormis un artefact au centre, la perception visuelle est plus précise vers le centre qu'en périphérie, et inversement pour l'audition. Dans le cas bimodal, la perte est généralement entre les pertes subies par chaque modalité seule. Même si l'amélioration n'est pas franche, elle reste intéressante étant donnée la forte dégradation de la topologie par un SWAE qui a comprimé la HRTF en 2D sans avoir de raison explicite de conserver les informations utiles à la localisation.

5 Conclusion et perspectives

Notre modèle étend une précédente contribution, en ajoutant un SWAE pour encoder des données en amont de la création d'un graphe multimodal, sur lequel un DNF peut faire de la fusion. Dans une certaine mesure, il est possible de réduire la dimensionnalité des données d'entrée par un auto-encodeur sans aucun impact à anticiper sur l'efficacité du modèle de fusion. Toutefois, il apparaît une borne à la force de la compression, au-delà de laquelle des informations pertinentes risquent d'être perdues. Mais la perte sera en partie compensée par l'autre modalité après la fusion.

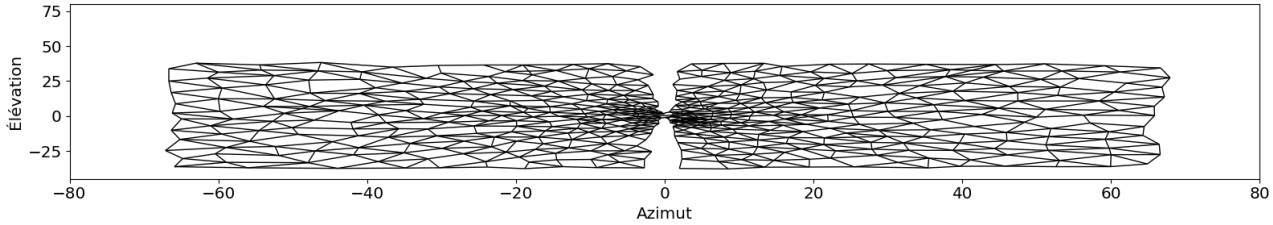


FIGURE 4 – GNG appris sur des données visuelles 2D déplacées par une transformation logpolaire

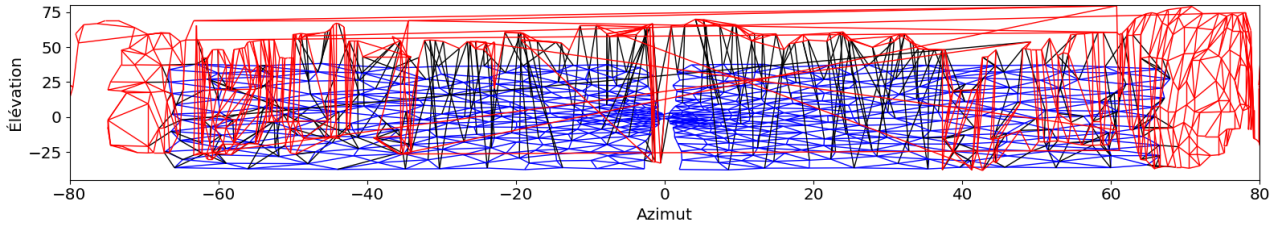


FIGURE 5 – Graphe audiovisuel obtenu en associant un GNG visuel et un GNG auditif 2D. Les connexions intra-visuelles sont affichées en bleu, intra-auditives en rouge, et inter-modales en noir.

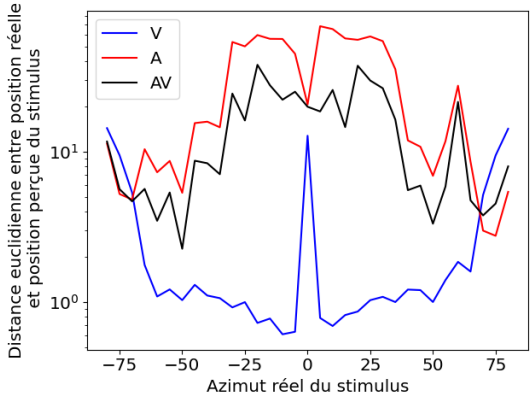


FIGURE 6 – Perte de précision dans la localisation d'un stimulus dans une topologie uniquement visuelle (V), uniquement auditive (A), ou bimodale (AV). L'ordonnée représente la distance en échelle logarithmique entre les positions réelle et perçue du stimulus. Pour chaque azimut, la distance est moyennée sur plusieurs élévations.

Les résultats présentés ici pourraient être améliorés de plusieurs façons. D'abord, dans la figure 6, il aurait été souhaitable que la fonction de distance dans le cas audiovisuel suive la meilleure des deux modalités, voire fasse mieux que les deux. Les résultats présentés précédemment dans [10] tendaient à montrer qu'il était possible de conserver les meilleurs propriétés des deux modalités. Mais il semblerait que la dégradation des informations de localisation auditives par le SWAE, quand l'espace latent est imposé en 2D sans supervision sur la pertinence des dimensions à conserver, soit trop détrimentale pour que la fusion des GNG rende les modalités vraiment complémentaires. C'est pour cette raison que nous avons eu besoin de limiter le nombre de connexions intermodales pour chaque nœud, car sinon les nœuds situés en haut du GNG auditif se connecteraient à toute une colonne de noœuds visuels, effaçant indirectement la perception de l'élévation dans tout le centre du champ de vision. Bien entendu, il serait envisageable de contraindre la structure de l'espace latent pour que, même en 2D, le SWAE apprenne à garder les propriétés de localisation sonore qui nous intéressent. L'optimisation du réseau de neurone pour accomplir cette tâche précise serait une perspective de prolongement de nos travaux. Cependant, des tests préliminaires que nous avons menés sur d'autres types d'auto-encodeurs (notamment VAE) semblent confirmer que les propriétés géométriques de l'espace latent formé par un SWAE sont bien indispensables à la création d'un graphe compatible avec notre méthode de fusion, même s'il se posera inévitablement la question de la superposition d'espaces latents de modalités différentes.

La principale nouveauté de notre modèle est que nous sommes désormais capables de coupler, d'une part, un apprentissage profond, à, d'autre part un modèle de fusion peu coûteux en apprentissage et ayant accès à des propriétés intéressantes (i.e., les capacités de sélection, attention, mémoire, etc. qui ont longuement été développées dans la littérature du DNF). L'ajout de réseaux de neurones ouvre la porte à la manipulation d'espaces d'entrée bien plus complexes. Un exemple serait la reconnaissance d'émotion. Les informations sur l'émotion d'un individu peuvent être perçues par plusieurs canaux : reconnaissance visuelle des expressions du visage, reconnaissance auditive du timbre de la voix, traitement du langage naturel… De nombreux travaux proposent de fusionner ces modalités dans le domaine de l'apprentissage profond, mais pas avec un DNF, car il

n'est pas capable d'intégrer des données aussi complexes. Notre méthode crée une opportunité de le faire.

Parmi les extensions possibles, et notamment dans le cas de tâches aussi complexes que la reconnaissance d'émotions, il serait intéressant d'étudier l'apprentissage simultané de plusieurs auto-encodeurs pour des modalités différentes. Une première piste, qui serait plus utile dans notre exemple de localisation, serait d'utiliser une modalité pour superviser l'autre. Nous entraînerions un auto-encodeur sur les données auditives, pour se conformer non pas à une distribution fixée arbitrairement, mais à la distribution latente des données visuelles, dont on sait qu'elles sont apprises avec une meilleure précision. Ce type de solution a déjà été exploré dans la littérature [26,37]. Ce serait une manière de privilégier lors de l'encodage les informations qui correspondent d'une modalité à une autre.

Cependant, il ne faut pas compter sur une correspondance systématique entre les modalités. Par exemple, il n'est pas garanti que tous les descripteurs d'une émotion soient accessibles aussi bien par reconnaissance visuelle que par traitement du langage : dans le langage seul, il est notoirement difficile de distinguer si un compliment est sarcastique ou non, là où l'expression du visage permet de faire plus facilement la distinction entre un sentiment heureux ou en colère. En forçant une corrélation visuo-langagière, nous risquerions d'entraîner un auto-encodeur (côté traitement du langage) à reconnaître dans du bruit un descripteur (existant en vision) auquel il n'a en fait pas accès. Une piste alternative serait de modifier la distance de Wasserstein pour qu'à l'entraînement, un auto-encodeur accorde moins d'importance à une dimension sur laquelle une autre modalité est plus efficace. En entraînant les modalités en parallèle, nous pourrions nous assurer de leur bonne complémentarité, et laisser ensuite le DNF réaliser la fusion et la prise de décision sans apprentissage supplémentaire.

Remerciements

Ces travaux ont été soutenus par la région Auvergne-Rhône-Alpes via l'initiative Pack Ambition Recherche (projet AMPLIFIER), ainsi que l'Agence Nationale de la Recherche – institut 3IA – MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

Une partie des calculs présentés dans cet article ont été réalisés grâce aux infrastructures de GRICAD (`https://gricad.univ-grenoble-alpes.fr`), soutenue par les communautés de recherche de Grenoble.

Références

[1] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 99–102. IEEE, 2001.

[2] S.-I. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2) :77–87, 1977.

[3] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan. Intrinsic dimension of data representations in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] S. Argentieri, P. Danès, and P. Souères. A survey on sound source localization in robotics : From binaural to array processing methods. *Computer Speech & Language*, 34(1) :87–112, 2015.

[5] Y. Bengio, A. Courville, and P. Vincent. Representation learning : A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8) :1798–1828, 2013.

[6] M. Cerda and B. Girau. Asymmetry in neural fields : a spatiotemporal encoding mechanism. *Biological cybernetics*, 107(2) :161–178, 2013.

[7] G. Dabane, L. U. Perrinet, and E. Daucé. What you see is what you transform : Foveated spatial transformers as a bio-inspired attention mechanism. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[8] A. Droniou, S. Ivaldi, and O. Sigaud. Deep unsupervised network for multimodal perception, representation and classification. *Robotics and Autonomous Systems*, 71 :83–98, 2015.

[9] J. Fix, N. Rougier, and F. Alexandre. A dynamic neural field approach to the covert and overt deployment of spatial attention. *Cognitive Computation*, 3(1) :279–293, 2011.

[10] S. Forest, J.-C. Quinton, and M. Lefort. Combining manifold learning and neural field dynamics for multimodal fusion. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[11] S. Forest, J.-C. Quinton, and M. Lefort. A dynamic neural field model of multimodal merging : application to the ventriloquist effect. *Neural Computation*, 34(8) :1701–1726, 2022.

[12] B. Fritzke. A growing neural gas network learns topologies. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995.

[13] N. Gonnier, Y. Boniface, and H. Frezza-Buet. Input prediction using consensus driven SOMs. In *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 38–42. IEEE, 2021.

[14] Q. Houbre, A. Angleraud, and R. Pieters. Balancing exploration and exploitation : a neurally inspired mechanism to learn sensorimotor contingencies. In *Human-Friendly Robotics 2020 : 13th International Workshop*, pages 59–73. Springer, 2021.

[15] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver : General perception with iterative attention. In M. Meila and T. Zhang,

editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR, 2021.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2014.

[17] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1) :59–69, 1982.

[18] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019.

[19] S. Krishnagopal and J. Bedrossian. Preserving data manifold structure in latent space for exploration through network geodesics. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[20] S. Lallee and P. F. Dominey. Multi-modal convergence maps : from body schema and self-representation to mental imagery. *Adaptive Behavior*, 21(4) :274–285, 2013.

[21] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch : Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[22] M. Lefort, Y. Boniface, and B. Girau. SOMMA : Cortically inspired paradigms for multimodal processing. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013.

[23] L. Manfredi, E. S. Maini, and C. Laschi. Neurophysiological models of gaze control in humanoid robotics. In B. Choi, editor, *Humanoid Robots*, chapter 10. IntechOpen, Rijeka, 2009.

[24] S. Marsland, J. Shapiro, and U. Nehmzow. A self-organising network that grows when required. *Neural networks*, 15(8-9) :1041–1058, 2002.

[25] T. Martinetz and K. Schulten. A "neural-gas" network learns topologies. *Artificial neural networks*, 1 :397–402, 1991.

[26] S. Moon, S. Kim, and H. Wang. Multimodal transfer deep learning with applications in audio-visual recognition. *arXiv preprint arXiv :1412.3121*, 2014.

[27] O. Ménard and H. Frezza-Buet. Model of multi-modal cortical processing : Coherent learning in self-organizing modules. *Neural Networks*, 18(5) :646–655, 2005.

[28] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34 :14200–14213, 2021.

[29] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[30] F. P. Ottes, J. A. V. Gisbergen, and J. J. Eggermont. Visuomotor fields of the superior colliculus : A quantitative model. *Vision Research*, 26(6) :857–873, 1986.

[31] G. I. Parisi, J. Tani, C. Weber, and S. Wermter. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research*, 43 :208–221, 2017.

[32] J.-C. Quinton and B. Girau. Predictive neural fields for improved tracking and attentional properties. In *The 2011 International Joint Conference on Neural Networks*, pages 1629–1636. IEEE, 2011.

[33] J.-C. Quinton and L. Goffart. A unified dynamic neural field model of goal directed eye movements. *Connection Science*, 30(1) :20–52, 2018.

[34] Y. Sandamirskaya. Dynamic neural fields as a step toward cognitive neuromorphic architectures. *Frontiers in Neuroscience*, 7 :276, 2014.

[35] C. Schauer and H. M. Gross. Design and optimization of Amari neural fields for early auditory-visual integration. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 4, pages 2523–2528, 2004.

[36] G. Schöner, J. Spencer, and DFT Research Group. *Dynamic Thinking : A Primer on Dynamic Field Theory*. Oxford Series in Developmental Cognitive Neuroscience. Oxford University Press, 2015.

[37] S. Stojanov, A. Thai, and J. M. Rehg. Using shape to categorize : Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1798–1808, 2021.

[38] W. Taouali, L. Goffart, F. Alexandre, and N. P. Rougier. A parsimonious computational model of visual target position encoding in the superior colliculus. *Biological Cybernetics*, 109(4) :549–559, 2015.

[39] J. Tekülve, A. Fois, Y. Sandamirskaya, and G. Schöner. Autonomous sequence generation for a neural dynamic robot : scene perception, serial order, and object-oriented movement. *Frontiers in neurorobotics*, 13 :95, 2019.

[40] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[41] M. Vavrečka and I. Farkaš. A multimodal connectionist architecture for unsupervised grounding of spatial language. *Cognitive Computation*, 6(1) :101–112, 2014.

[42] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[43] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. Deep multimodal representation learning from temporal data. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5066–5074, 2017.

Détection de la controverse : une approche basée sur les réseaux de neurones, appliquée aux graphes et aux textes

S. Benslimane[1], J. Aze[1], S. Bringay[2,1], C. Mollevi[3,4], M. Servajean[2,1]

[1] LIRMM UMR 5506, CNRS, University of Montpellier, Montpellier, France
[2] AMIS, Paul Valery University, Montpellier, France
[3] Institut du Cancer Montpellier (ICM), Montpellier, France
[4] Institut Desbrest d'Epidémiologie et de Santé Publique, UMR Inserm

prenom.nom@lirmm.fr

Résumé

Cet article propose une approche de détection de la controverse dans les réseaux sociaux, basée sur la structure d'une discussion et de ses caractéristiques textuelles. La méthode proposée s'appuie sur les réseaux de neurones graphiques (Graph Neural Networks ou GNN) pour encoder la représentation graphique de la discussion (y compris les textes) dans un vecteur multidimensionnel. Ce dernier est utilisé pour classer les fils de discussions comme étant controversés ou non. Les expériences menées sur différents jeux de données montrent l'impact positif de la combinaison des caractéristiques textuelles et structurelles.

Mots-clés

Détection de controverse, Réseaux de neurones graphiques, Traitement naturel du langage, Reddit

Abstract

This paper proposes a controversy detection approach based on both graph structure of a discussion and text features. Our proposed approach relies on Graph Neural Network (GNN) to encode the graph representation (including its texts) in an embedding vector before performing a graph classification task. The latter will classify the post as controversial or not. Conducted experiments using different real-world datasets show the positive impact of combining textual features and structural information.

Keywords

Controversy detection, Graph neural networks, Natural language processing, Reddit

1 Introduction

Cet article [1] est un résumé de la publication réalisée pour la conférence WISE 2021 [1] et étendue dans le journal World Wide Web [2]. La disponibilité d'un grand nombre de sources de données et l'émergence de réseaux sociaux, tels que Twitter et Reddit, ont accru la connectivité sociale des personnes, ce qui leur permet d'exprimer, de propager, de partager et de contester facilement des opinions. Dans cet article, nous étudions le phénomène social de la controverse. Un contenu controversé peut être défini comme un contenu attirant des avis et opinions divergents, tant positifs que négatifs [4]. La détection précoce de ces sujets est importante, pour éviter par exemple la désinformation ou les discussions haineuses. Cependant, il s'agit d'une tâche difficile qui doit être effectuée sur un grand nombre de contenus, en constante évolution et couvrant un large éventail de thématiques. La controverse évolue au cours du temps et selon les communautés engagées.

Pour résoudre cette tâche, on trouve dans la littérature trois types de travaux : **(i)** les approches basées sur le contenu, **(ii)** celles basées sur la structure et **(iii)** celles considérées comme hybrides. Les premières utilisent uniquement les caractéristiques textuelles des messages [5]. Cependant, l'interprétation des messages et des termes utilisés étant subjectif, l'information comprise dans ces textes peut être différente selon certains facteurs, tels que la culture ou la langue des communautés, et doit donc être traitée avec précaution. Le second type d'approche se base sur les interactions entre utilisateurs, révélées par des informations structurelles issues du graphe des interactions de ces utilisateurs (e.g. propriété de connectivité ou de centralité) [3]. Enfin, des études récentes combinent les informations issues du contenu et de la structure [9].

Nous présentons dans cet article une nouvelle approche hybride de détection de la controverse, basée sur les réseaux de neurones graphiques (Graph Neural Networks) afin de combiner les informations textuelles et structurelles. L'originalité de notre approche réside dans l'utilisation de méthodes GNNs pour représenter les utilisateurs (nœuds) dans un espace euclidien à faible dimension, en tenant compte des informations structurelles. Les deux architectures proposées, l'une basée sur une représentation hiérarchique du graphe, l'autre sur des mécanismes d'attention, diffèrent largement de l'approche de [10]. Nos expérimentations se focalisent sur des données réelles du réseau social Reddit, même si notre méthode est applicable à tout autre média social suivant quelques adaptations lors de la construction du graphe.

1. L'article a reçu le prix du meilleur article de la conférence WISE.

2 Méthode

Notre approche se décompose en 4 étapes :

Étape 1 : Construction du graphe. Un fil de discussion est représenté sous la forme d'un graphe non orienté où un nœud représente un utilisateur et une arête entre 2 nœuds représente une réponse d'une personne à une autre.

Étape 2 : Caractéristiques des utilisateurs. Chaque utilisateur est représenté par les contenus qu'il a publiés dans la discussion. Récemment, différents modèles de langage NLP tels que BERT, pré-entraînés sur un large corpus, ont été proposés pour améliorer la représentation dynamique du texte. Nous extrayons, pour chaque texte, un vecteur le représentant à partir du modèle BERT, et nous agrégeons ensuite ces vecteurs par utilisateur.

Étape 3 : Encodage du graphe. Cette étape vise à représenter l'ensemble du graphe sous la forme d'un vecteur à faible dimension. Ce dernier sera utilisé en entrée de la dernière étape de classification du graphe. Récemment, différentes approches basées sur les GNNs ont été proposées pour adapter les architectures d'apprentissage profond aux données de type graphe [6, 7]. L'idée principale est de considérer chaque nœud du graphe comme un nœud de calcul, et d'apprendre à partir des GNNs un plongement dans un espace vectoriel représentant les nœuds. Cette étape exploite à la fois les caractéristiques des nœuds de la couche précédente ainsi que la représentation de ces voisins. Ensuite, la représentation de ces nœuds est agrégée afin de représenter le graphe complet. Deux stratégies sont proposées pour cette représentation vectorielle du graphe : la première basée sur les représentations hiérarchiques d'un graphe par des réseaux convolutifs [8] et la seconde basée sur des scores d'attention entre les nœuds [7].

Étape 4 : Classification du graphe. À l'aide du vecteur représentant le graphe et d'un réseau de neurones, le fil de discussion est ensuite classifié, controversé ou non.

3 Expériences

Les expériences ont été menées sur plusieurs jeux de données (subreddits) provenant de Reddit [4], chacun de ces jeux comprenant des milliers de fils de discussions, composés de leurs messages respectifs. En comparant nos résultats à des méthodes utilisant ces données et se basant soit seulement sur la structure [4], soit sur la structure combinée à des informations textuelles [4, 10], les deux méthodes proposées obtiennent des résultats équivalents en termes de précision de classification des fils de discussions controversés ou non, voire supérieurs sur certains jeux de données, notamment pour notre approche basée sur la représentation hiérarchique du graphe. Des expériences, omettant les informations textuelles dans le graphe, ont montré que la précision de la classification dans certains jeux de données donnaient de moins bons résultats.

4 Conclusion

Nous avons présenté dans ce résumé nos travaux autour de la détection de la controverse sur Reddit, combinant les informations structurelles et textuelles autour de l'utilisation des réseaux de neurones graphiques (GNNs). Nous prévoyons d'étendre ces travaux pour quantifier la controverse et prendre en compte la temporalité.

Références

[1] Samy Benslimane, Jérôme Azé, Sandra Bringay, Maximilien Servajean, and Caroline Mollevi. Controversy detection : a text and graph neural network based approach. In *Web Information Systems Engineering*, 2021.

[2] Samy Benslimane, Jérôme Azé, Sandra Bringay, Maximilien Servajean, and Caroline Mollevi. A text and GNN based controversy detection method on social media. *World Wide Web*, 26(2) :799–825, 2023.

[3] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1) :3 :1–3 :27, 2018.

[4] Jack Hessel and Lillian Lee. Something's brewing ! early prediction of controversy-causing posts from discussion features. In *ACL Human Language Technologies, Volume 1*, pages 1648–1659, 2019.

[5] Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. Probabilistic approaches to controversy detection. In *25th ACM International Conference on Information and Knowledge Management, CIKM*, pages 2069–2072, 2016.

[6] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Int. Conf. on Learning Representations*, 2017.

[7] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Int. Conf. on Learning Representations*, 2018.

[8] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, pages 4805–4815, 2018.

[9] Juan Manuel Ortiz De Zarate and Esteban Feuerstein. Vocabulary-based method for quantifying controversy in social media. In *Int. Conf. on Conceptual Structures*, volume 12277, pages 161–176, 2020.

[10] Lei Zhong, Juan Cao, Qiang Sheng, Junbo Guo, and Ziang Wang. Integrating semantic and structural information with graph convolutional network for controversy detection. In *Proceedings of the 58th Annual Meeting of ACL*, pages 515–526, July 2020.

Remerciements

Ce projet a été soutenu par des subventions du fond de dotation Janssen Horizon. L'accès aux ressources HPC de l'IDRIS a été accordé dans le cadre de l'allocation AD011012604 par GENCI.

Session 3 : Motifs et sémantique

Extraction de co-localisations sous contrainte de la structure spatiale

R. Govan[1], N. Selmaoui-Folcher[1], A. Giannakos[2], P. Fournier-Viger[3]

[1] Université de la Nouvelle-Calédonie, ISEA
[2] Université de Picardie Jules Verne, EPROAD
[3] Shenzhen University, Big Data Institute

{rodrigue.govan, nazha.selmaoui}@unc.nc

Résumé

Une co-localisation est un sous-ensemble de caractéristiques géographiquement proches les unes des autres. La majorité des méthodes existantes utilise des mesures standards de proximité (par exemple la distance euclidienne). Cependant, ces mesures ne sont pas les plus adaptées selon la zone d'étude. La structure spatiale doit être prise en compte. Cet article propose CSS-Miner, une approche d'extraction de co-localisations sous la contrainte de la structure spatiale. Ici, nous utilisons comme contrainte le réseau routier d'une ville. CSS-Miner a été appliqué sur deux jeux de données des villes de Paris et Chicago en sélectionnant différents points d'intérêt.

Mots-clés

co-localisation, extraction de connaissances, données spatiales, structure spatiale.

Abstract

Spatial co-location pattern is a subset of object features that are geographically close to one another. The majority of existing methods employ standard proximity measures (e.g. Euclidean distance). However, depending on the study area, these standard measures do not work well. The spatial structure has to be considered. This article proposes CSS-Miner, a co-location pattern mining approach under the spatial structure constraint. In this case, we use the street network of a city as a constraint. CSS-Miner has been applied to two datasets from the cities of Paris and Chicago by selecting different POIs.

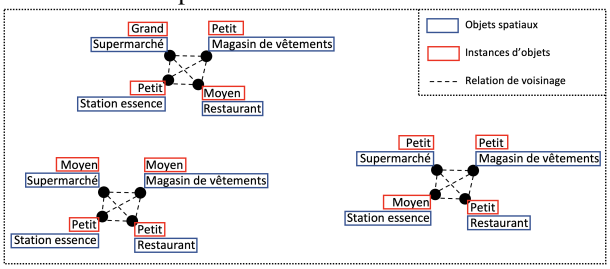
Keywords

co-location, knowledge mining, spatial data, spatial structure.

1 Introduction

Dans le domaine de la fouille des données, l'extraction de co-localisations est une des méthodes permettant d'extraire de l'information et des connaissances prenant en compte la dimension spatiale des données et pouvant aider les décideurs. Une co-localisation (ou motif spatial) est un sous-ensemble de caractéristiques spatiales qui sont fréquemment localisées dans une même zone géographique. Si l'on prend l'exemple des centres commerciaux dans une ville, ils contiennent fréquemment de grands supermarchés, des petits restaurants, des magasins de vêtements et une station essence. De ce fait, si l'on considère les centres commerciaux autour d'une ville comme une co-localisation, ses objets spatiaux sont décrits par les supermarchés, restaurants, magasins de vêtements et stations essence. Un supermarché peut être un objet spatial avec des attributs/instances tels que petit, moyen et grand. Un graphe représentant cet exemple est illustré dans la Fig. 1.

FIGURE 1 – Exemple d'une co-localisation avec ses objets spatiaux et leurs instances.



De nombreux travaux liés à l'extraction de co-localisations ont été menés [15, 19, 27]. Les méthodes d'extraction de co-localisations ont été appliquées dans divers domaines d'études, telles que l'analyse de concentration d'entreprises [6], l'explication de phénomènes anthropiques [1] et plus précisément l'analyse de l'érosion des sols [18]. Cependant, malgré un grand nombre de cas d'usage, la plupart de ces méthodes utilise des fonctions standards de distance (par exemple la distance euclidienne) pour mesurer la proximité des objets spatiaux. Pour certains cas d'usage, il est conseillé d'utiliser d'autres mesures de distance. Dans le cas d'une analyse de comportements de la population d'une ville à travers ses lieux d'intérêt, la distance euclidienne entre deux objets spatiaux n'a plus lieu d'être, car la longueur du chemin parcouru entre ces deux objets peut être significativement différente de sa distance euclidienne.
Selon la zone d'étude, il peut s'avérer essentiel de prendre en compte la structure spatiale, puisqu'elle impacte la dis-

tribution des objets spatiaux dans un espace donné. En employant la distance euclidienne dans l'analyse d'une zone urbaine, nous perdons totalement l'information sur la structure spatiale de cette zone. Afin de garder cette structure, il est nécessaire d'utiliser d'autres mesures de distance. Cependant, garder l'information de la structure spatiale de la zone d'étude peut augmenter la complexité de l'analyse, en termes de pré-traitement de données, mais aussi au niveau des paramètres à définir.

Dans cet article, nous proposons CSS-Miner (CSS pour **C**o-localisation sous contrainte de la **S**tructure **S**patiale), une méthode d'extraction de co-localisations sous contrainte de la structure spatiale de la zone d'étude. En premier lieu, la méthode construit un graphe sous cette contrainte en utilisant un algorithme de recherche du plus court chemin. Puis, CSS-Miner extrait les cliques maximales pour obtenir les motifs spatiaux. Pour les tests, la méthode proposée a été appliquée sur deux jeux de données des villes de Paris et Chicago, nous permettant d'extraire de nouveaux motifs pertinents, mais aussi de filtrer des motifs non pertinents.

L'article est organisé comme suit. La section 2 présente un bref état de l'art sur l'extraction de motifs spatiaux, en particulier avec l'approche par les événements. La section 3 décrit l'approche proposée qui prend en compte la contrainte de la structure spatiale. Puis, la section 4 présente les données utilisées dans cet article et les motifs extraits. Enfin, une conclusion est tirée et des perspectives sont discutées.

2 État de l'art

Dans leur article, Huang et al. [12] ont présenté deux approches d'extraction de motifs spatiaux : l'approche par les transactions et l'approche par les événements.

L'approche par les transactions consiste à transformer les objets spatiaux en données séquentielles dans le but d'appliquer les algorithmes standards d'extraction d'*itemsets* fréquents. Cette approche par les transactions a été initialement introduite par Koperski et al. [15].

L'approche par les événements se focalise sur la localisation des objets spatiaux et leurs proximités. Initialement proposé par Shekhar et al. [19], cette approche extrait tous les sous-ensembles d'objets qui sont géographiquement proches les uns des autres, aussi appelés co-localisations. De la même manière que l'approche par les transactions, des mesures d'intérêt ont été définies afin de ne garder que les co-localisations les plus pertinentes. Dans la littérature, nous pouvons observer que l'approche par les transactions est plus fréquemment utilisée que celle par les événements. Ces méthodes utilisent la distance euclidienne pour définir la relation de voisinage. Dans cet article, nous proposons une autre mesure de distance liée à une contrainte que nous avons nommé, la contrainte de la structure spatiale.

Pour cela, nous avons utilisé l'approche par les événements afin de tirer profit de la dimension spatiale de nos objets et leurs proximités. Pour appliquer l'approche par les événements sous notre contrainte de la structure spatiale, nous utilisons une méthode d'extraction de cliques maximales afin d'obtenir nos co-localisations. De ce fait, les sous-sections 2.1 et 2.2 suivantes donnent respectivement, un aperçu sur les approches d'extraction de cliques maximales et un aperçu des principales études menées sur l'extraction de co-localisations et leurs mesures d'intérêt.

2.1 Extraction de cliques maximales

(**Graphe complet**) Soit $G = (V, E)$ un graphe avec $V = \{v_1, v_2, \ldots, v_n\}$ l'ensemble des sommets et $E \subseteq \{(v_i, v_j) \in V^2 \mid \forall i, j \in \{1, \ldots, n\} \text{ et } i \neq j\}$ l'ensemble des arêtes. Si deux sommets v_i et v_j sont liés i.e., $(v_i, v_j) \in E$, alors v_i et v_j sont adjacents. Un graphe est dit complet si chaque paire de sommets du graphe est liée par une arête (adjacent).

(**Clique**) Soit $G = (V, E)$ un graphe et $g = (V_g, E_g)$ un sous-graphe tel que $V_g \subseteq V$ et $E_g \subseteq \{(v_{g,i}, v_{g,j}) \in E \mid v_{g,i} \in V_g \wedge v_{g,j} \in V_g\}$. Une clique de G est un sous-graphe $g \subseteq G$ tel que g est complet.

(**Clique maximale**) Pour $G = (V, E)$ un graphe donné et $g \subset G$ une clique, la clique g est dite maximale si et seulement s'il n'existe pas de clique g' telle que $g \subset g' \subseteq G$.

Avec l'approche par les événements, il est possible d'extraire nos motifs spatiaux par l'extraction de cliques maximales. Valiant [24] a démontré qu'énumérer toutes les cliques maximales est un problème #P-complet. De la même manière que les méthodes d'extraction d'itemsets fréquents, l'extraction des cliques maximales s'effectue en testant chaque combinaison de sommets d'un graphe afin d'obtenir les cliques maximales. En particulier, nous pouvons mentionner les algorithmes proposés par Bron et al. [4] et Tomita et al. [21] pour leur complexité de $O(3^{n/3})$ dans le pire scénario avec un graphe à n noeuds qui est optimal en fonction de n, mais aussi Moon et al. [17] et Cazals et al. [5] qui considèrent un appel récursif dans leur algorithme pour améliorer l'extraction des cliques maximales.

Dans la littérature, les méthodes d'extraction de cliques maximales sont communément utilisées afin d'extraire les co-localisations [2, 16, 22, 26]. En effet, en définissant un graphe où les sommets représentent des objets spatiaux et les arêtes représentant leurs voisinages puis en appliquant une méthode d'extraction de cliques maximales, nous pouvons obtenir les sous-ensembles d'objets qui sont tous voisins entre eux. Ainsi, dans cet article, nous utiliserons l'approche proposée dans [4] puis adaptée dans [21] pour sa rapidité étant donné la taille de nos jeux de données détaillés dans la section 4.1.

2.2 Extraction de co-localisations et leurs mesures d'intérêt

Le principe de l'approche par les événements est de projeter les données spatialisées par leurs coordonnées et de définir la proximité entre ces objets spatiaux dans le but d'extraire les motifs. Dans cette section, nous rappelons le formalisme de l'extraction de co-localisations proposé par Shekhar et Huang [19], Huang et al. [12] et Yoo et Shekhar [27]. Soit $\mathcal{F}$ un ensemble de caractéristiques et $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ une base de données d'objets spatiaux. Chaque objet dans $\mathcal{O}$ se compose d'un triplet `<object_id, localisation, c>`, où `c` $\in \mathcal{F}$.

Par exemple, dans la Fig. 2.2, $\mathcal{F} = \{A, B, C\}$, $\mathcal{O} = \{A_1, B_2, \ldots, C_3\}$ avec $A_1 =< 1, (x_1, y_1), A >$, $B_2 =< 2, (x_2, y_2), B >$, etc. Une co-localisation $\mathcal{C}$ est un sous-ensemble de caractéristiques de $\mathcal{F}$ associées à des objets spatiaux appartenant à $\mathcal{O}$. Ces co-localisations représentent des caractéristiques apparaissant fréquemment dans des objets voisins. La relation de voisinage est définie par une relation binaire $\mathcal{R}(o, o')$ entre deux objets spatiaux o et o'. En fonction des besoins de l'utilisateur et des cas d'usage, $\mathcal{R}$ peut se baser sur un seuil de distance entre deux objets ou sur l'intersection de ces objets. Plusieurs travaux ont été menés, incluant Yoo et Shekhar [27], Wang et al. [25] et Kim et al. [14]. La plupart de ces travaux se basent généralement sur la distance euclidienne pour quantifier la proximité entre les objets spatiaux. Mais plus récemment, des travaux ont été menés sur l'extraction de co-localisations utilisant différentes relations de voisinage. Yu [28] a proposé dans son article la longueur du plus court chemin comme mesure de distance. Cependant, en proposant cette méthode, l'auteur ajoute un paramètre qui est le nombre maximum d'objets voisins. En définissant ce paramètre, cela assure un algorithme d'extraction rapide mais cela limite aussi la taille des co-localisations ce qui peut passer outre certains motifs qui peuvent s'avérer pertinents. Puis, Yu et al. [29] ont ajouté une fonction de décroissance de la distance afin de déterminer la dépendance spatiale entre les objets spatiaux. La fonction consiste à pondérer la contribution d'une co-localisation dans la mesure d'intérêt.

L'approche par les événements est basée sur la définition d'un seuil de voisinage. Pour déterminer si deux objets sont géographiquement proches, nous fixons un seuil de distance maximale d. Une fois que le voisinage est défini, le graphe est construit avec les objets spatiaux représentant les sommets. Deux sommets sont adjacents si leur distance associée respecte le seuil d (i.e., si la mesure de distance entre deux sommets est inférieure à d).

Pour les méthodes d'extraction de motifs spatiaux, des mesures d'intérêt ont été développées afin de quantifier la pertinence des motifs. Pour mesurer si une co-localisation est pertinente ou non, l'indice de participation, basé sur le ratio de participation est utilisé. L'indice de participation est aussi appelé la prévalence. Nous parlons ainsi de motif spatial prévalent.

(**Ratio de participation**) Soit $\mathcal{C}$ une co-localisation. Pour une instance $f_i \in \mathcal{C}$, le ratio de participation est défini par :

$$Pr(f_i, \mathcal{C}) = \frac{|\{ \text{ instances de } f_i \text{ participant à } \mathcal{C})\}|}{|\{ \text{ instances de } f_i \ \}|} \quad (1)$$

En prenant l'exemple de la Fig. 2, soit $\mathcal{C} = \{A, B\}$ une co-localisation et $I_{\mathcal{C}} = \{(A_1, B_1), (A_1, B_2), (A_3, B_4)\}$ l'ensemble des instances de $\mathcal{C}$. Avec A et B, deux caractéristiques ayant respectivement, 3 et 4 instances, nous avons :
$Pr(A, \{A, B\}) = \frac{|\{A_1, A_3\}|}{|\{A_1, A_2, A_3\}|} = \frac{2}{3}$ et
$Pr(B, \{A, B\}) = \frac{|\{B_1, B_2, B_4\}|}{|\{B_1, B_2, B_3, B_4\}|} = \frac{3}{4}$.

(**Indice de participation**) Soit $\mathcal{C}$ une co-localisation, $I_{\mathcal{C}} = \{I_1^{\mathcal{C}}, ..., I_k^{\mathcal{C}}\}$ l'ensemble des instances de $\mathcal{C}$ et $\mathcal{F} = \{f_1, \ldots, f_n\}$ l'ensemble des caractéristiques de la base de données $\mathcal{O}$. L'indice de participation est défini par :

$$Pi(\mathcal{C}) = \min_{f_i \in \mathcal{C}} Pr(f_i, \mathcal{C}) \quad (2)$$

En utilisant l'exemple précédent, nous avons pour indice de participation :

$$\begin{aligned} Pi(\{A, B\}) &= \min_{f_i \in \{A,B\}} Pr(f_i, \{A, B\}) \\ &= \min(\frac{2}{3}, \frac{3}{4}) = \frac{2}{3} \end{aligned}$$

Dans cet article, la mesure de prévalence sera utilisée afin de déterminer si les co-localisations dans la section 4 sont pertinentes ou non.

Cet indice de participation a été défini pour des données spatialisées de type point. Cependant, avec la croissance considérable de collecte de données, nous avons aujourd'hui différents types de données (lignes, polygones, ...). Dans leur contribution, Akbari et al. [1] ont proposé une variante de l'indice de participation pour chaque type de données. Pour prendre en compte tout type de données, les auteurs ont proposé de restreindre chaque région d'intérêt en appliquant le diagramme de Voronoï à partir des attributs cibles. Dans leur cas, la cible est une caractéristique spatiale qui est de type point. Une fois que le diagramme de Voronoï est appliqué, chaque cellule correspond à une instance de co-localisation. Puis, lors des calculs de prévalence, pour prendre en compte les données de type ligne/polygone, les auteurs pondèrent chaque objet spatial par la proportion de l'objet présente dans la cellule de Voronoï.

Puisque nous n'appliquons pas le diagramme de Voronoï dans cet article, nous n'allons pas utiliser la variante de la prévalence proposée dans [1]. De ce fait, nous allons devoir réduire nos données de type polygone à un seul point en prenant le centre de gravité du polygone (la moyenne de toutes les coordonnées du polygone).

Comme mentionné précédemment, les travaux autour de l'approche par les événements utilisent généralement la distance euclidienne comme relation de voisinage, ignorant la structure spatiale de la zone d'étude. Dans cet article, nous allons donc utiliser la longueur du plus court chemin existant comme mesure de proximité.

2.3 Recherche du plus court chemin

Au cours des dernières décennies, la recherche du plus court chemin a été un problème majeur dans la théorie des graphes. La vitesse de recherche dépend entièrement du nombre de sommets et d'arêtes dans un graphe. Une des premières solutions a été présentée par Dijkstra [7]. Pour un graphe de $|V|$ sommets et $|E|$ arêtes, l'algorithme de Dijkstra a une complexité polynomiale de $O((|V| + |E|) \log n)$. Puis, de nouvelles méthodes ont été développées afin d'accélérer la recherche du plus court chemin [13, 9, 20].

Plus récemment, Varia et Kurasova [23] ont proposé une version accélérée de l'algorithme de Dijkstra, en ajoutant deux composantes : le recherche bidirectionnelle et la parallélisation. Afin de chercher le plus court chemin entre

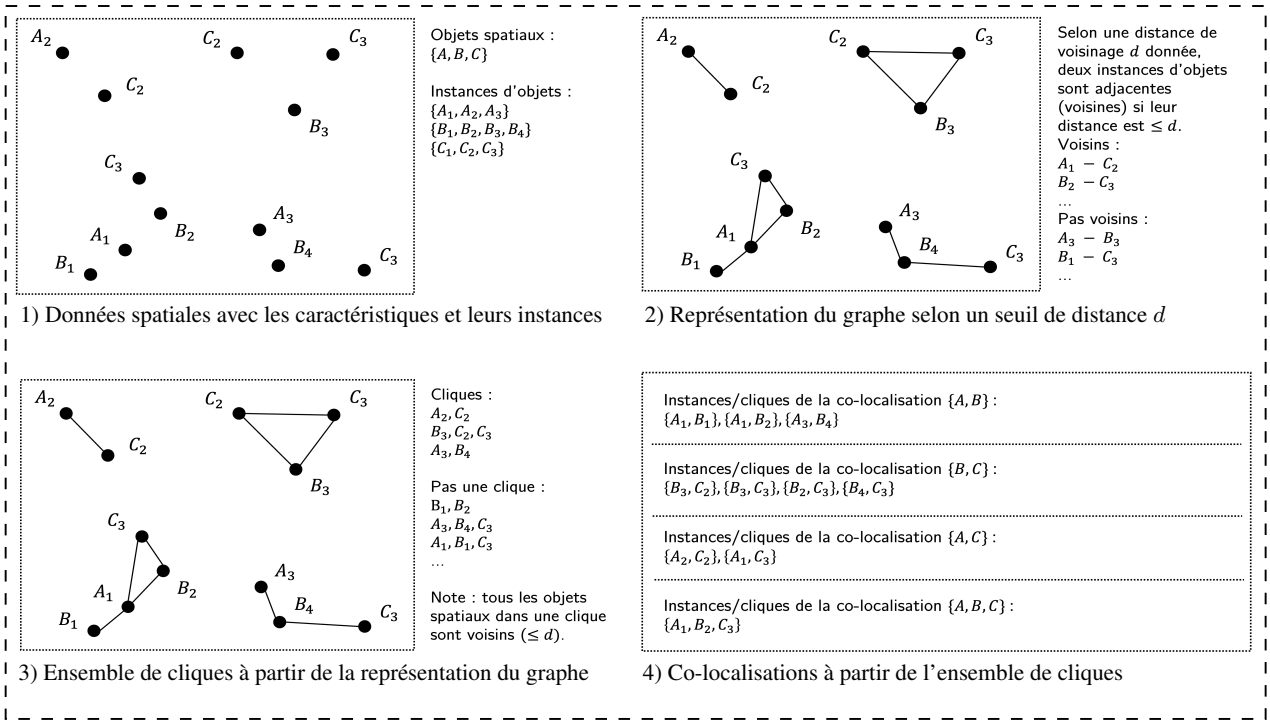


FIGURE 2 – Exemple de co-localisations basées sur un ensemble de cliques d'un jeu de données spatialisées.

deux sommets v_i et v_j, les auteurs appliquent l'algorithme de Dijkstra de v_i vers v_j et de v_j vers v_i. Puisque l'algorithme de Dijkstra est basé sur une file d'attente avec priorité, la composante bidirectionnelle utilise deux files d'attente. Cependant, lors des deux recherches, chaque recherche avance l'une après l'autre. L'avantage étant que les deux recherches seront plus courtes, mais elles avancent au tour par tour. Pour palier ce problème, les auteurs ont donc ajouté la parallélisation. Avec cette composante, les deux recherches avancent en même temps. Par l'ajout de ces deux composantes, selon les auteurs, le temps d'exécution de l'approche proposée est au minimum divisé par deux par rapport à la méthode initiale, selon le nombre de sommets dans un graphe.
Afin de prendre en compte la contrainte de la structure spatiale et d'accélérer notre processus, l'algorithme de Dijkstra bidirectionnel parallélisé sera donc utilisé.

3 Méthodes

Considérons un ensemble d'objets spatiaux $\mathcal{O}$ avec un ensemble de caractéristiques $\mathcal{F}$. Soit G_S un graphe représentant la structure spatiale telle que $G_S = (V_S, E_S)$ où V_S est l'ensemble des sommets et E_S est l'ensemble des arêtes.

3.1 Prise en compte de la contrainte de la structure spatiale

Pour analyser des points d'intérêt dans une structure spatiale (par exemple une zone urbaine), la longueur du plus court chemin parcouru entre deux localisations (x_i, y_i) et (x_j, y_j) associées respectivement, aux objets spatiaux o_i et o_j, semble la plus adaptée. Afin d'extraire les co-localisations, nous avons cherché à inclure la contrainte de la structure spatiale.
L'intégration de la contrainte est réalisée suivant plusieurs étapes :

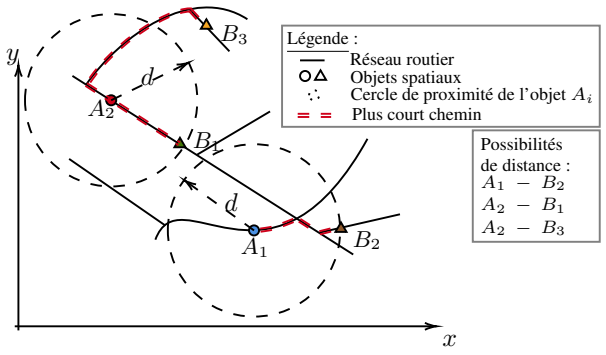
1. Pour chaque objet spatial $o_i \in \mathcal{O}$, nous l'associons dans la structure spatiale G_S au plus proche objet $o_S \in V_S$ (par la distance euclidienne) ;
2. Nous déterminons le plus court chemin pour chaque objet de V_S aux autres objets localisés dans un cercle de rayon d selon la distance euclidienne ;
3. Si la longueur du plus court chemin entre deux objets de V_S est inférieure ou égale à d, alors ces objets sont considérés comme voisins.

Afin de ne pas déterminer inutilement des plus courts chemins, nous n'appliquons pas l'algorithme de recherche du plus court chemin entre deux objets de V_S si ces deux objets ne sont pas respectivement associés à deux objets spatiaux de $\mathcal{O}$. Même si la distance euclidienne ne définit pas la relation de voisinage dans notre approche, nous l'utilisons tout de même dans le but d'élaguer le nombre de plus courts chemins calculés. Appliquer un seuil de distance par un cercle de rayon d va nous éviter de calculer des plus courts chemins qui nous le savons déjà, seront supérieurs à notre seuil. En effet, par inégalité triangulaire, un objet spatial localisé en dehors du cercle de rayon d d'un autre objet spatial a un plus court chemin qui sera supérieur à d.

3.2 Construction du graphe

Pour extraire nos motifs spatiaux (co-localisations) qui sont les cliques maximales, nous avons choisi de construire le

FIGURE 3 – Trois possibilités de distances que CSS-Miner peut rencontrer.



graphe $G = (\mathcal{O}, E_{\mathcal{O}})$ (sous la contrainte de la structure spatiale) où $E_{\mathcal{O}} = \{(o_i, o_j) \mid \exists(o_{S,i}, o_{S,j}) \in E_S, D_{sp}(o_{S,i}, o_{S,j}) \leq d, \forall(i,j) \in [\![1,n]\!]^2, i \neq j\}$ avec $o_{S,i}$ représentant l'objet de la structure spatiale associé à l'objet spatial $o_i \in \mathcal{O}$ et D_{sp} représentant la distance obtenue par l'algorithme du plus court chemin de Dijkstra.

La Fig. 3 illustre les trois possibilités que CSS-Miner peut rencontrer où les objets A_i et B_i sont des objets de V_S expliqué dans la section 3.1. Avec d en tant que rayon de distance et seuil de chemin le plus court, nous avons pour distance euclidienne, $d_2(A_2, B_3) > d$, donc CSS-Miner ne lancera pas l'algorithme de recherche du plus court chemin et ne considérera pas A_2 et B_3 comme voisins sous la contrainte de la structure spatiale. De l'autre côté, nous avons $d_2(A_1, B_2) \leq d$, donc notre algorithme lancera l'algorithme de Dijkstra. Cependant, nous avons $D_{sp}(A_1, B_2) > d$, donc nous ne considérerons pas A_1 et B_2 comme géographiquement proches (comme voisins) sous contrainte. Enfin, nous avons dans la Fig. 3, l'objet spatial B_1 localisé dans le cercle de rayon d et de centre A_2. CSS-Miner cherchera le plus court chemin, obtiendra $D_{sp}(A_2, B_1) < d$ et ainsi, considérera ces deux objets spatiaux comme voisins. Ainsi, leur valeur associée dans la matrice d'adjacence sera égale à 1.

Au final, dans CSS-Miner, nous manipulons deux graphes : Le premier représentant la structure spatiale et le second représentant le voisinage de nos jeux de données construit à partir du premier graphe.

4 Résultats

Dans cet article, nous appliquons notre approche sur deux jeux de données réelles. Le premier est créé en collectant des données de la plateforme *OpenData* de Paris [1] et sa périphérie [2] (voir la description des données dans le tableau 1). Le second jeu de données est aussi créé en collectant des données de la plateforme *OpenData* de Chicago [3] (voir la description des données dans le tableau 2).

Pour chaque jeu de données, le processus entier a été réalisé via le langage Python, sur un ordinateur avec un processeur AMD Ryzen 7 3700X 8-core, 64Go de RAM et une carte graphique NVIDIA GeForce RTX 2060 SUPER avec 8Go de RAM dédiée. Les temps d'exécution de tout le processus sur les données de Paris et Chicago ont été respectivement, d'environ 2 et 5 heures.

Cette étude de cas vise à analyser et comprendre le comportement de la jeunesse dans une grande ville. Cette approche reste tout de même générique, puisque nous pouvons l'appliquer dans une analyse de population plus large selon leurs catégories socio-professionnelles, par exemple : Quelles sont les habitudes quotidiennes d'un cadre face à un étudiant ? Une autre analyse de lieux d'intérêt peut aussi être pertinente pour le développement d'un outil d'aide à la décision afin de contribuer au développement du tourisme d'une ville. Finalement, l'analyse de lieux d'intérêt reste un sujet d'étude très varié.

TABLE 1 – Description des données de Paris.

Variable	Attributs	# Modalités	# Objets
Lycées	Type	7	239
Cinémas	# Sièges (**)	5	85
(*) Vélos	Capacité (**)	8	996
Parcs	Type	9	722
(*) Métros	Ligne	16	326

(*) : La variable concernent des stations.
(**) : Les données ont été discrétisées par quantile.

TABLE 2 – Description des données de Chicago.

Variable	Attributs	# Modalités	# Objets
Lycées	Type	13	142
(*) Bus	# Lignes	12	5 606
(*) Tramway	# Lignes	6	124
Fast Food		1	877
(*) Vélos	Capacité (*)	8	1 402
Parcs	Type	13	613

(*) : La variable concernent des stations.
(**) : Les données ont été discrétisées par quantile.

4.1 Pré-traitement des données

Pour intégrer notre contrainte de la structure spatiale, il est nécessaire d'avoir accès à cette information. Dans notre cas, nous avons utilisé le réseau routier comme structure spatiale. Ici, nous supposons que le parcours entre deux objets spatiaux s'effectue à pied. Ce choix est dû au fait que nous avons souhaité utiliser des données disponibles uniquement en libre accès, là où le trafic routier n'est pas toujours disponible.

Pour accéder aux réseaux routiers de Paris et Chicago, nous avons utilisé la méthode `OSMnx` [3]. Les auteurs ont rendu `OSMnx` simple d'utilisation. En effet, nous pouvons récupérer un réseau routier à partir du nom de la ville ou en fournissant les coordonnées de la zone d'étude via sa librairie Python. Une fois le réseau routier récupéré, il peut être représenté par un graphe avec les arêtes représentant les routes et les sommets représentant les intersections des

1. https ://opendata.paris.fr/
2. https ://data.iledefrance.fr/
3. https ://data.cityofchicago.org/

routes. Au final, le graphe associé au réseau routier de Paris contient 42 870 sommets et 241 016 arêtes tandis que le graphe associé au réseau routier de Chicago contient 184 476 sommets et 1 217 928 arêtes.

Pour les données de Paris, nous avons récupéré les données de Cinémas, Lycées, Stations de vélos, Parcs et Stations de métros. Pour restreindre le zone d'étude, nous n'avons gardé que les objets spatiaux de Paris intra-muros. Comme mentionné précédemment, nous avons choisi d'analyser que les données de type point. De ce fait, les Parcs qui sont initialement des polygones ont été réduits en un point, en l'occurrence le centre gravité. De plus, puisque les co-localisations ne sont compatibles qu'avec des données catégorielles, nous avons discrétisé deux variables (Cinémas et Stations de vélos) par quantile. Au final, nous avons 2 968 objets spatiaux décrits dans le tableau 1.

Pour les données de Chicago, nous avons récupéré les données de Tramway, Parcs, Lycées, Stations de vélos et Chaînes de *Fast Food* de la ville. Le même processus de pré-traitement des données de Paris a été appliqué aux données de Chicago. Ainsi, nous avons 8 764 objets spatiaux qui composent nos données de Chicago.

Pour les deux jeux de données et leur structure spatiale associée, nous avons projeté toutes les coordonnées dans le système de coordonnées WGS 84 / Pseudo-Mercator (EPSG : 3857). Ce système de coordonnées nous permet de pouvoir calculer les distances en mètres.

À l'étape d'élagage, nous avons fixé un seuil sur le rayon de 500m ($d = 500$). Chaque objet ne sera comparé qu'aux objets contenus dans ce cercle de rayon d. Lors de la construction du graphe et de sa matrice d'adjacence associée, pour déterminer si deux objets (deux sommets) sont contigus, nous avons fixé le seuil de la distance à pied au même seuil du rayon ($d = 500$). Ainsi, si le plus court chemin trouvé entre deux objets est inférieur à 500m, alors leur valeur associée dans la matrice d'adjacence est égale à 1. Sinon, elle est égale à 0. De plus, pour éviter toute boucle dans le graphe (une arête reliant un sommet à lui-même), nous avons mis toute la diagonale de la matrice d'adjacence à 0.

4.2 Données de Paris

Suite au pré-traitement des données et à la construction du graphe, nous avons extrait les cliques maximales. Puisque l'article cherche à analyser le comportement de la population jeune de Paris, le tableau 3 ne montre que les co-localisations contenant la variable Lycées.

Le tableau 3 nous montre les possibles activités à proximité des lycées pour la population jeune de Paris, notamment par la présence des parcs et des cinémas. De plus, nous observons par ces co-localisations, l'omniprésence des variables Lycées et Vélos (stations de vélos en libre-service), ce qui montre aussi que la ville de Paris aide au mieux la jeunesse parisienne à se déplacer librement et en même temps pratiquer une activité sportive. Il serait intéressant d'appliquer CSS-Miner sur d'autres villes de France proposant ce service afin de pouvoir confirmer cette tendance.

Puisque CSS-Miner considère le réseau routier comme contrainte de structure spatiale, l'idée est de voir s'il y a une différence par rapport aux co-localisations sans cette contrainte, c'est-à-dire avec la distance euclidienne comme relation de voisinage. Les résultats nous montrent que les motifs extraits sous contrainte n'ont pas systématiquement une prévalence supérieure à la prévalence obtenue sans contrainte. Nous pouvons l'expliquer comme suit. Les motifs extraits sans contrainte ont été obtenus en utilisant le même seuil de distance (i.e., 500m), tout comme CSS-Miner. Par inégalité triangulaire, un chemin parcouru à pied entre deux objets spatiaux est supérieur à sa distance euclidienne. De ce fait, sans la contrainte, les cliques maximales contiennent plus d'objets, augmentant la probabilité d'avoir un grand nombre d'instances par variable, ce qui peut réduire leur prévalence. Cela explique aussi pourquoi la co-localisation {Parcs, Lycées, Vélos} avec une prévalence de 0.89, voit sa prévalence diminuer à 0.56 si la variable Cinémas y est ajoutée. En effet, en ajoutant une variable dans une co-localisation, cela augmente le nombre d'objets spatiaux contenus dans cette co-localisation, ce qui peut diminuer sa prévalence.

TABLE 3 – Prévalences des co-localisations de Paris.

Co-localisation	Prévalence sous contrainte	Prévalence sans contrainte
{Parcs, Lycées, Vélos}	0.89	0.89
{Lycées, Vélos}	0.86	0.86
{Parcs, Lycées, Vélos, Métros}	0.78	0.89
{Lycées, Cinémas, Vélos}	0.71	0.71
{Lycées, Vélos, Métros}	0.71	0.71
{Lycées, Cinémas, Vélos, Métros}	0.71	0.71
{Parcs, Lycées, Cinémas, Vélos}	**0.56**	0.44

Enfin, sans prendre en compte la contrainte de la structure spatiale, l'algorithme a extrait des motifs que CSS-Miner n'a pas extraits. Ces motifs sont : {Lycées, Métros} et {Parcs, Lycées, Cinémas, Métros} avec des prévalences égales à 0.31 et 0.14 respectivement. Ces deux motifs ont une prévalence nulle si l'on prend en compte la contrainte. Cela montre que même si les objets spatiaux sont proches par la mesure euclidienne, la longueur de leur plus court chemin ne vérifie pas notre critère de voisinage. Ces objets ne peuvent donc pas être considérés comme voisins. Au final, en prenant en compte la structure spatiale de la zone d'étude, nous pouvons extraire les motifs pertinents et filtrer les motifs non pertinents, selon la zone d'étude.

4.3 Données de Chicago

De la même manière qu'avec les données de Paris, le tableau 4 ne montre que les co-localisations contenant la variable Lycées de Chicago.

Les prévalences du tableau 4 montrent que la majorité des Lycées de Chicago ont une chaîne de *Fast Food* à proximité, donc la population jeune de Chicago sera plus tentée

FIGURE 4 – Données de Paris avec les cliques maximales extraites.

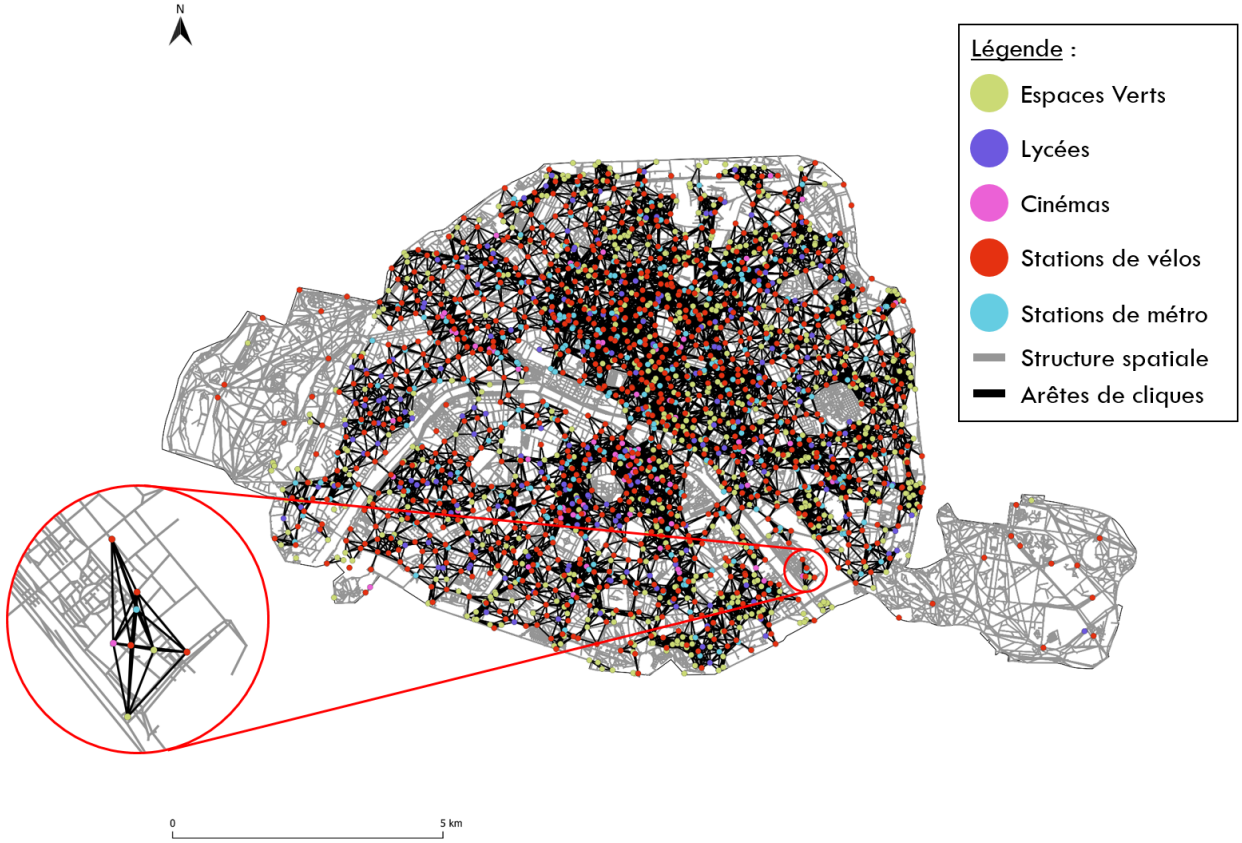


TABLE 4 – Prévalences des co-localisations de Chicago.

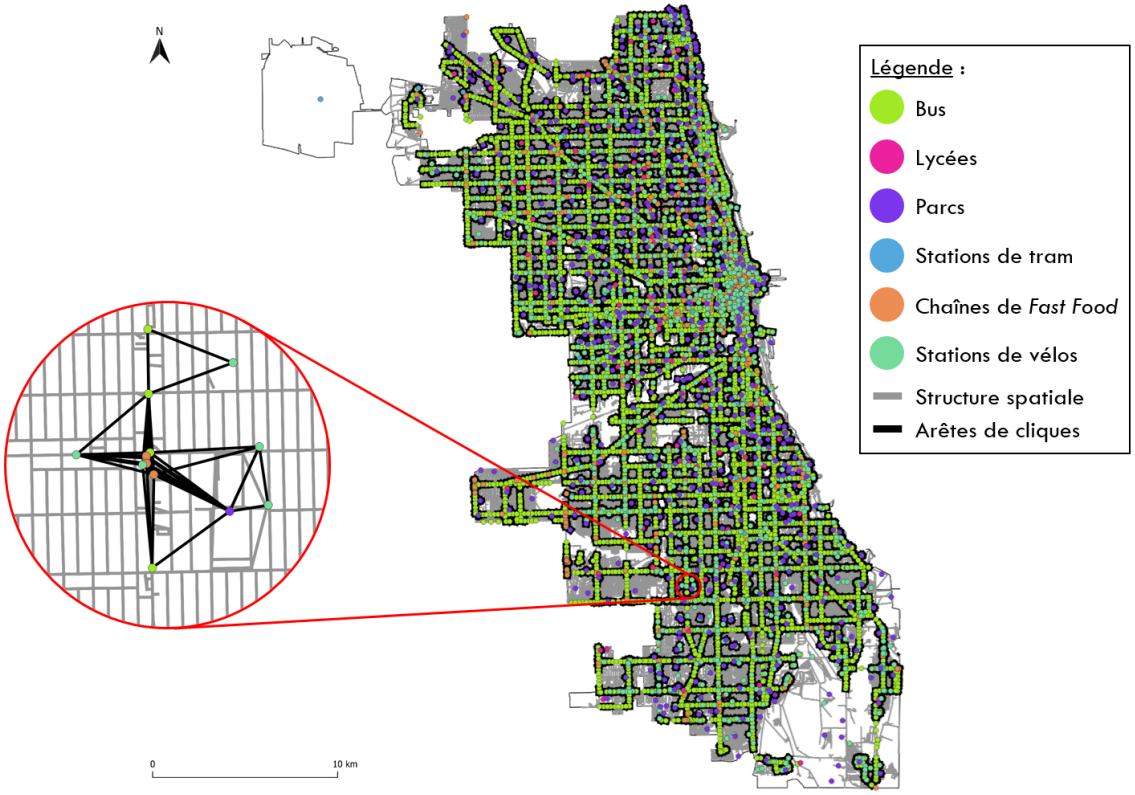
Co-localisation	Prévalence sous contrainte	Prévalence sans contrainte
{Bus, *Fast Food*, Lycées, Vélos}	**0.58**	0.5
{Bus, *Fast Food*, Lycées, Tramway, Vélos}	0.38	0.38
{Bus, *Fast Food*, Lycées}	**0.33**	0.17
{Bus, *Fast Food*, Lycées, Tramway}	0.3	0.3
{Bus, *Fast Food*, Lycées, Parcs}	0.17	0.17
{Bus, *Fast Food*, Lycées, Tramway, Vélos, Parcs}	0.15	0.15

de manger dans un *Fast Food* à midi ou en sortant d'école. L'omniprésence des variables Lycées et *Fast Food* dans nos co-localisations peut aussi alarmer la population sur la malnutrition des américains, ou du moins de la population jeune de Chicago. Pour confirmer cette affirmation, il serait intéressant d'appliquer cette approche sur les grandes villes des États-Unis et vérifier si nous pouvons extraire ces mêmes co-localisations. Il serait également intéressant d'obtenir un jeu de données recensant les *Fast Food* de Paris afin de démontrer si les chaînes de restauration rapide à Paris ciblent la jeune population de la même manière qu'à Chicago. Cependant, ce jeu de données n'est malheureusement pas disponible sur les plateformes *OpenData*.
Enfin, nous notons qu'à partir de ces co-localisations, l'omniprésence d'un moyen de transport en commun à proximité des Lycées de Chicago, ce qui peut supposer que la ville de Chicago est bien desservie. Tout comme les données de Paris, en se basant sur les résultats de prévalence, nous avons plus de motifs pertinents sous contrainte que sans la contrainte pour les mêmes raisons énoncées précédemment.

5 Conclusion et perspectives

Dans cet article, nous avons présenté CSS-Miner, une approche d'extraction de co-localisations sous contrainte de la structure spatiale de la zone d'étude. Nous avons décrit cette contrainte et comment nous l'avons prise en compte, en l'occurrence avec un réseau routier et une recherche du plus court chemin. Pour extraire nos motifs spatiaux, nous avons utilisé l'approche de l'extraction de cliques maximales avec une recherche de voisins restreinte par un cercle de rayon d modifiable par l'utilisateur, selon le cas d'étude.

FIGURE 5 – Données de Chicago avec les cliques maximales extraites.



Au final, grâce aux plateformes *OpenData* de Paris et Chicago, nous avons pu créer deux jeux de données réelles.

Cependant, durant l'étape de pré-traitement de données, nous avons choisi de réduire tous nos objets spatiaux en points. Nos futurs travaux seront garder le type initial de nos données (points, lignes, polygones, ...). De plus, à l'étape de la recherche du plus court chemin, un croisement entre la structure spatiale et nos objets spatiaux est effectué. Afin d'optimiser notre recherche du plus court chemin, une phase additionnelle d'élagage semble nécessaire. Plusieurs travaux ont été menés sur l'élagage de graphes, par exemple un élagage par filtre léger sur les réseaux de neurones convolutifs [11] ou encore un élagage de graphe appliqué sur une grille [10]. Donc une des prochaines étapes de nos travaux sera d'élaguer le graphe associé à notre structure spatiale afin d'accélérer notre étape de recherche du plus court chemin.

Par ailleurs, CSS-Miner reste une méthode d'analyse exploratoire, la prochaine étape de nos travaux sera donc d'intégrer la connaissance d'experts métier [8], tels que des urbanistes, des démographes et géographes, afin de vérifier la pertinence de nos co-localisations extraites.

Enfin, dans cet article, nous avons supposé que le chemin effectué entre deux objets spatiaux était à pied. En perspective, pour pouvoir considérer la distance en voiture (ou à vélo), nous prévoyons de considérer la structure spatiale par un graphe orienté étant donné que toutes les routes en voiture ne sont pas bidirectionnelles. Par la suite, pour que la distance en voiture soit pertinente, il serait donc nécessaire d'intégrer la dynamique temporelle avec les heures d'affluences impactant le trafic routier. Cependant, ces données intégrant la dimension temporelle ne sont pas forcément disponibles en libre accès. De ce fait, cette tâche nécessitera d'utiliser des *API* fournies par Google et d'autres entreprises de gestion du trafic routier.

Remerciements

Ces travaux ont été réalisés dans le cadre du projet SPIraL (ANR-19-CE35-0006-02) et financés par l'Agence Nationale de Recherche. Nous remercions Dr. Cyrille Goarant de l'Institut Pasteur de Nouvelle-Calédonie pour ses conseils durant la réalisation de cet article.

Références

[1] Mohammad Akbari, Farhad Samadzadegan, and Robert Weibel. A generic regional spatio-temporal co-occurrence pattern mining model : a case study for air pollution. *Journal of Geographical Systems*, 17(3) :249–274, 2015.

[2] Xuguang Bao and Lizhen Wang. A clique-based approach for co-location pattern mining. *Information Sciences*, 490 :244–264, 2019.

[3] Geoff Boeing. Osmnx : New methods for acquiring, constructing, analyzing, and visualizing complex

street networks. *Computers, Environment and Urban Systems*, 65 :126–139, 2017.

[4] Coen Bron and Joep Kerbosch. Algorithm 457 : finding all cliques of an undirected graph. *Communications of the ACM*, 16(9) :575–577, 1973.

[5] Frédéric Cazals and Chinmay Karande. A note on the problem of reporting maximal cliques. *Theoretical computer science*, 407(1-3) :564–568, 2008.

[6] Jeffrey Chiu, Amin Vahedian Khezerlou, and Xun Zhou. Understanding business location choice pattern : A co-location analysis on urban poi data. In *Proceedings of the 2nd INFORMS Workshop on Data Science, Phoenix, AZ, USA*, volume 3, 2018.

[7] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1) :269–271, 1959.

[8] Frédéric Flouvat, Jean-François N'guyen Van Soc, Elise Desmier, and Nazha Selmaoui-Folcher. Domain-driven co-location mining : Extraction, visualization and integration in a gis. *Geoinformatica*, 19 :147–183, 2015.

[9] M.L. Fredman and R.E. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. In *25th Annual Symposium onFoundations of Computer Science, 1984.*, pages 338–346, 1984.

[10] Daniel Harabor and Alban Grastien. Online graph pruning for pathfinding on grid maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1114–1119, 2011.

[11] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv :1808.06866*, 2018.

[12] Yan Huang, Shashi Shekhar, and Hui Xiong. Discovering colocation patterns from spatial data sets : a general approach. *IEEE Transactions on Knowledge and data engineering*, 16(12) :1472–1485, 2004.

[13] Donald B. Johnson. Efficient algorithms for shortest paths in sparse networks. *J. ACM*, 24(1) :1–13, jan 1977.

[14] Seung Kwan Kim, Jee Hyung Lee, Keun Ho Ryu, and Ungmo Kim. A framework of spatial co-location pattern mining for ubiquitous gis. *Multimedia tools and applications*, 71(1) :199–218, 2014.

[15] Krzysztof Koperski and Jiawei Han. Discovery of spatial association rules in geographic information databases. In *International Symposium on Spatial Databases*, pages 47–66. Springer, 1995.

[16] Seung Kwan Kim, Younghee Kim, and Ungmo Kim. Maximal cliques generating algorithm for spatial colocation pattern mining. In *Secure and Trust Computing, Data Management and Applications : 8th FIRA International Conference, STA 2011, Loutraki, Greece, June 28-30, 2011. Proceedings 8*, pages 241–250. Springer, 2011.

[17] J.W. Moon and L. Moser. On cliques in graphs. *Israel J. Math.*, 3 :23—-28, 1965.

[18] Nazha Selmaoui-Folcher, Frédéric Flouvat, Dominique Gay, and Isabelle Rouet. Spatial pattern mining for soil erosion characterization. In *New Technologies for Constructing Complex Agricultural and Environmental Systems*, pages 190–210. IGI Global, 2012.

[19] Shashi Shekhar and Yan Huang. Discovering spatial co-location patterns : A summary of results. In *International symposium on spatial and temporal databases*, pages 236–256. Springer, 2001.

[20] Mikkel Thorup. Undirected single-source shortest paths with positive integer weights in linear time. *J. ACM*, 46(3) :362–394, may 1999.

[21] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363 :28–42, 2006.

[22] Vanha Tran, Lizhen Wang, Hongmei Chen, and Qing Xiao. Mcht : A maximal clique and hash table-based maximal prevalent co-location pattern mining algorithm. *Expert Systems with Applications*, 175 :114830, 2021.

[23] Gintaras Vaira and Olga Kurasova. Parallel bidirectional dijkstra's shortest path algorithm. *Databases and Information Systems VI, Frontiers in Artificial Intelligence and Applications*, 224 :422–435, 2011.

[24] L.G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3) :410—-421, 1979.

[25] Lizhen Wang, Yuzhen Bao, and Zhongyu Lu. Efficient discovery of spatial co-location patterns using the icpi-tree. *The Open Information Systems Journal*, 3(1), 2009.

[26] Xiaojing Yao, Ling Peng, Liang Yang, and Tianhe Chi. A fast space-saving algorithm for maximal co-location pattern mining. *Expert Systems with Applications*, 63 :310–323, 2016.

[27] Jin Soung Yoo and Shashi Shekhar. A joinless approach for mining spatial colocation patterns. *IEEE Transactions on Knowledge and Data Engineering*, 18(10) :1323–1337, 2006.

[28] Wenhao Yu. Spatial co-location pattern mining for location-based services in road networks. *Expert Systems with Applications*, 46 :324–335, 2016.

[29] Wenhao Yu, Tinghua Ai, Yakun He, and Shiwei Shao. Spatial co-location pattern mining of facility points-of-interest improved by network neighborhood and distance decay effects. *International Journal of Geographical Information Science*, 31(2) :280–296, 2017.

Analyse d'une enquête sur la sémantique des motifs séquentiels avec négation

Thomas Guyet[1]

[1] Inria – Centre de Lyon, AIstroSight

thomas.guyet@inria.fr

Résumé

Un motif séquentiel avec négation prend la forme d'un motif séquentiel pour lequel le symbole de négation peut être utilisé devant certains des itemsets. Dans ce cas, l'itemset qui suit doit être absent d'une séquence pour que le motif apparaisse dans cette séquence. Des travaux récents ont montré que différentes sémantiques pouvaient être attribuées à ces formes de motif. Ces travaux ont ainsi mis en évidence que les algorithmes d'extraction de ces motifs n'extrayaient pas les mêmes ensembles de motifs et ils soulèvent la question de l'interprétabilité des résultats. Dans ce travail, nous nous sommes demandés si certaines sémantiques étaient plus intuitives que d'autres et si celles-ci correspondaient à celles d'un ou plusieurs algorithmes de l'état de l'art. Pour cela, nous avons procédé sous la forme d'un questionnaire. Cet article présente ce questionnaire et l'analyse des 124 réponses. Les résultats montrent que deux sémantiques sont majoritaires mais qu'aucune d'elles ne correspond à celles des algorithmes principaux de l'état de l'art. Des recommandations sont faites pour tenir compte de ce résultat.

Mots-clés

Extraction de motifs, motifs séquentiels, négation, interprétation, enquête.

Abstract

A sequential pattern with negation takes the form of a sequential pattern for which the negation symbol can be used before some of the itemsets. In this case, the following itemset must be absent in a sequence for the pattern to appear in this sequence. Recent work has shown that these patterns have different semantics and raises the question of the interpretability of pattern mining algorithms. This article presents a questionnaire about the intuitiveness of some semantics. The analysis of the 124 answers shows that there are mainly two semantics that are mostly intuitive but that none of them corresponds to those of the main algorithms of the state of the art. Recommendations are made to address this outcome.

Keywords

Pattern mining, sequential patterns, negation, survey.

1 Introduction

L'extraction de motifs séquentiels est une classe de méthodes classiques de la fouille de données. Elle vise à extraire des sous-séquences (motifs) qui apparaissent fréquemment dans une grande base de séquences. Le motif apparaît fréquemment s'il apparaît dans au moins σ séquences, où σ est défini par l'utilisateur. Par exemple, prenons le motif $\langle e\ (ca)\ d\rangle$ désignant que "l'item e est suivi de a et c en même temps (*itemset*) puis de d"[1]. Dans le tableau ci-dessous, ce motif apparaît dans 4 séquences ($\boldsymbol{p_0}$, $\boldsymbol{p_2}$, $\boldsymbol{p_3}$ et $\boldsymbol{p_4}$).

Table 1: Exemple de base de séquences. La case à cocher sur la droite permet au lecteur de répondre lui-même aux questions. Voir Question 1 pour ce tableau.

id	*Séquence*
$\boldsymbol{p_0}$	$\langle e\ (caf)\ d\ b\ e\ d\rangle$
$\boldsymbol{p_1}$	$\langle c\ a\ d\ b\ e\ d\rangle$
$\boldsymbol{p_2}$	$\langle e\ (ca)\ d\rangle$
$\boldsymbol{p_3}$	$\langle d\ e\ (ca)\ b\ d\ b\ e\ f\rangle$
$\boldsymbol{p_4}$	$\langle c\ e\ b\ (fac)\ d\ e\ c\rangle$

Ces motifs fréquents peuvent être énumérés efficacement, grâce à la propriété d'anti-monotonie du support (*i.e.* le nombre d'occurrences d'un motif). Intuitivement, le support d'un motif décroît avec la taille des motifs. Cette propriété, utilisée par la plupart des algorithmes de la littérature, évite d'énumérer les motifs qui sont plus grands que des motifs qu'on sait a priori ne pas être fréquents.

Plusieurs travaux [4, 7] ont enrichi le domaine des motifs séquentiels par l'ajout d'information sur l'absence de la survenue d'un évènement. On parle alors de motifs séquentiels *avec négation*. Les motifs séquentiels avec négation prennent la forme de motifs séquentiels pour lesquels un symbole de négation, $\neg$, devant un itemset indique que ce dernier doit être absent d'une séquence pour y apparaître. Intuitivement, le motif $\langle a\ \neg b\ c\rangle$ sera reconnu dans une séquence si cette dernière comporte un a puis un c et que b est absent entre les occurrences de a et c.

Néanmoins, il a été constaté que les deux algorithmes principaux eNSP [4] et NegPSpan [7] n'extraient pas les

[1] On suppose ici que (ca) et (ac) désignent le même ensemble d'items.

mêmes ensembles de motifs négatifs. Ceci s'explique par le fait que ces deux algorithmes n'attribuent pas la même sémantique au symbole de négation [2].[2] Pour un motif $\boldsymbol{p}$ et une séquence $\boldsymbol{s}$ donnés, eNSP et NEGPSPAN ne seront pas forcément d'accord sur le fait que $\boldsymbol{p}$ apparaisse ou non dans $\boldsymbol{s}$. Les comptages d'apparition dans la base sont donc différents pour ces algorithmes et les motifs qu'ils considèrent comme effectivement fréquents peuvent ainsi être différents.

Les deux sémantiques sont toutes aussi intéressantes l'une que l'autre. La question qui se pose réside alors sur le partage de la sémantique entre l'utilisateur et l'outil qu'il utilise. Autrement dit : l'utilisateur à qui sont délivrés les motifs a-t-il une interprétation similaire à celle de l'algorithme utilisé ? Si ce n'est pas le cas, il peut y avoir une mauvaise interprétation des résultats de l'extraction de motifs. Un utilisateur non-expert ne cherchant pas forcément à comprendre les subtilités de ces motifs, il semble utile d'identifier une possible disparité entre la sémantique utilisée dans un algorithme et celle utilisée "intuitivement" par un utilisateur. Si cette disparité existe, il sera alors nécessaire de proposer des solutions évitant une mauvaise interprétation des résultats.

Dans cet article, nous nous sommes donc principalement posé trois questions :

1. existe-t-il une sémantique "intuitive" pour les motifs avec négation ?
2. la sémantique "intuitive" correspond-elle à celle qui est effectivement utilisée par l'un des algorithmes de l'état de l'art ?
3. quelles recommandations faire sur l'usage des motifs avec négation ?

Pour répondre à ces questions, la méthodologie a consisté à proposer un questionnaire pour révéler la sémantique qui est intuitivement appliquée par les utilisateurs. Le détail de la méthodologie de cette enquête est décrit dans la Section 3. La Section 5 présente les questions qui ont été posées aux utilisateurs et explicite les interprétations alternatives qui sont possibles. La Section 6 présente et analyse les résultats qui ont été collectés auprès de 124 participants. Avant cela, on commence par un bref état de l'art des méthodes d'extraction de motifs séquentiels avec négation.

2 État de l'art sur l'extraction de motifs séquentiels avec négation

Les premiers travaux sur l'extraction de motifs négatifs ont été proposés par Savasere et *al.* [11] dans le cadre de la fouille d'itemsets. Les premiers travaux sur les motifs séquentiels avec négation ont été proposés par Wu et *al.* [14] pour des règles d'association. Plusieurs approches récentes ont été proposées pour bénéficier également des avancées dans le domaine de l'extraction de motifs. L'algorithme eNSP extrait des motifs négatifs en exploitant des opérations ensemblistes entre motifs séquentiels fréquents [4]. Il évite ainsi l'énumération directe des motifs avec négation, car l'ensemble des motifs qui sont extraits ne bénéficient pas de la propriété d'antimonotonie. De nombreuses variantes de cet algorithme ont été proposées depuis, s'intéressant à l'utilité des items [15], aux répétitions [5], aux contraintes multiples de supports [16], etc. NEGPSPAN [7] est une approche concurrente à eNSP qui utilise une sémantique de motifs différente. Cette sémantique bénéficie de la propriété d'antimonotonie. Ceci permet une extraction efficace et complète selon les principes classiques de l'extraction de motifs. Récemment, Wang et *al.* [13] ont proposé VM-NSP, un algorithme qui utilise une représentation verticale pour améliorer l'efficacité des algorithmes. Le lecteur intéressé par un état de l'art plus complet des approches récentes d'extraction de motifs séquentiels avec négation peut se référer à Wang et *al.* [12].

Les premières approches se sont comparées entre elles bien qu'elles n'utilisent pas les mêmes sémantiques de motifs. L'identification des différentes sémantiques a conduit à clarifier le domaine [2]. Plus précisément, huit sémantiques des motifs avec négation ont été identifiées. Ces huit sémantiques sont issues de choix possibles d'interprétation de la notion de non-inclusion, d'occurrence et de relation d'inclusion. La section 5 détaille ces notions.

3 Enquête sur la perception des motifs avec négation

L'enquête[3] mise en place vise à identifier une sémantique qui serait plus naturellement utilisée par les utilisateurs d'algorithmes d'extraction de motifs. Cette enquête est organisée en trois parties (la section suivante revient plus en détail sur les questions des phases 2 et 3 de l'enquête) :

1. estimation du niveau de connaissance du domaine de la fouille de motifs et de la logique. Dans cette partie, on demande si l'utilisateur est familier des notions d'extraction de motifs, et également s'il est informaticien/logicien/chercheur. L'objectif de cette question est de disposer d'informations pour caractériser d'éventuels biais de l'ensemble des enquêtés.
2. vérification de la compréhension des principes des motifs séquentiels afin de limiter les biais de compréhension dans la suite des questions. Tout d'abord, un texte explique et illustre les principes des motifs séquentiels. Une première question évalue la compréhension de la sémantique des motifs séquentiels (sans négation), notamment les notions d'*itemset*, le séquencement et la possibilité de *gaps*[4].

[2] Ce n'est pas la seule raison de la divergence entre les algorithmes. Mais les autres différences sont mineures.

[3] Enquête : `http://people.irisa.fr/Thomas.Guyet/negativepatterns/Survey/survey.php`

[4] La reconnaissance de la sous-séquence permet l'insertion d'itemsets au milieu d'une occurrence.

Tant que la réponse à cette question n'est pas correcte, l'utilisateur ne peut pas poursuivre le questionnaire. Une seconde question vérifie que la portée des négations est comprise telle que définie par notre cadre d'analyse [2]. Par exemple, pour le motif $\langle a\ \neg b\ c\rangle$, la négation du b ne porte pas au-delà d'une occurrence de c. Ainsi, ce motif est considéré comme apparaissant dans la séquence $\langle a\ e\ c\ b\rangle$ même si un b apparaît après le c. Les utilisateurs ne répondant pas correctement à cette question seront écartés de l'analyse des réponses.

3. identification de la sémantique « intuitive » des motifs séquentiels avec négation. Pour chacune de ces questions, on demande à l'utilisateur de cocher les séquences dans lesquelles il pense qu'un motif apparaît (voir exemple de la Figure 1). Le groupe de séquences cochées associe donc un utilisateur à une sémantique donnée.

L'enquête a été diffusée au travers de listes de diffusion de recherche ainsi que dans des cercles non liés à la recherche pour collecter des réponses de non-experts. Elle est accessible via un navigateur web standard. Le questionnaire est rédigé en anglais et s'adresse donc à des anglophones. Les explications relatives aux principes de la notion de motif séquentiel sont détaillées en début de questionnaire. Afin d'éviter le biais de maîtrise des représentations mathématiques, le questionnaire peut être joué en deux versions : avec notations sous forme de lettres ou de symboles colorés (cf. Figure 1).
Le questionnaire est totalement anonyme. Seule la date de saisie du questionnaire a été collectée en sus des réponses aux questions.

4 Cadre général

On commence par introduire la syntaxe des motifs séquentiels avec négation. Dans toute la suite, $[n] = \{1, \ldots, n\}$ désigne l'ensemble des n premiers entiers, et $\mathcal{I}$ désigne un ensemble d'items (alphabet). Un sous-ensemble $A = \{a_1\ a_2\ \cdots\ a_m\} \subseteq \mathcal{I}$ est nommé un *itemset*. Une *séquence* $\boldsymbol{s}$ est de la forme $\boldsymbol{s} = \langle s_1\ s_2\ \cdots\ s_n\rangle$ où s_i est un itemset.

Définition 1 (Motif séquentiel avec négation)**.** *Un motif séquentiel avec négation* $\boldsymbol{p} = \langle p_1\ \neg q_1\ p_2\ \neg q_2\ \cdots\ p_{n-1}\ \neg q_{n-1}\ p_n\rangle$ *est telle que* $p_i \in 2^{\mathcal{I}}\backslash\{\emptyset\}$ *pour tout* $i \in [n]$ *et* $q_i \in 2^{\mathcal{I}}$ *pour tout* $i \in [n-1]$. $\boldsymbol{p}^+ = \langle p_1\ p_2\ \cdots p_n\rangle$ *désigne la partie positive de* $\boldsymbol{p}$.

La sémantique des motifs repose sur la relation d'inclusion. Cette relation précise comment considérer si un motif apparaît (est inclus) ou non dans une séquence. Cette relation utilise la notion d'occurrence d'un motif dans une séquence, formellement définie ainsi :

Définition 2 (Occurrence d'un motif séquentiel)**.** *Soit une séquence* $\boldsymbol{s} = \langle s_1\ \ldots\ s_n\rangle$ *et* $\boldsymbol{p} = \langle p_1\ \ldots\ p_m\rangle$ *un motif séquentiel.* $\boldsymbol{e} = (e_i)_{i\in[m]} \in [n]^m$ *est une* occurrence *du motif* $\boldsymbol{p}$ *dans la séquence* $\boldsymbol{s}$ *ssi* $\forall i \in [m],\ p_i \subseteq s_{e_i}$ *et* $e_i < e_{i+1}$ *pour tout* $i \in [m-1]$.

Partant de cette définition, nous avons construit la question suivante pour en valider sa compréhension avant de poursuivre la suite du questionnaire.

Question 1 (Occurrence d'un motif séquentiel)**.** *Soit le motif séquentiel* $\boldsymbol{p} = \langle(ca)\ d\ e\rangle$*, indiquer dans quelles séquences de la Table 1 apparaît le motif* $\boldsymbol{p}$.[5]

Les réponses attendues à cette question sont les séquences $\boldsymbol{p_0}$, $\boldsymbol{p_3}$ et éventuellement $\boldsymbol{p_4}$. La séquence $\boldsymbol{p_0}$ permet de vérifier la compréhension que (ca) apparaît dans (caf) selon nos définitions. La séquence $\boldsymbol{p_1}$ permet de vérifier qu'il faut que tous les éléments de (ca) apparaissent ensemble. La séquence $\boldsymbol{p_2}$ permet vérifier la compréhension de l'importance de l'ordre dans la séquence. La séquence $\boldsymbol{p_3}$ permet vérifier la compréhension de la notion de *gap* : il est possible d'avoir des itemsets au milieu d'une occurrence (par exemple, la survenue de b entre le d et le e). Finalement, la dernière séquence présente un itemset dont les items ne sont pas ordonnés. Dans le cas où $\boldsymbol{p_4}$ ne serait pas jugé contenir $\boldsymbol{p}$ alors on serait informé d'une sensibilité de l'utilisateur à l'ordre présenté dans un itemset (ce qui n'est classiquement pas le cas).
De la même manière, la sémantique des motifs séquentiels avec négation repose sur une relation d'inclusion. Un motif avec négation, $\boldsymbol{p}$, est inclus dans une séquence $\boldsymbol{s}$ si $\boldsymbol{s}$ contient une sous-séquence $\boldsymbol{s'}$ telle que chaque ensemble positif de $\boldsymbol{p}$, *i.e.* p_i, est inclus dans un itemset de $\boldsymbol{s'}$ (en respectant l'ordre) et que toutes les contraintes de négations exprimées par les $\neg q_i$ sont également satisfaites. La contrainte de négation de q_i s'appliquant alors à la sous-séquence de $\boldsymbol{s'}$ située entre l'occurrence de l'itemset positif précédant $\neg q_i$ dans $\boldsymbol{p}$ et l'occurrence de l'itemset positif suivant $\neg q_i$ dans $\boldsymbol{p}$.
Cette définition détermine la portée de la négation. Cette définition est propre au cadre dans lequel nous travaillons par la suite. Aussi, il est important de vérifier qu'il est partagé par les utilisateurs. La question suivante permet de s'en assurer.

Question 2 (Portée de la négation)**.** *On considère un motif* $\boldsymbol{p} = \langle c\ \neg d\ e\rangle$*. Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous,* $\boldsymbol{p}$ *apparaît.*

id	*Séquence*
$\boldsymbol{s_0}$	$\langle f\ f\ c\ b\ d\ a\ e\rangle$
$\boldsymbol{s_1}$	$\langle f\ c\ b\ f\ a\ e\rangle$
$\boldsymbol{s_2}$	$\langle b\ f\ c\ b\ a\rangle$
$\boldsymbol{s_3}$	$\langle b\ c\ b\ e\ d\rangle$
$\boldsymbol{s_4}$	$\langle f\ a\ c\ e\ b\rangle$

Dans cette question, il est raisonnable de considérer que $\boldsymbol{p}$ apparaît dans $\boldsymbol{s_1}$, $\boldsymbol{s_3}$ (le d est hors de la portée supposée de la négation) et $\boldsymbol{s_4}$. Les enquêté(e)s qui ne cochent pas $\boldsymbol{s_4}$ ont probablement interprété la contrainte $\neg d$ comme : l'apparition d'un élément qui n'est pas d (ce qui n'est pas dans les définitions proposées par la suite).

[5] Le lecteur est invité à remplir lui-même les réponses aux questions dans les tableaux avant de lire les explications.

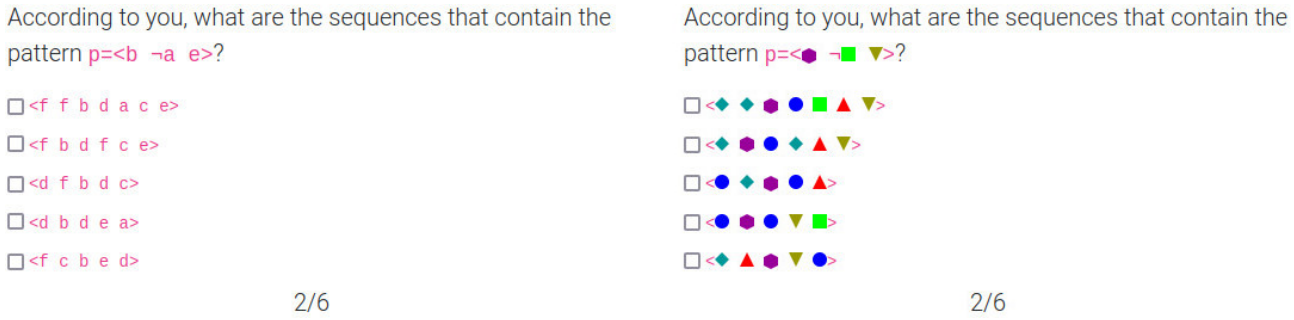


Figure 1: Illustration des deux versions du questionnaire : sur la gauche, version avec lettres, sur la droite, version avec symboles colorés. La question consiste à désigner les séquences pour lesquels l'utilisateur pense qu'elle contient un motif.

Si $\boldsymbol{p_0}$ est considéré comme contenant $\boldsymbol{p}$ il est probable que la contrainte $\neg d$ soit comprise comme devant suivre strictement après c (de nouveau, ce n'est pas une situation considérée dans l'analyse de Besnard et Guyet [2]).

5 Questions sur la sémantique des négations

Dans cette section, nous reprenons les questions de la troisième partie du questionnaire et nous expliquons les différentes interprétations que révèlent les réponses faites par les enquêtés.

5.1 Non-inclusion d'un itemset

Question 3. *Soit le motif séquentiel* $\boldsymbol{p} = \langle d\ \neg(af)\ b\rangle$. *Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous,* $\boldsymbol{p}$ *apparaît.*

id	*Séquence*
$\boldsymbol{i_0}$	$\langle e\ e\ d\ a\ b\ e\rangle$
$\boldsymbol{i_1}$	$\langle d\ (af) b\ c\rangle$
$\boldsymbol{i_2}$	$\langle e\ d\ (fc)\ b\rangle$
$\boldsymbol{i_3}$	$\langle e\ c\ d\ (ec)\ b\rangle$
$\boldsymbol{i_4}$	$\langle d\ (fa)\ b\ e\rangle$

Cette question est construite de telle sorte que chacune des séquences contienne la partie positive du motif[6] avec seulement un itemset entre les occurrences de d et de b. Ces séquences posent donc la question de la non-inclusion de l'itemset (af) dans a, (af), (fc), (ec) ou (fa). Dans le cas où l'enquêté(e) coche les séquences $\boldsymbol{i_0}$, $\boldsymbol{i_2}$ et $\boldsymbol{i_3}$, on peut en déduire qu'il/elle considère que la présence d'au moins un élément de l'itemset (af) "active" la négation. On parle alors d'une *non-inclusion partielle*. En revanche, si seule la séquence $\boldsymbol{i_3}$ est cochée, alors on peut en déduire qu'il/elle considère qu'il faut la présence de tous les items de l'itemset pour "activer" la négation. On nomme cela une *non-inclusion totale*. En complément, la séquence $\boldsymbol{i_4}$ visait à voir si l'ordre dans l'itemset pouvait importer aux enquêté(e)s et si cela était cohérent avec leur réponse à la séquence $\boldsymbol{p_4}$ de la question 1.

Plus formellement, cette question discrimine entre deux choix d'inclusion entre deux itemsets $P \in 2^{\mathcal{I}} \setminus \{\emptyset\}$ et $I \in 2^{\mathcal{I}}$:

- non-inclusion partielle : $P \not\subseteq_G I \Leftrightarrow \exists e \in P, e \notin I$
- non-inclusion totale : $P \not\subseteq_D I \Leftrightarrow \forall e \in P, e \notin I$

La non-inclusion partielle signifie que $P \setminus I$ est non-vide tant que la non-inclusion totale signifie que P et I sont disjoints (noté D). Dans la suite, le symbole $\not\subseteq_*$ dénote une relation de non-inclusion entre itemsets, indifféremment $\not\subseteq_G$ ou $\not\subseteq_D$.

5.2 Occurrence d'un motif avec négation

Question 4 (Occurrence d'un motif avec négation)**.** *Soit le motif séquentiel* $\boldsymbol{p} = \langle f\ \neg(ea)\ d\rangle$. *Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous,* $\boldsymbol{p}$ *apparaît.*

id	*Séquence*
$\boldsymbol{e_0}$	$\langle b\ b\ f\ c\ e\ d\ b\rangle$
$\boldsymbol{e_1}$	$\langle b\ f\ e\ a\ c\ b\ d\rangle$
$\boldsymbol{e_2}$	$\langle f\ c\ (ea)\ b\ c\ d\ c\rangle$
$\boldsymbol{e_3}$	$\langle b\ c\ f\ b\ c\ c\ d\rangle$

Dans cette question, la forme du motif séquentiel $\boldsymbol{p} = \langle f\ \neg(ea)\ d\rangle$ est la même à une permutation près des lettres à la précédente et chaque séquence de la base de séquences contient la partie positive de $\boldsymbol{p}$. Mais cette fois, il y a plusieurs itemsets entre les occurrences de f et de d. L'objet de cette question est donc de voir comment cet ensemble d'itemsets positionnés dans la portée de la négation est considéré par un(e) enquêté(e). On estime tout d'abord que, quel que soit l'enquêté(e), $\boldsymbol{p}$ apparaît dans $\boldsymbol{e_3}$ (il n'y a clairement ni e ni a ici) mais $\boldsymbol{p}$ n'apparaît pas dans $\boldsymbol{e_2}$ (on retrouve l'itemset (ea) dans la portée de la négation). La séquence la plus révélatrice est en fait $\boldsymbol{e_1}$. La spécificité de cette séquence est de contenir les deux éléments de l'itemset nié (e et a), mais dans deux itemsets différents. L'enquêté(e) qui ne la coche pas (*i.e.* qu'il/elle considère que $\boldsymbol{p}$ n'apparaît pas dans $\boldsymbol{e_1}$) utilise la notion d'*occurrence souple* : il faudrait que e et a apparaissent ensemble pour "activer" la négation (comme le cas de $\boldsymbol{e_2}$). L'enquêté(e) qui la coche estime que c'est globalement sur toute la période que s'applique la contrainte de négation, on parle d'*occurrence stricte*.

[6] $\langle d\ b\rangle$ est la partie positive de $\langle d\ \neg(af)\ b\rangle$.

Quant à la séquence $\boldsymbol{e_0}$, elle révèle la notion de non-inclusion vu précédemment : en cas de non-inclusion partielle, $\boldsymbol{p}$ apparaît dans $\boldsymbol{e_0}$, mais pas si on considère une non-inclusion totale.

Deux sémantiques ont été distinguées : les occurrences strictes et les occurrences souples. Elles peuvent être formellement définies comme suit : Soit une séquence $\boldsymbol{s} = \langle s_1 \dots s_n \rangle$ et un motif avec négation $\boldsymbol{p} = \langle p_1 \neg q_1 \dots \neg q_{m-1} p_m \rangle$. On dit que $\boldsymbol{e} = (e_i)_{i\in[m]} \in [n]^m$ est une occurrence souple de $\boldsymbol{p}$ dans la séquence $\boldsymbol{s}$ ssi :

- $p_i \subseteq s_{e_i}$ pour tout $i \in [m]$
- $q_i \not\subseteq_* s_j,\ \forall j \in [e_i+1, e_{i+1}-1]$ pour tout $i \in [m-1]$

On dit que $\boldsymbol{e} = (e_i)_{i\in[m]} \in [n]^m$ est une occurrence stricte de $\boldsymbol{p}$ dans la séquence $\boldsymbol{s}$ ssi :

- $p_i \subseteq s_{e_i}$ pour tout $i \in [m]$
- $q_i \not\subseteq_* \bigcup_{j\in[e_i+1,e_{i+1}-1]} s_j$ pour tout $i \in [m-1]$

Intuitivement, la contrainte souple considère la non-inclusion de q_i pour chacun des itemsets situés dans l'intervalle de position $[e_i + 1, e_{i+1} - 1]$ tandis que la contrainte stricte considère la non-inclusion sur l'union de l'ensemble des itemsets à ces mêmes positions. L'intervalle correspond aux itemsets de la séquence strictement entre les occurrences des itemsets entourant q_i.

5.3 Occurrences multiples dans une séquence

Question 5 (Occurrences multiples d'un motif avec négation)**.** *Soit le motif séquentiel $\boldsymbol{p} = \langle b \neg e\ f \rangle$. Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous, $\boldsymbol{p}$ apparaît.*

id	*Séquence*
$\boldsymbol{o_0}$	$\langle b\ a\ f\ d\ b\ d\ f \rangle$
$\boldsymbol{o_1}$	$\langle b\ a\ f\ d\ e\ b\ d\ f \rangle$
$\boldsymbol{o_2}$	$\langle d\ b\ e\ c\ a\ d\ f\ b\ d\ e\ f \rangle$
$\boldsymbol{o_3}$	$\langle b\ a\ f\ b\ a\ e\ f \rangle$

Dans cette question, les séquences ci-dessous contiennent chacunes plusieurs occurrences de la partie positive du motif. Pour rendre plus visible cette situation, il y a même plusieurs occurrences non imbriquées de $\langle b\ f \rangle$. Dans la mesure où la contrainte de négation porte uniquement sur un seul item (e), les choix relatifs aux dimensions précédentes – non-inclusion d'un itemset et type d'occurrence – n'ont a priori pas d'impact. Ceci permet donc de focaliser la question sur la perception de ces occurrences multiples. Deux comportements sont alors attendus :

- La première interprétation consiste à considérer que dès qu'il existe une occurrence de la partie positive, $\langle b\ f \rangle$, qui satisfait la contrainte de négation, alors la séquence est reconnue. On parle alors d'*occurrence faible*. Cette interprétation est révélée par la sélection des séquences $\boldsymbol{o_0}$, $\boldsymbol{o_1}$ et $\boldsymbol{o_3}$.
- Le second comportement consiste à considérer que dès qu'une occurrence de la partie positive ne satisfait pas la contrainte de négation, alors la séquence n'est pas reconnue. On parle alors de *non-occurrence forte*. Pour la question 5, cela correspond aux enquêté(e)s qui ont coché uniquement la séquence $\boldsymbol{o_0}$, toutes les autres ayant au moins une occurrence de $\langle b\ f \rangle$ avec un e interstitiel. On peut néanmoins constater que la séquence $\boldsymbol{o_1}$ est piégeuse pour ceux qui ont cette intuition, puisqu'il y a deux occurrences minimales de $\langle b\ f \rangle$ (au sens de Mannila et *al.* [9]) qui satisfont la contrainte de négation, mais il y a aussi une occurrence impliquant le premier b et le dernier f. Cette dernière occurrence de la partie positive ne satisfait pas la contrainte de négation. Pour des novices dans l'utilisation des séquences, cette subtilité peut être difficile à détecter. Il semble donc plus judicieux de ne juger de l'interprétation que sur l'absence de $\boldsymbol{o_3}$.

Cette question nous amène de nouveau à deux alternatives. Soit une séquence $\boldsymbol{s}$ et un motif $\boldsymbol{p}$. Pour $\not\subseteq_* \in \{\not\subseteq_D, \not\subseteq_G\}$ et $\bullet \in \{\circ, \bullet\}$,

- $\boldsymbol{p} \preceq^*_\bullet \boldsymbol{s}$ signifie que le motif $\boldsymbol{p}$ est inclus dans la séquence $\boldsymbol{s}$ ssi il existe au moins une occurrence (souple ou stricte) de $\boldsymbol{p}$ dans $\boldsymbol{s}$ avec la non-inclusion $\not\subseteq_*$.
- $\boldsymbol{p} \sqsubseteq^*_\bullet \boldsymbol{s}$ signifie que le motif $\boldsymbol{p}$ est inclus dans la séquence $\boldsymbol{s}$ ssi il existe au moins une occurrence de $\boldsymbol{e}$ dans $\boldsymbol{p}^+$ et que pour chaque occurrence $\boldsymbol{e}$ de $\boldsymbol{p}^+$ dans $\boldsymbol{s}$, $\boldsymbol{e}$ est également une occurrence (souple ou stricte) de $\boldsymbol{p}$ dans $\boldsymbol{s}$ avec la non-inclusion $\not\subseteq_*$.

Les trois dimensions interprétatives de la négation se combinent donc en huit sémantiques possibles définies par leurs relations d'inclusion : $\preceq^D_\circ, \preceq^D_\bullet, \preceq^G_\circ, \preceq^G_\bullet, \sqsubseteq^D_\circ, \sqsubseteq^D_\bullet, \sqsubseteq^G_\circ, \sqsubseteq^G_\bullet$ étudiées dans [2]. Comme illustré, les trois questions ci-dessus ont été construites pour explorer indépendamment chacune des trois dimensions de la sémantique de la négation dans un motif séquentiel. En particulier, nous avons illustré comment la construction des questions permet d'associer, en fonction de la réponse donnée, un(e) enquêté(e) à une sémantique.

6 Analyse et résultats de l'enquête

À l'issue de la période d'enquête, nous avons collecté 124 questionnaires complets. L'expertise auto-estimée dans le domaine de l'extraction de motifs se répartit en 40 novices, 54 ayant des connaissances en science des données et 27 se déclarants familiers avec l'extraction de motifs. 79 se déclarent comme informaticiens, 82 comme chercheurs et 23 comme logiciens. Le nombre de tentatives pour la compréhension de la notion d'occurrence d'un motif est en moyenne de 1.27 ± 0.49 (entre 1 et 5 tentatives). 102 ont correctement répondu dès la première tentative. On peut noter que 6 enquêté(e)s ayant des connaissances en analyse de données (sur 24) ont eu besoin de plus d'une tentative pour avoir la réponse correcte.

Le résultat de l'enquête comporte les réponses booléennes (séquence cochée ou non cochée) pour chaque séquence des questions. Dans l'objectif d'identifier les sémantiques les plus naturelles chez les enquêté(e)s, on peut voir ce problème comme un problème d'extraction d'itemsets fréquents ou de co-clustering. On cherche à identifier des groupes d'individus qui ont coché les mêmes réponses.
Pour l'analyse des réponses, nous procédons en deux temps :

1. on commence par analyser les résultats question par question, *i.e.* indépendamment pour chacune des dimensions de la sémantique des motifs ($\not\subseteq_D$ ou $\not\subseteq_G$, $\circ$ ou $\bullet$, $\preceq$ or $\sqsubseteq$).

2. on complète l'analyse par une analyse globale des sémantiques.

Dans la section précédente, nous avons identifié pour chaque question les grandes classes de réponse attendue. On donne donc par la suite les statistiques d'apparition pour chacune, mais comme les réponses ne correspondent pas forcément exactement à ce qui est attendu (soit par inattention de l'enquêté(e), soit par une interprétation différente), nous proposons d'utiliser l'analyse de concepts formels (*Formal Concept Analysis* ou FCA) [6] pour donner une vision globale des résultats. La FCA est une technique d'analyse de données qui identifie des concepts d'un jeu de données. Chaque concept est décrit, d'une part, par son intention qui est ici un ensemble de réponses cochées et, d'autre part, son extension qui liste tous les individus qui ont choisi ces réponses. Les concepts extraits sont *fermés*, c'est-à-dire que leur extension est maximale pour leur intention et réciproquement. Un des intérêts de la FCA est de représenter de manière synthétique les données dans le treillis de concepts. Au travers de ce treillis, il est possible d'analyser précisément des groupes d'individus ayant fait les mêmes réponses. On peut noter que la FCA a déjà été utilisée pour l'analyse de questionnaires [1]. L'outil utilisé pour construite les treillis est GALACTIC [3].

6.1 Analyse de chaque dimension de la sémantique

Dans cette partie, on analyse les réponses à quatre questions : on s'intéresse tout d'abord aux réponses à la question sur la portée des négations, ensuite, on analyse les trois dimensions de la sémantique des motifs avec négation : la non-inclusion des itemsets, les occurrences et les relations d'inclusion. Les Tableaux 2 à 5 donnent de manière synthétique les nombres de chacune des interprétations. Les Figures 2 à 4 illustrent les treillis de concepts obtenus pour chacune de ces questions pour donner une image plus globale des réponses.
Concernant la compréhension de la portée des négations, 101 personnes ont coché des réponses correspondant à l'attendu pour cette question de vérification (cf. Table 2). Il est intéressant de constater que 9 personnes qui avaient coché $\boldsymbol{s_1}$ et $\boldsymbol{s_3}$ n'ont pas coché $\boldsymbol{s_4}$ laissant penser que, pour elles, la négation d'un itemset signifie qu'il s'agit d'un

Table 2: Résultat sur la question de la portée de la négation.

Portée	Nombre	Pourcentage
Conforme	101	81.4%
Conforme sauf $\boldsymbol{s_4}$	9	7.3%
Alternatif	14	11.3%

Table 3: Réponses à la question des non-inclusions (en nombre et en pourcentage).

Interprétation	Nombre	Pourcentage
Non-inclusion partielle	100	90.9%
Non-inclusion totale	3	2.7%
Autre	7	6.4%

évènement qui n'est pas l'évènement nié. Autrement dit, il faut au moins un évènement qui ne soit pas l'évènement nié pour activer la négation.[7] Pour les autres différences marginales (14 personnes), nous considérons qu'il s'agit d'oublis ou d'erreurs. Les réponses de ces personnes ont été écartées de la suite de l'analyse des résultats, leur compréhension possiblement différente de la portée de la négation ne permet d'exploiter leurs réponses. La suite des analyses porte donc sur 110 questionnaires complétés.
Concernant les non-inclusions d'itemsets (Table 3 et Figure 2), on constate que les enquêté(e)s ont très majoritairement (100) sélectionné le triplet de réponse $\boldsymbol{i_0}$, $\boldsymbol{i_2}$ et $\boldsymbol{i_3}$ correspondant à l'interprétation de non-inclusion partielle (concept `$8` dans la Figure 2). Seulement 3 personnes ont considéré l'interprétation de la non-inclusion totale. De manière plutôt inattendue, 22 enquêté(e)s ont considéré que la séquence $\boldsymbol{i_4}$ contenait le motif et donc que (fa) n'était pas incompatible avec (af). Ces enquêté(e)s se répartissent dans les différents niveaux d'expertises (8, 11 et 3 respectivement pour les niveaux 0, 1 et 2). Il ne s'agit donc pas plus particulièrement des personnes qui ne sont pas biaisées par l'habitude des notations de la fouille de motifs.

Table 4: Réponses à la question des occurrences.

Interprétation	Nombre	Pourcentage
Occurrence stricte	97	88.2%
Occurrence souple	7	6.3%
Autre	6	5.5%

Concernant l'analyse des occurrences (Table 4 et Figure 3), seule la séquence $\boldsymbol{e_1}$ permet de discriminer l'intuition des enquêté(e)s. Pour la Table 4, on s'assure aussi que les réponses sont correctes pour $\boldsymbol{e_2}$ et $\boldsymbol{e_3}$, sinon on place la réponse en "autre". De nouveau, on obtient un résultat très marqué pour l'interprétation dite souple : 97 personnes y adhèrent (concept `$7` dans la Figure 3). Le concept `$3` correspond aux individus qui n'ont pas coché $\boldsymbol{e_1}$. Il s'agit donc des interprétations d'occurrence stricte.

[7] NB : dans les questions suivantes, toutes les séquences ont au moins un évènement "neutre" là où un itemset avec négation est attendu. On peut donc conserver des personnes sans biaiser les réponses suivantes.

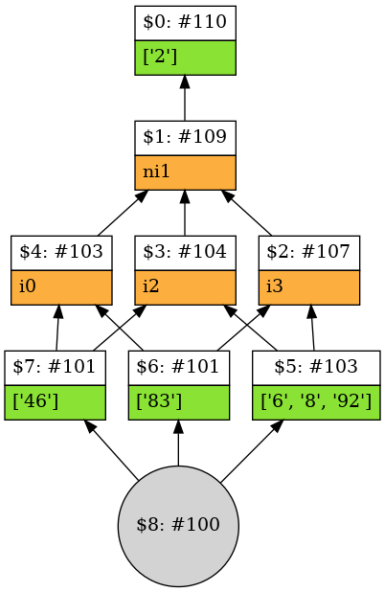


Figure 2: Concepts extraits à partir des réponses à la Question 3 : non-inclusion d'un itemset. Chaque concept est illustré par une boîte contenant différents éléments : les générateurs sur fond orange (réponses possibles aux questions), les prototypes sur fond vert. Le symbole `$` désigne un numéro de concept (un identifiant). La taille de l'extension est précisée avec un `#`. La liste entre crochets pour un prototype désigne une liste d'exemples (ici des numéros de questionnaires) "complémentaire" par rapport au concept inférieur. Par exemple, le concept `$7` couvre 101 exemples: les 100 du concept `$8` plus l'exemple `46`. Chaque concept indique l'intention comme un ensemble de séquences cochées (se reporter aux tables présentées dans les exemples). Dans les réponses aux questions, `i0` désigne que d'enquêté(e) a coché la séquence $\boldsymbol{i_0}$, et `ni1` (préfixe avec `n`) désigne que d'enquêté(e) **n'a pas** coché la séquence $\boldsymbol{i_1}$. Le choix entre les deux représentations d'une réponse a été fait en considérant la lisibilité du treillis obtenu.

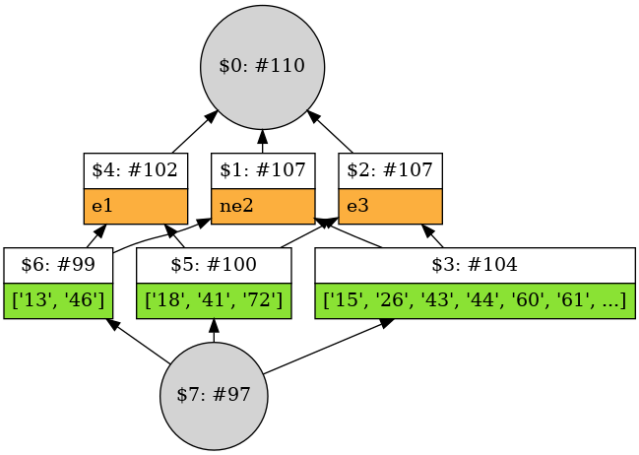


Figure 3: Concepts extraits à partir des réponses à la Question 4 relative aux occurrences (cf Figure 2 pour légende).

Table 5: Réponses à la question des relations d'inclusion.

Interprétation	Nombre	Pourcentage
Relation faible	75	69.2%
Relation forte	31	28.2%
Autre	4	3.6%

Finalement, concernant l'analyse des relations d'inclusion (Table 5 et Figure 4), le résultat est ici plus partagé. 75 personnes ont exclusivement identifié les trois séquences correspondant à la notion de relation faible (relation $\preceq$). Ils sont représentés dans le concept `$3` de la Figure 4. En revanche, 31 personnes ont exclusivement sélectionné la séquence $\boldsymbol{o_0}$ (concept `$1`). Ces derniers ont ainsi préféré l'interprétation de la relation forte (relation $\sqsubseteq$). Ces 31 personnes comprennent 14 qui n'ont pas coché la séquence $\boldsymbol{o_1}$ et 17 qui l'ont cochée. Ces derniers adhèrent plus à la notion d'occurrence minimale de Mannila et *al.* [10].

6.2 Fréquence des sémantiques

Les questions 3, 4 et 5 attribuent chaque enquêté(e) à une interprétation d'une des trois dimensions qui constituent la sémantique d'un motif avec négation. On cherche maintenant à voir s'il existe des sémantiques (comme combinaison des choix d'interprétation pour les trois dimensions) qui sont dominantes parmi les 8 possibles.

La Figure 5 synthétise les réponses à notre enquête. Elle illustre le treillis de concepts représentant les sémantiques des motifs avec négation. Les cinq prototypes du niveau inférieur décrivent les 5 sémantiques (et leur représentation dans les données) qui ont été effectivement utilisées par les enquêté(e)s. Ils sont définis par un choix d'interprétation pour chacune des trois dimensions. De gauche à droite, on identifie les sémantiques $\preceq^D_\circ$, $\preceq^G_\bullet$, $\sqsubseteq^G_\bullet$, $\sqsubseteq^G_\circ$ et $\sqsubseteq^D_\circ$. Sachant que parmi les huit sémantiques, il y a en fait deux paires équivalentes [2] ($\sqsubseteq^D_\circ \sim \sqsubseteq^D_\bullet$ et $\preceq^D_\circ \sim \preceq^D_\bullet$), la seule sémantique qui n'est pas représentée dans cette enquête est $\preceq^G_\circ$.

Sur les 110 enquêté(e)s, le questionnaire a permis d'attribuer une sémantique intuitive à 96 personnes. Les 14 autres personnes ont au moins une question pour laquelle il n'a pas été clairement identifiée une des interprétations attendues. Ces individus se trouvent dans les concepts intermédiaires (prototypes `5`, `6` et `10`, générateur `15` et concepts `3`, `9` et `13`).

Une première constatation est qu'il est possible d'attribuer une sémantique à une grande partie des enquêté(e)s. C'est-à-dire que ce sont probablement les mêmes personnes qui ont faits des réponses "alternatives" aux différentes questions. Ce résultat nous conforte sur l'exploitabilité des résultats collectés.

Ensuite, cette figure met en évidence le résultat principal de cette étude : il existe principalement deux sémantiques qui sont intuitivement utilisées : $\sqsubseteq^G_\circ$ à 23.9% et $\preceq^G_\circ$ à 69.8%. Les autres sémantiques sont marginalement représentées.

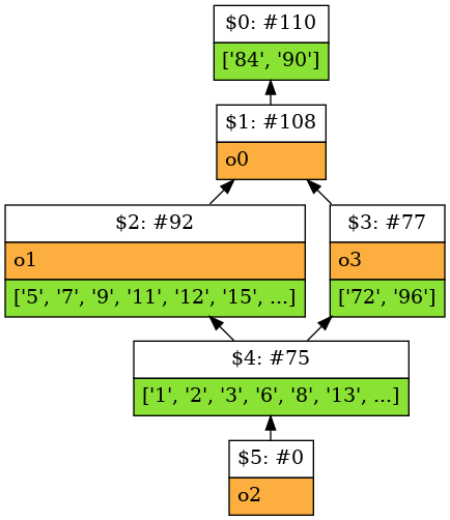


Figure 4: Concepts extraits à partir des réponses à la Question 5 relative aux relations d'inclusion (cf. Figure 2 pour légende).

Nous nous sommes ensuite intéressés à la comparaison des populations définies par le choix des sémantiques en nous intéressant à leurs réponses aux questions sur leur profil. Nous avons pour cela procédé à un test statistique pour comparer les distributions des niveaux d'expertises (test de Student). Les résultats ne montrent aucune différence significative entre les groupes. La conclusion que nous tirons est que l'intuition d'une sémantique n'est globalement pas liée à une expertise particulière en informatique ou en science des données.

7 Sémantiques préférées et algorithmes

On peut conclure de ces analyses qu'il n'y a pas une seule sémantique partagée pour les enquêté(e)s, mais plutôt que deux dominent : $\sqsubseteq_{\mathrm{O}}^{G}$ et $\preceq_{\mathrm{O}}^{G}$. Il est intéressant de comparer ce résultat avec les choix des deux algorithmes majeurs du domaine, eNSP et NegPSpan dont les sémantiques sont respectivement $\sqsubseteq_{\mathrm{O}}^{D}$ et $\preceq_{\mathrm{O}}^{D}$/$\preceq_{\bullet}^{D}$.

Tout d'abord, aucun des algorithmes ne répond à l'intuition des enquêté(e)s puisque les deux s'appuient sur une non-inclusion totale des itemsets tandis que c'est la non-inclusion partielle qui semble la plus intuitive. Une explication du choix algorithmique vient du fait que la non-inclusion partielle est antimonotone tandis que la relation totale est monotone. Cette dernière est moins facile à exploiter algorithmiquement. Les sémantiques les plus intuitives ne sont donc pas celles qui sont les plus appropriées algorithmiquement.

En pratique, on peut craindre des erreurs d'interprétation des motifs extraits par les algorithmes. Sans explicitation de la sémantique de ces derniers, les résultats de cette étude montrent que les motifs seront interprétés avec une sémantique différente de celle qui a servi à les extraire. Ceci constitue donc un problème important sur l'utilisation de ces algorithmes.

Une première recommandation serait alors de n'avoir que des singletons dans les négations d'un motif. Auquel cas, les non-inclusions partielles et totales sont équivalentes.

Une seconde solution serait de développer un algorithme alternatif adapté à une interprétation partielle de la non-inclusion. Ces adaptations sont algorithmiquement faisables. Il faudrait alors comparer leurs performances de calcul pour s'assurer que de tels algorithmes restent efficaces.

Néanmoins, les résultats montrent que le choix effectué par NegPSpan concernant la gestion des occurrences multiples répond à l'intuition d'un plus grand nombre. La seconde recommandation serait donc d'étendre préférentiellement l'algorithme NegPSpan.

Finalement, la troisième recommandation serait de promouvoir l'utilisation de syntaxes différenciées pour chaque sémantique. Cette recommandation avait été également suggérée dans [2].

8 Discussion

Cette enquête est probablement perfectible sur sa méthodologie. En particulier, il y a eu peu de questions pour décrire précisément le profil des enquêté(e)s. Ceci ne permet pas de savoir si la population enquêtée correspond bien à celle des utilisateurs potentiels d'algorithmes de fouille de motifs. De plus, la diffusion du questionnaire a été majoritairement effectuée via des canaux académiques. Ceci peut introduire un biais dans les réponses.

Une seconde limite sur la forme du questionnaire est la non-redondance des questions. En effet, chaque dimension d'interprétation de la sémantique des motifs avec négation ne fait l'objet que d'une seule question. Cela peut être jugé sensible à des erreurs. Nous avons opté pour un questionnaire plus court ne répétant par les questions. De plus, nous avons conçu le questionnaire pour séparer au mieux les différentes dimensions et ainsi éviter toute ambiguïté dans l'analyse des résultats.

La troisième limite est que le nombre de réponse au questionnaire peut sembler faible. La collecte de 124 questionnaires a nécessité plusieurs mois. La récupération d'un nombre significativement supérieur aurait nécessité d'autres stratégies de diffusion. De plus, ce nombre nous est apparu suffisant, au regard des questions et des résultats, pour faire une analyse solide statistiquement. En effet, les différences marquées dans les résultats permettent de fournir des résultats significatifs.

La qualité des réponses collectées est attestée par deux questions : une question préliminaire éliminatoire ainsi qu'une seconde question portant sur la portée de la négation qui nous sert à filtrer les enquêté(e)s qui viendraient biaiser les résultats. Le très faible nombre de ces personnes laissent penser que le jeu de réponses est de bonne qualité, *i.e.* que les personnes enquêtées ont consciencieusement répondu aux questions. Ceci est conforté par le fait qu'on ait pu attribuer une sémantique à 88% des enquêté(e)s

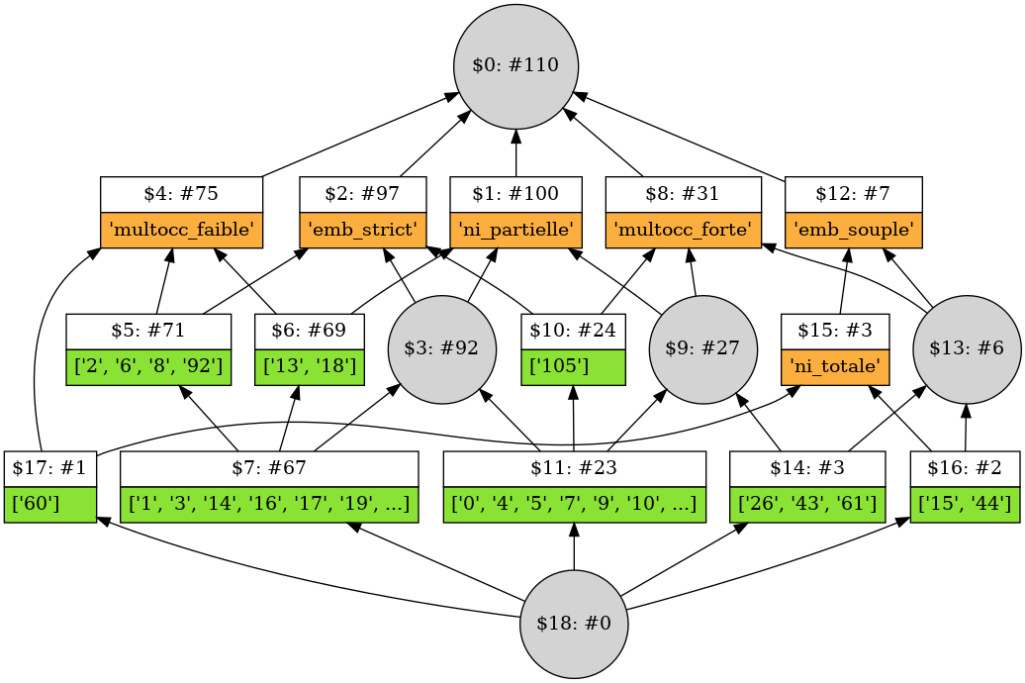


Figure 5: Concepts extraits à partir des réponses attributions faites pour chaque dimension.

montrant que les interprétations alternatives ou erreurs sont concentrées sur un petit nombre de personnes.

Un autre biais possiblement important de ce questionnaire est la présentation des notions élémentaires de motifs séquentiels qui pourraient avoir induit certaines réponses plutôt que d'autres. On peut en effet s'interroger sur le fait que les réponses pour les questions 3 et 4 aient des réponses aussi peu diverses. On s'attendait à avoir des perceptions plus hétérogènes de la notion de non-inclusion d'itemsets, mais cette diversité ne se retrouve pas dans le panel de personnes interrogées. Dans la mesure où la question 5 montre de la diversité dans les réponses, nous estimons que si l'hétérogénéité était réellement marquée dans les questions précédentes, elle serait apparue dans les réponses au questionnaire. Parmi les biais de présentation, la forme de présentation du questionnaire avec des symboles (et non des lettres) nous a été rapportée comme intéressante par certain(e)s enquêté(e)s. En effet, l'utilisation de lettres présuppose un ordre dans les items qui n'existe pas. En pratique, on a observé que seul 22.6% des enquêté(e)s étaient sensibles à l'ordre dans les itemsets. L'utilisation de symboles géométriques retranscrit mieux cette idée d'*itemset*. Malheureusement, nous n'avons pas collecté l'information sur le mode graphique du questionnaire effectivement utilisé. Nous ne pouvons donc pas vérifier cette hypothèse.

Finalement, le questionnaire est intimement lié au cadre d'analyse proposé par Besnard et Guyet [2] qui fait quelques hypothèses sur la forme des motifs avec négation, et leur sémantique. En particulier, nous avons vu ci-dessus que la question de l'insensibilité à l'ordre dans un itemset est une hypothèse forte que nous faisons. La seconde hypothèse est sur la portée de la négation. On s'aperçoit que 18.5% des enquêté(e)s n'ont pas répondu comme attendu à cette question. Comme nous avons écarté ces personnes de l'analyse, cela n'impacte pas les conclusions, mais cela soulève des questions sur l'"intuition" qu'ont eu ces personnes. Des interviews plus précises seraient ici nécessaires. Une troisième hypothèse est sur la syntaxe des motifs avec négation. Une étude plus complète pourrait s'intéresser à des syntaxes plus étendues : en permettant, par exemple, plusieurs négations consécutives, ou des négations en tête ou en queue d'un motif. Ces dernières possibilités existent chez certains algorithmes d'extraction de motifs de l'état de l'art [8].

9 Conclusion

Dans cet article, nous nous sommes intéressés à la sémantique des motifs séquentiels avec négation du point de vue des utilisateur(trice)s potentiels d'algorithmes d'extraction de tels motifs. L'intérêt de ce travail est de savoir si les motifs qui sont extraits par les algorithmes de l'état de l'art sont bien interprétés par les utilisateur(trice)s. En effet, les travaux de l'état de l'art avaient mis en évidence une ambiguïté dans ces notations [2]. Pour répondre à cette question, nous avons mené une enquête auprès d'utilisateur(trice)s potentiels ayant des profils variés. Cette enquête visait à comprendre les sémantiques auxquelles les utilisateurs adhéraient plus favorablement parmi celles qui avaient été identifiées.

L'analyse des réponses à l'enquête montre que deux sémantiques, dénotées $\sqsubseteq_{\mathrm{O}}^{G}$ et $\preceq_{\mathrm{O}}^{G}$, dominent dans le panel de 124 personnes interrogées. Il est tout d'abord intéressant

de constater qu'il n'existe pas une sémantique intuitive partagée uniformément. Les résultats sont également particulièrement intéressants par le fait que la préférence pour $\not\subseteq_G$ ne correspond pas à ce qui est utilisé dans les algorithmes majeurs de l'extraction de motifs avec négation (eNSP et NEGPSPAN). Cette relation intervenant lorsque la négation porte sur des ensembles d'item (*i.e.* $\neg(ab)$), une information particulière devrait être donnée aux utilisateurs sur les motifs comportant ce type de contrainte. Ensuite, $\preceq$ est majoritaire (à $\approx 69\%$) dans le panel et correspond au choix de l'algorithme NEGPSPAN. C'est également à la sémantique qui dispose des propriétés d'antimonotonie si les négations ne portent que sur des singletons.
À la suite de cette enquête, nous formulons les recommandations suivantes pour les méthodes d'extraction de motifs séquentiels avec négation :

- limiter l'usage de négation d'itemsets et privilégier l'utilisation de négation d'items,
- sinon, explorer l'extension l'algorithme NEGPSPAN dont la sémantique de la relation d'inclusion correspond à l'intuition majoritaire,
- proposer des syntaxes différenciées pour les différentes sémantiques.

References

[1] Radim Belohlavek, Erik Sigmund, and Jiří Zacpal. Evaluation of IPAQ questionnaires supported by formal concept analysis. *Information Sciences*, 181(10):1774–1786, 2011.

[2] Philippe Besnard and Thomas Guyet. Semantics of negative sequential patterns. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 1009–1015. IOS Press, 2020.

[3] Salah Eddine Boukhetta, Christophe Demko, Karell Bertet, Jérémy Richard, and Cécile Cayèré. Temporal sequence mining using FCA and GALACTIC. In *Graph-Based Representation and Reasoning*, pages 185–199. Springer International Publishing, 2021.

[4] Longbing Cao, Xiangjun Dong, and Zhigang Zheng. e-NSP: Efficient negative sequential pattern mining. *Artificial Intelligence*, 235:156–182, 2016.

[5] Xiangjun Dong, Yongshun Gong, and Longbing Cao. e-RNSP: An efficient method for mining repetition negative sequential patterns. *IEEE transactions on cybernetics*, 50(5):2084–2096, 2018.

[6] Bernhard Ganter and Rudolf Wille. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[7] Thomas Guyet and René Quiniou. NegPSpan: efficient extraction of negative sequential patterns with embedding constraints. *Data Mining and Knowledge Discovery*, 34(2):563–609, 2020.

[8] Sue-Chen Hsueh, Ming-Yen Lin, and Chien-Liang Chen. Mining negative sequential patterns for e-commerce recommendations. In *Proceedings of the Asia-Pacific Services Computing Conference*, pages 1213–1218. IEEE, 2008.

[9] Heikki Mannila and Hannu Toivonen. Discovering generalized episodes using minimal occurrences. In *Proceedings of the Conference on Knowledge Discovery and Delivery (KDD)*, volume 96, pages 146–151, 1996.

[10] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery*, 1(3):259–289, 1997.

[11] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 494–502, 1998.

[12] Wei Wang and Longbing Cao. Negative sequence analysis: A review. *ACM Computing Survey*, 52(2):32:1–32:39, 2019.

[13] Wei Wang and Longbing Cao. VM-NSP: vertical negative sequential pattern mining with loose negative element constraints. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–27, 2021.

[14] Xindong Wu, Chengqi Zhang, and Shichao Zhang. Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS)*, 22(3):381–405, 2004.

[15] Tiantian Xu, Xiangjun Dong, Jianliang Xu, and Xue Dong. Mining high utility sequential patterns with negative item values. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 31(10):1750035, 2017.

[16] Tiantian Xu, Xiangjun Dong, Jianliang Xu, and Yongshun Gong. E-msNSP: Efficient negative sequential patterns mining based on multiple minimum supports. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 31(02):1750003, 2017.

Sémantique Formelle à Deux Joueurs pour Arbres d'Attaque

T. Brihaye[1], S. Pinchinat[2], A. Terefenko[1,2]

[1] Université de Mons, Belgique
[2] Université de Rennes, IRISA, France

Résumé

Les arbres d'attaques sont un formalisme utilisé en sécurité pour l'évaluation de menace en analyse de risque. En 2017, M. Audinot et al. ont introduit une sémantique de chemins sur un système de transition pour les arbres d'attaque. Cette approche ne permet pas de considérer des systèmes avec plusieurs acteurs. Inspiré par ce travail, nous proposons une interprétation à deux joueurs de ce formalisme en généralisant la sémantique de chemins à une sémantique de stratégies. Dans ce cadre, nous montrons que le problème de la vacuité d'un arbre d'attaque est PSPACE*-complet, alors que le problème d'appartenance d'une stratégie à la dénotation d'un arbre est* CONP*-complet.*

Mots-clés

arbre d'attaque, sémantique, arène de jeu, stratégie.

Abstract

Attack trees are a fromalism used in security for the evaluation of threat in risk analysis. In 2017, M. Audinot et al. introduced a path semantics over a transition system for attack trees. This approach does not allow to consider multi-agent systems. Inspired by the latter, we propose a two-player interpretation of this formalism by generalising the path semantics to a strategy semantics. We then show that the emptiness problem for an attack tree is PSPACE*-complete and the membership problem for a strategy to the description of an attack tree is* CONP*-complete.*

Keywords

Attack tree, semantics, game arena, strategy.

1 Introduction

La sécurité est un sujet d'attention croissante dans notre société actuelle pour protéger les ressources critiques de divulgation d'informations, de vol ou de dégâts. Le modèle informel d'arbre d'attaque a été d'abord introduit pas Schneier [3] pour représenter les menaces possibles sur un système informatique. Les arbres d'attaques ont depuis été grandement utilisés dans l'industrie et sont conseillés dans le rapport de l'OTAN de 2008 pour régir l'évaluation des menaces dans l'analyse de risque. Le modèle des arbres d'attaque est un sujet d'intérêt croissant dans la communauté des méthodes formelles avec de nombreuses d'approches différentes (voir le survey [4]).

Le premier modèle formel d'arbre d'attaque introduit dans [3] visait à décrire les attaques possibles sur un système par raffinement de l'objectif principal en sous-objectifs coordonnés soit avec l'opérateur OR soit avec l'opérateur AND. La sémantique sera ensuite augmentée ([1]) avec l'opérateur $SAND$ (pour "AND séquentiel"), qui exprime que les sous-objectifs doivent être atteints dans un ordre donné. En particulier, les auteurs de [1] introduisent une sémantique de chemins sur un système de transition.

Dans notre article [2], notre objectif est de proposer une nouvelle sémantique des arbres d'attaque plus réaliste : nous voulons que nos attaquants soient capables d'adapter leurs actions en fonction de l'environnement, donnant naturellement une sémantique à deux joueurs. Notre approche généralise [1] à un cadre de la théorie des jeux, menant à une sémantique de stratégie. Nous présentons ici un résumé de notre contribution.

2 Syntaxe des arbres d'attaque

Un *arbre d'attaque* est un modèle qui spécifie l'objectif d'un des deux joueur (l'*attaquant*). Etant donné un ensemble de propositions $Prop$, un arbre d'attaque sur $Prop$ est :

- soit une feuille composée d'une unique formule propositionelle ϕ sur $Prop$,
- soit une expression $OP(\tau_1, ..., \tau_n)$ où $\tau_1, ..., \tau_n$ sont des arbres d'attaque et OP est un opérateur parmis OR, AND ou $SAND$.

Nous modélisons un système multi-joueur, par une *arène de jeu concurrente* : un graphe fini sur lequel deux joueurs jouent un jeu de durée non bornée et où ils choisissent une action à chaque tour.

Exemple 2.1. Supposons que l'attaquant essaie de rentrer dans un bâtiment à deux portes d'entrée (d_1 et d_2), sans être vu par le garde qui contrôle l'accès à une des portes. Ce garde peut se déplacer d'une porte à l'autre, mais ne peut pas contrôler les deux portes en même temps. La situation est représentée par l'arène en Figure 1 où, dans chaque état, la première lettre représente la position de l'attaquant (o il est à l'extérieur du bâtiment, d_1 il est à la première porte, et d_2 il est à la seconde porte) et la deuxième lettre représente la position du garde (m il est en mouvement entre les deux portes, d_1 il est à la première porte, et d_2 il est à la seconde porte). Considérons l'ensemble $Prop = \{seen, d_1, d_2\}$ où d_1 et d_2 décrivent la position de l'attaquant et $seen$ est vrai dans les états (d_1, d_1) et (d_2, d_2), c'est-à-dire lorsque garde et le voleur sont à la même porte. L'objectif de l'attaquant est décrit par l'arbre $\tau = OR(d_1 \wedge \neg seen, d_2 \wedge \neg seen)$.

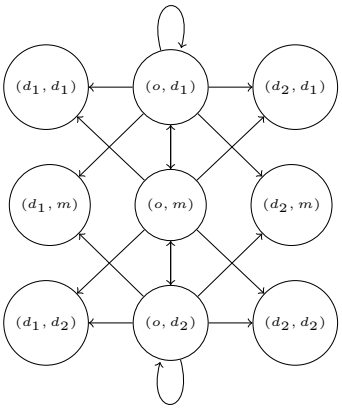


FIGURE 1 – Position du voleur et du garde

3 Sémantique de chemins

Remarquons qu'une arène de jeu peut être vue comme un système de transition $\mathcal{S}$ si on oublie les joueurs. La sémantique de chemins d'un arbre d'attaque τ est une sémantique (1 joueur) consistant en l'ensemble des chemins (suite d'états) du système de transition constituant un scénario favorable pour l'attaquant :

Définition 3.1 ([1])**.** La *sémantique de chemins* $Paths_{\mathcal{S}}(\tau)$ de τ sur le système $\mathcal{S}$ est définie inductivement :

- $Paths_{\mathcal{S}}(\phi)$ est l'ensemble des chemins terminant par un état où ϕ est vrai,
- $Paths_{\mathcal{S}}(OR(\tau_1, ..., \tau_n))$ est l'union des sémantiques des enfants,
- $Paths_{\mathcal{S}}(SAND(\tau_1, ..., \tau_n))$ est la concaténation des sémantiques des enfants,
- $Paths_{\mathcal{S}}(AND(\tau_1, ..., \tau_n))$ est obtenu par shuffle des sémantiques des enfants.

Exemple 3.2. Dans l'Exemple 2.1, la séquence d'états $(o, d_1)(o, m)(d_1, d_2)$ est un chemin dans la sémantique de la feuille $d_1 \wedge \neg seen$. Ce chemin est donc également dans la sémantique de $\tau = OR(d_1 \wedge \neg seen, d_2 \wedge \neg seen)$.

4 Sémantique de stratégies

Obtenir une sémantique compositionnelle représentant les stratégies que notre attaquant peut appliquer pour atteindre son objectif, peu importe le comportement du défenseur n'est pas immédiat, comme le montre l'exemple suivant.

Exemple 4.1. Dans l'Exemple 2.1, il n'existe aucune stratégie, ni pour l'objectif $d_1 \wedge \neg seen$, ni pour l'objectif $d_2 \wedge \neg seen$, puisque le garde peut décider de rester indéfiniment à la même porte. Toutefois, il existe bien une stratégie pour l'objectif $OR(d_1 \wedge \neg seen, d_2 \wedge \neg seen)$: il suffit d'attendre une unité de temps pour voir vers quelle porte le garde va se déplacer et ensuite d'entrer par l'autre porte.

La situation de l'Exemple 2.1 écarte une définition compositionnelle de la semantique de stratégies car nous ne pouvons pas déduire inductivement une sémantique non vide pour un arbre d'attaque alors que ses sous-arbres ont une sémantique vide. Nous utilisons donc la définition suivante.

Définition 4.2. Pour un arbre d'attaque τ, sa sémantique de stratégie $Strat_{\mathcal{G}}(\tau)$ est l'ensemble des stratégies assurant à l'attaquant de jouer des parties (chemins) se trouvant dans $Paths_{\mathcal{S}}(\tau)$.

5 Problèmes de décisions

Nous avons étudié deux problèmes de décisions.

Définition 5.1. Le problème de *non-vacuité* (NV) pour une sémantique fixée d'arbre d'attaque $[\![\cdot]\!]_{\mathcal{G}}$ est le problème de décision suivant :
Entrée : $\mathcal{G}$, une arène de jeu, τ, un arbre d'attaque.
Sortie : Oui si $[\![\tau]\!]_{\mathcal{G}} \neq \emptyset$, Non sinon.

Définition 5.2. Le problème de *l'appartenance* (A) pour une sémantique fixée d'arbre d'attaque $[\![\cdot]\!]_{\mathcal{G}}$ de type X (chemins ou stratégies) est le problème de décision suivant :
Entrée : $\mathcal{G}$, une arène de jeu, τ, un arbre d'attaque et $x \in X$.
Sortie : Oui si $x \in [\![\tau]\!]_{\mathcal{G}}$, Non sinon.

Où $[\![\tau]\!]_{\mathcal{G}}$ fait référence à soit la sémantique de chemins, soit la sémantique de stratégie. Les résultats sont repris dans la table suivante où [2] est notre contribution.

	sémantique de chemin	sémantique de stratégie
NV	NP-**complet** [1] (SMP[*] et réduction de SAT)	PSPACE-**complet** [2] (algorithme alternant polynomial et réduction de QBF)
A	P [2] (Backward induction)	coNP-**complet** [2] (SMP et réduction de UNSAT)

[*] Small model property.

Ce travail a été en partie soutenu par le Fonds de la Recherche Scientifique - FNRS sous la subvention n°T.0027.21.

Références

[1] Maxime Audinot, Sophie Pinchinat, and Barbara Kordy. Is my attack tree correct ? In *European Symposium on Research in Computer Security*, pages 83–102. Springer, 2017.

[2] Thomas Brihaye, Sophie Pinchinat, and Alexandre Terefenko. Adversarial formal semantics of attack trees and related problems. In Pierre Ganty and Dario Della Monica, editors, *Proceedings of the 13th International Symposium on Games, Automata, Logics and Formal Verification, GandALF 2022, Madrid, Spain, September 21-23, 2022*, volume 370 of *EPTCS*, pages 162–177, 2022.

[3] Bruce Schneier. Attack trees. *Dr. Dobb's journal*, 24(12) :21–29, 1999.

[4] Wojciech Wideł, Maxime Audinot, Barbara Fila, and Sophie Pinchinat. Beyond 2014 : Formal methods for attack tree–based security modeling. *ACM Computing Surveys (CSUR)*, 52(4) :1–36, 2019.

Session 4 : Explicabilité et équité 1

Normes techniques et éthique de l'IA

M. Gornet[1], W. Maxwell[1]

[1] Télécom Paris, IP Paris - Institut Polytechnique de Paris
NOS - Numérique, Organisation et Société
i3 - institut interdisciplinaire de l'innovation

melanie.gornet@telecom-paris.fr

Résumé

Les progrès technologiques de l'IA engendrent des avancées majeures, mais ont aussi des conséquences sociales importantes qui nécessitent de réguler les pratiques. Le projet de règlement européen sur l'IA, également connu sous le nom d'AI Act, imposera aux fournisseurs de systèmes d'IA à haut risque un nombre d'obligations pouvant être qualifiées d'« éthiques » [1]*, notamment le respect de l'équité et des droits fondamentaux. Pour s'assurer de la conformité de ces systèmes aux exigences éthiques, le projet d'AI Act prévoit le recours aux normes harmonisées, soulevant ainsi la question de la compatibilité entre les normes techniques et les enjeux éthiques. Dans cette étude, nous contribuons à ce débat en rappelant le rôle des normes et de la certification en Europe, avant de présenter les différents acteurs de la normalisation travaillant actuellement sur des normes « éthiques » pour l'IA. Nous montrons à travers cet inventaire la diversité de leurs travaux ainsi que la concurrence émergente entre différentes visions de l'éthique de l'IA. Enfin, nous discutons des risques que soulèvent ces normes, tels que la difficulté à définir des critères objectifs et la possibilité de confusion avec une garantie de l'éthique.*

Mots-clés

normes, certification, éthique de l'AI, droit de l'IA, régulation, AI Act

Abstract

Technological advancements in AI are generating major breakthroughs, but they also have significant social consequences that require regulation. The proposed European regulation on AI, also known as the AI Act, will impose a number of obligations on providers of high-risk AI systems, that can be considered as "ethical" obligations, including respect for fairness and fundamental rights. To ensure compliance of these systems with ethical requirements, the proposed AI Act plans to use harmonized standards, raising the question of compatibility between technical standards and ethical issues. In this paper, we contribute to this debate by recalling the role of standards and certification in Europe, before presenting the actors currently working on "ethical" AI standards. We show through this inventory the diversity of their work and the competition that is emerging between different visions of AI ethics. Finally, we discuss the risks raised by these standards, such as the difficulty of defining objective criteria and the possibility that citizens may be misled.

Keywords

normes, certification, AI ethics, AI law, regulation, AI Act

1 Introduction

Les systèmes dits d'« intelligence artificielle » [2] (IA), ont été l'objet de nombreuses controverses ces dernières années, renforcées par leur couverture médiatique. Entre autres, l'opacité des systèmes est fortement critiquée car elle serait un obstacle à la compréhension des prises de décision. Les nombreux biais indésirables des systèmes d'IA ont également été pointés du doigt, renforçant les discriminations dans le domaine de la sécurité [3], de la justice [4], dans l'accès à l'emploi [5] ou aux aides sociales [6]. De nombreuses initiatives ont ainsi vu le jour, appelant à prendre en considération un certain nombre de principes éthiques lors du cycle de vie des systèmes d'IA. Parmi ces principes, l'équité, l'explicabilité, la transparence ou encore la vie privée s'érigent en préceptes universels [33]. Pourtant, si ces principes font consensus, l'absence d'instructions claires, concrètes et opérationnelles sur la manière d'atteindre réellement ces objectifs rend leur adoption complèxe.

1. L'adjectif « éthique » est définit dans le dictionnaire Larousse par : « Qui concerne la morale », `https://www.larousse.fr/dictionnaires/francais/\%C3\%A9thique/31388`; le dictionnaire Le Robert ajoute un second sens : « Qui intègre des critères moraux dans son fonctionnement », `https://dictionnaire.lerobert.com/definition/ethique`.

2. Le terme « intelligence artificielle » est très controversé, notamment pour son caractère anthropomorphe, c'est-à-dire rappelant des caractéristiques normalement réservés aux êtres humains [56].

3. Par exemple l'arrestation à tort de personnes afro-américaines, mal reconnues par les algorithmes de reconnaissance faciale des forces de l'ordre aux États-Unis [29].

4. Par exemple le logiciel COMPAS utilisé par la justice américaine pour prédire le taux de récidive de criminels, qui semble prédire un risque plus élevé pour les personnes afro-américaines [5].

5. Par exemple le logiciel de tri des CVs d'Amazon qui rejetait plus facilement les candidates femmes [53].

6. Par exemple l'algorithme de détection de fraudes aux allocations familiales de l'autorité fiscale néerlandaise, signalant davantage les personnes issues de l'immigration [55].

En avril 2021, la Commission Européenne a proposé un projet de règlement établissant des règles harmonisées en matière d'intelligence artificielle, nommé *AI Act*, et inspiré pour partie de ses travaux préliminaires sur l'éthique de l'IA. Le texte prévoit le recours à des normes harmonisées pour respecter les exigences essentielles [30]. Alors que ces normes ont jusque-là été utilisées pour codifier des critères techniques, elles vont, dans le cadre de l'IA, devoir adresser des problèmes éthiques. Mais est-il possible de réguler des questions éthiques par des normes techniques ?

La question des normes pour traiter de l'éthique de l'IA est quasiment absente de la littérature. En 2017, [7] examinait les premières initiatives de normes éthiques, en particulier les initiatives de l'IEEE, mais ces initiatives se sont depuis démultipliées. D'autres travaux ont recensé les projets de normes pour l'IA, sans différencier les normes traitant des aspects éthiques [22, 64] [7].

Dans cette étude, nous proposons d'examiner les différentes normes en préparation en matière d'éthique de l'IA, les acteurs qui les développent, et les dynamiques qui se dégagent. Cet inventaire permettra d'apporter des premières réponses sur le rôle des normes et de la certification en matière d'éthique de l'IA.

Nous commençons par rappeler dans la Section 2 le fonctionnement des procédures de normalisation et de certification, ainsi que la manière dont l'*AI Act* introduit le recours à des normes dites « éthiques ». Nous dressons dans la Section 3 une vue d'ensemble des activités des différents acteurs de la normalisation et de leurs visions de l'éthique, mettant en lumière leurs désaccords et leur rivalité. Nous discutons dans la Section 4 des problèmes posés par ces normes, notamment la difficulté de définir ce qu'est, ou n'est pas une norme « éthique », la subjectivité de leur contenu, et la certification vis à vis de ces normes qui s'érige insidieusement comme une garantie de l'éthique. Enfin, en conclusion, nous dégageons quelques pistes de réflexion pour contribuer au débat plus large sur la coexistence entre normes techniques et débats éthiques.

2 Rôle des normes techniques et de la certification

2.1 Normes techniques et organismes de normalisation

Les normes sont des documents techniques destinés à établir des solutions communes à des exigences données [8] [10]. Elles permettent notamment de *« définir un langage commun entre les acteurs, de clarifier, d'harmoniser les pratiques et de définir le niveau de qualité [...] des produits [et des] services »* [2]. Leur application, que ce soit dans le secteur privé ou public, est volontaire : il n'est jamais exigé légalement d'appliquer une norme, mais cela peut faciliter la certification [9]. Toutefois, l'accès aux normes est souvent payant [10].

Les normes peuvent remplir plusieurs rôles : elles aident les entreprises à optimiser les coûts et à augmenter leur efficacité [19], elles contribuent à dynamiser l'économie [25], à stimuler l'innovation [4], à encourager la concurrence entre les entreprises [6], et à assurer la protection des consommateurs [52].

Les plus connues sont les normes ISO, du nom de l'organisme les développant : l'Organisation internationale de normalisation, ou *International Organization for Standardization* en anglais. Les normes ISO sont couramment utilisées pour certifier la qualité des produits ou des services d'une entreprise [11], sa performance environnementale [12], ou encore la sécurité de ses systèmes informatiques [13]. Dans le cadre des systèmes numériques, l'ISO collabore souvent avec la Commission électrotechnique internationale (IEC) [14]. Enfin, une troisième organisation, nommé l'Union internationale des télécommunications (ITU) [15] qui est une agence des Nations Unies, participe également à établir des normes pour les technologies de l'information et de la communication. L'ISO, l'IEC et l'ITU sont les trois agences de normalisation principales participant à l'élaboration de normes pour l'IA. Mais d'autres organismes de normalisation du monde du numérique comme le World Wide Web Consortium (W3C) [16], lancent également des groupes de travail sur l'IA [17].

Au niveau européen, trois organismes se chargent du développement de normes : le Comité européen de normalisation (CEN), le Comité européen de normalisation électrotechnique (CENELEC) [18] et l'Institut européen des normes de télécommunication (ETSI) [19]. Ils sont parfois nommés ESOs, pour *European Standardisation Organisations* [20].

Les normes destinées à soutenir la législation européenne sont publiées au Journal Officiel de l'Union Européenne [21], elles sont alors appelées « normes harmonisées » [22]. Ces

7. Voir également l'initiative du AI Standards Hub : https://aistandardshub.org/ai-standards-search/
8. Traduction des auteurs.
9. Voir Section 2.2
10. C'est le cas notamment des normes ISO.
11. Par exemple, la norme ISO 9001 :2015 - Quality management systems et ses normes connexes : https://www.iso.org/iso-9001-quality-management.html
12. Par exemple, la norme ISO 14001 :2015 - Environmental management systems et ses normes connexes : https://www.iso.org/iso-14001-environmental-management.html
13. Par exemple, la norme ISO 27001 :2022 - Information security, cybersecurity and privacy protection et ses normes connexes : https://www.iso.org/isoiec-27001-information-security.html
14. https://iec.ch/homepage
15. https://www.itu.int/en/Pages/default.aspx
16. https://www.w3.org/
17. https://www.w3.org/blog/2021/04/w3c-launches-the-web-machine-learning-working-group/
18. CEN et CENELEC travaillent régulièrement ensemble pour le développement des normes, https://www.cencenelec.eu/
19. https://www.etsi.org/
20. Voir le site de la Commission Européenne recensant les acteurs clés de la normalisation : https://single-market-economy.ec.europa.eu/single-market/european-standards/key-players-european-standardisation_en
21. https://eur-lex.europa.eu/oj/direct-access.html?locale=fr
22. Voir le site de la Commission Européenne sur les normes harmonisées : https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards_en

normes harmonisées peuvent être des normes européennes développées par la CEN, la CENELEC ou l'ETSI, mais peuvent également être des normes internationales adoptées par les ESOs après vote des membres. Une fois adoptées au niveau européen, ces normes peuvent alors être adoptées, par ricochet, par les organismes nationaux.

En effet, chacun des pays membres de l'Union possède son propre organisme de normalisation. Par exemple, en France il s'agit de l'Association Française de Normalisation (AFNOR) et en Allemagne du *Deutsches Institut für Normung* (DIN). Ces organismes nationaux sont parfois nommés NSBs, pour *National standardisation bodies* [23]. Ils sont constitués d'experts réunis en différentes commissions, comme la Commission de Normalisation sur l'IA (CN IA) de l'AFNOR, qui votent pour l'adoption de normes, le lancement de projets de normes, et participent à la création de normes dans les instances européennes, comme la CEN-CENELEC, ou internationales, comme l'ISO. Ces experts sont souvent des acteurs industriels, représentant les intérêts de leurs entreprises d'origines, ces dernières ayant passé un contrat de prestation avec un NSB. Mais les experts peuvent également être issus d'instituts de recherche ou d'établissements publics par exemple. Tout le monde peut ainsi demander à rejoindre un NSB pour participer à l'élaboration de normes et aux votes des CN, en échange de frais d'adhésion.

2.2 Processus de certification

Les processus de certification interviennent indépendamment de la création de normes : ni l'ISO, ni les ESOs, ni les NSBs ne se chargent de les mener. Une fois une norme publiée, un organisme indépendant peut proposer une assurance écrite, appelée « certificat », attestant que le produit ou service répond aux exigences définies dans la norme [24]. Ces organismes de certification peuvent être des organismes officiels [25] ou non.

En Europe, il existe un type particulier de certification appelé « marquage CE », pour « Conformité Européenne ». Pour certains groupes de produits, cette marque est obligatoire pour pouvoir entrer sur le marché européen. C'est le cas, par exemple, des produits électriques et électroniques, des jouets, des dispositifs médicaux, des machines, etc [27]. Elle indique la conformité aux exigences essentielles définies dans les directives européennes et est appliquée directement par le constructeur. En effet, le fabriquant est le seul responsable de la déclaration de conformité. Il peut choisir d'évaluer lui-même son produit, ou de faire appel à un organisme notifié, i.e. une organisation désignée par un pays de l'UE pour évaluer la conformité de certains produits avant leur mise sur le marché [28]. Le recours à un organisme notifié est même obligatoire pour certaines catégories de produits. Dans le cas d'un recours à un organisme notifié, le fabriquant doit constituer un dossier technique documentant la conformité, puis signer la déclaration de conformité européenne. Il doit être en mesure de fournir les documents justificatifs du marquage CE sur demande de l'autorité nationale compétente [29]. Toutefois, le marquage ne signifie pas que l'UE a approuvé un produit comme étant sûr ou conforme [15], son apposition sur le produit n'engage que le fabriquant.

Bien que les normes soient volontaires, en Europe les produits fabriqués dans le respect des normes harmonisées bénéficient d'une présomption de conformité vis-à-vis de la législation. Les fabricants peuvent alors bénéficier de procédures d'évaluation de la conformité simplifiées. S'ils choisissent de ne pas appliquer les normes harmonisées, ils doivent démontrer par d'autres moyens que le produit satisfait aux exigences essentielles définies dans la directive ou dans le règlement européen correspondant [28].

En plus de ces certifications encadrées par la loi, certaines entreprises ou associations développent elles-mêmes leurs propres critères pour juger de la qualité d'un produit. Ces marques délivrées par des organismes privés sont appelées « labels » et ont l'avantage d'être plus rapides à mettre en place que la certification aux normes. Ils sont notamment très utilisés pour le respect de critères environnementaux [30].

Ainsi, la certification publique, basée sur la mise en conformité à partir normes harmonisées, est destinée à faciliter le commerce tout en garantissant le respect des exigences réglementaires. La certification privée quant à elle, qu'elle soit basée sur des normes européennes non-harmonisées, des normes internationales ou sur des cadres d'évaluation privés, est davantage présentée comme un repère pour le consommateur, qui sert à démontrer la qualité supérieur d'un produit et à améliorer son potentiel commercial.

2.3 *AI Act* et normes « éthiques »

L'*AI Act* [23] est aujourd'hui encore en discussion au sein des instances européennes, le texte n'est donc pas définitif et est amené à évoluer. Nous nous baserons ici sur la proposition de la Commission européenne datant d'avril 2021. Dans la proposition de règlement, les systèmes d'IA sont classés en plusieurs catégories, selon leur niveau de risques : risque minimal (art. 69), risque faible (art. 52), haut risque (art. 6 et suivants) et risque inacceptable (art. 5). Ces derniers seront interdits d'utilisation, tandis que les systèmes à risque minimal ou faible seront autorisés respectivement sans restriction, et simplement avec des obligations

23. Voir supra note 20

24. Voir le site de l'ISO sur la certification : https://www.iso.org/certification.html

25. Les organismes de certification peuvent être accrédités. L'accréditation consiste à recevoir une reconnaissance officielle par un organisme indépendant, appelé organisme d'accréditation, que l'organisme de certification fonctionne bien conformément aux normes internationales. Toutefois, l'accréditation n'est pas obligatoire pour devenir organisme de certification [26].

27. Une exception notable est le cas des véhicules tels que les voitures, les bus, les camions ou les tracteurs qui suivent une procédure spécifique.

28. Voir le site de la Commission Européenne sur les organismes notifiés : https://single-market-economy.ec.europa.eu/single-market/goods/building-blocks/notified-bodies_en

29. Voir le site de la Commission Européenne sur le marquage CE : https://europa.eu/youreurope/business/product-requirements/labels-markings/ce-marking/index_en.htm

30. Voir le guide du ministère de la transition écologique sur les labels environnementaux : https://www.ecologie.gouv.fr/labels-environnementaux

d'information et de transparence. Les systèmes soumis à une évaluation de conformité seront seulement les systèmes jugés à haut risque, qui devront être certifiés CE.

La Commission insiste sur sa volonté d'intégrer des considérations « éthiques » pour l'encadrement des systèmes d'IA. Ainsi, selon le considérant 5 [31], le règlement contribue à *« faire de l'Union un acteur mondial de premier plan dans le développement d'une intelligence artificielle sûre, fiable et éthique, et il garantit la protection de principes éthiques »*. Selon l'exposé des motifs, le règlement a pour but de renforcer *« la contribution de l'Union à la définition de normes mondiales et à la promotion d'une IA digne de confiance qui soit conforme aux valeurs et aux intérêts de l'Union »*. Il est également précisé que les exigences minimales proposées s'inspirent des lignes directrices en matière d'éthique du Groupe d'Experts de Haut Niveau sur l'IA de la Commission européenne (GEHN) [32] [49].

Ces considérations éthiques se reflètent dans l'attention particulière portée aux violations des droits individuels. Ainsi un système peut être considéré à haut risque s'il présente *« un risque de préjudice pour la santé et la sécurité, ou un risque d'incidence négative sur les droits fondamentaux »* (art. 7).

Pour réguler l'IA, la Commission a fait le choix du modèle de la conformité [9] : les systèmes d'IA devront respecter les critères définis par la réglementation *avant* de pouvoir être distribués sur le marché européen [30]. Ce choix est également celui de nombreux textes en droit du numérique [9]. La normalisation devrait alors *« jouer un rôle essentiel [...] afin de garantir la conformité »* (considérant 61). Cependant, les exigences minimales étant inspirées de critères éthiques, les normes conçues pour les évaluer touchent par là-même à des aspects éthiques.

Les normes dites « éthiques » se sont fortement développées ces dernières années. La première norme a traiter explicitement d'éthique en robotique date de 2016, elle a été suivie depuis par d'autres initiatives comme celle de l'IEEE [7]. Aujourd'hui, de nombreux acteurs développent des normes touchant de près ou de loin à des aspects éthiques. Nous étudions cet écosystème dans la Section 3.

Néanmoins, ces normes restent marginales. Par exemple, le Règlement n°1025/2012 relatif à la normalisation européenne liste les éléments pouvant être considérés comme des spécifications techniques (art. 2.4.a) [1], comme la protection de l'environnement, la santé, ou encore la sécurité, mais n'inclut pas les critères éthiques.

31. Selon le dictionnaire Larousse, un considérant est un alinéas dans un arrêt d'une cour ou dans une décision de juridiction administrative qui motive la décision. https://www.larousse.fr/dictionnaires/francais/consid\%C3\%A9rant/18384

32. Plus connu sous le nom de *High-Level Expert Group on Artificial Intelligence*, ou *HLEG*, en anglais.

3 Une multitude d'acteurs et d'initiatives en matière de normes éthiques pour l'IA

3.1 Une ruée vers les normes pour accompagner les progrès techniques

Il existe un décalage entre la vitesse de déploiement des produits et services basés sur l'IA et le développement des normes. Cela crée une pression sur les organismes de normalisation afin de publier des normes le plus rapidement possible. À cela s'ajoute les contraintes dues au calendrier de l'*AI Act* et de sa mise en application.

L'ISO est en avance dans la création de normes techniques pour l'IA : 17 standards publiés et 24 en développement aujourd'hui [33], en comptant simplement les normes transverses [34]. Parmi ces normes pour l'IA, de nombreuses initiatives adressent des questions éthiques larges comme les impacts sociaux de la technologie [35], tentent de définir ce qu'est une IA « de confiance » [36], ou adaptent le management de risques à l'IA [37]. D'autres normes choisissent une approche plus pratique, définissant des méthodes de conception pour éviter le traitement des biais [38], ou des critères de qualité des données [39].

Au contraire, les européens CEN et CENELEC démarrent à peine leur campagne de création de normes relative à l'*AI Act*. À ce jour, aucune norme n'a encore été publiée dans ce cadre. Seule leur feuille de route [11] laisse transparaître les objectifs d'adoption de normes. Dans cette feuille de route, plusieurs thèmes clés sont identifiés pour les travaux de normalisation futurs : la terminologie, la confiance, l'éthique, la sécurité et la sûreté, la résilience et la souveraineté, le respect du droit [40]. De son côté, l'ETSI travaille également sur la création de normes dans le cadre de l'*AI Act*, par exemple sur l'évaluation des systèmes d'IA, la définition de paramètres de qualité, l'explicabilité et la transparence des traitements, ou encore la traçabilité des modèles [41].

33. Chiffres relevés en février 2023, selon le site de l'ISO : https://www.iso.org/committee/6794475.html

34. L'ISO comporte plusieurs sous-comités, comme l'ISO JTC1/SC42 travaillant sur des normes transverses à tous les systèmes d'IA, mais également des sous-comités dédiés à certaines applications tel que l'ISO JTC1/SC37 travaillant sur les systèmes biométriques. Nous décomptons ici simplement le travail du JTC1/SC42.

35. ISO/IEC TR 24368 :2022, Information technology — Artificial intelligence — Overview of ethical and societal concerns : https://www.iso.org/standard/78507.html

36. ISO/IEC TR 24028 :2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence : https://www.iso.org/standard/77608.html

37. ISO/IEC 23894 :2023 Information technology — Artificial intelligence — Guidance on risk management : https://www.iso.org/standard/77304.html

38. ISO/IEC TR 24027 :2021, Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making : https://www.iso.org/standard/77607.html ; ISO/IEC CD TS 12791 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks : https://www.iso.org/standard/84110.html

39. ISO 8000-1 :2022 Data quality — Part 1 : Overview : https://www.iso.org/standard/81745.html

40. Voir Annexe D de la feuille de route [11].

Le développement de normes harmonisées pour l'*AI Act*, comprenant la création de nouvelles normes et l'adoption par les organismes des pays membres de normes ISO déjà publiées, doit se poursuivre jusqu'en 2025. Les normes européennes jouent donc avec un calendrier très serré.

L'une des problématiques liées aux contraintes temporelles est que les normes publiées en premier ont tendance à être adoptées plus facilement que celles qui suivent. Toutefois, si les normes arrivent trop tôt, elles peuvent conduire à une sélection prématurée et inefficace de la technologie [6]. C'est le cas du clavier QWERTY qui, malgré ses nombreux inconvénients, s'est imposé sur le marché au détriment d'autres bien meilleures solutions [4]. Il y a un risque que cela se reproduise pour l'IA, en normalisant des pratiques encore imparfaites et empêchant de nouvelles pratiques, plus éthiques, de se mettre en place.

De plus, les normes et les processus de certification sont généralement longs à mettre en place. Ainsi, les labels indépendants ont plus de chances de se développer rapidement, avant la création de normes harmonisées.

3.2 La coexistence de normes européennes et internationales cache des enjeux géopolitiques

Les organismes européens et internationaux ont l'habitude de collaborer sur l'élaboration de normes. Notamment, les accords de Vienne et de Francfort, conclus respectivement entre la CEN et l'ISO, et entre la CENELEC et l'IEC, facilitent les échanges d'informations entre les organismes et évitent les doublons dans les travaux [32, 24]. Cette collaboration se poursuit jusque dans l'adoption de normes puisque les normes ISO et IEC peuvent intégrer le catalogue des normes européennes par ratification par le CEN-CENELEC. Actuellement, près de 33% des publications du CEN sont issues de l'ISO, et 73% de celles du CENELEC sont issues de l'IEC. En ce qui concerne les normes harmonisées, les normes ISO et IEC ont la priorité lorsqu'elles existent, à moins qu'il ne soit prouvé que la demande de la Commission ne peut être satisfaite par les normes issues de ces organismes internationaux [16].

Dans le cadre de l'IA, les normes européennes sont en retard par rapport aux normes internationales [41]. Ainsi, des appels sont lancés pour d'approfondir les liens avec l'ISO [44] et converger avec les normes internationales [31]. Cependant, le modèle de conformité présenté par la Commission dans l'*AI Act* renforce la distinction entre les normes européennes harmonisées, bénéficiant d'une présomption de conformité, et les normes internationales non-ratifiées [37]. Il existe donc une tension considérable quant à savoir qui des instances européennes ou internationales développera les normes qui façonneront l'IA en Europe. Cette tension est d'autant plus importante en ce qui concerne les normes relatives aux aspects éthiques des systèmes, où les visions peuvent diverger entre l'Union européenne et d'autres régions du monde.

Ainsi, certaines parties prenantes remettent en question la confiance dans les normes internationales, car selon elles, rien ne garantit que ces normes soient conformes aux droits et valeurs de l'UE [21]. L'ANEC [42], une organisation qui défend les intérêts des consommateurs européens dans les processus de normalisation et de certification, se dit ainsi préoccupée par l'adoption en Europe de normes auxquelles ont participé des pays ou des entreprises non européens [54]. Elle propose que la Commission précise si une norme harmonisée peut être confiée à l'ISO, ou si elle doit être développée au sein des organismes de normalisation européens afin de *« préserver les valeurs ou l'éthique européennes »* [43] [54]. De plus, elle insiste sur l'importance de ne pas mettre en péril les valeurs fondamentales européennes dans le seul but de réduire les délais de développement [54]. Cela implique de ne pas laisser aux organismes internationaux la souveraineté des normes et de prendre le temps de développer des normes européennes [13].

La composition de ces organismes de normalisation est également au coeur du débat. Si une grande partie des membres de l'ISO viennent d'Europe de l'ouest, près de la moitié viennent d'ailleurs dans le monde, particulièrement d'Asie et d'Amérique du Nord [40]. De plus, le plus grand groupe de parties prenantes de l'ISO est l'industrie [40]. Certains acteurs considèrent que cela permet à l'ISO de disposer d'une expertise industrielle supérieure à celle des ESOs qui serait bénéfique aux normes européennes [37]. D'autres, critiquent l'absence de représentation de certaines parties prenantes. De fait, les associations représentant les intérêts des consommateurs telles que l'ANEC, ainsi que celles représentant les travailleurs ou les petites entreprises, ne disposent pas officiellement du droit de participer aux travaux de l'ISO et de l'IEC. Par conséquent, elles n'ont pas de voix dans l'élaboration de ces normes internationales qui, à terme, seront adoptées comme normes européennes [16]. Si cette composition revêt une certaine importance dans le cadre général, l'enjeu est encore plus grand lorsqu'il s'agit des normes liées à l'éthique, afin de ne pas laisser l'industrie dicter les codes sociaux et éthiques.

3.3 La diversité des initiatives en dehors des instances européennes et internationales

En dehors du cadre bien défini des organismes de normalisation, certaines entités développent leurs propres initiatives pour normaliser ou certifier l'éthique de l'IA. C'est le cas des laboratoires nationaux de métrologie qui travaillent sur le développement de cadres d'évaluation pour l'IA, indépendemment des instances européennes et internationales de normalisation.

Laboratoires nationaux En France, le Laboratoire National de Métrologie et d'Essais (LNE), a développé son propre « Référentiel de certification de processus pour l'IA » [17]. Il définit un certain nombre d'exigences à res-

41. Voir le retour de DEKRA sur la stratégie de normalisation de la Commission européenne : https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13099-Standardisation-strategy/F2662668_en

42. https://www.anec.eu/

43. Traduction des auteurs.

pecter lors de la conception, le développement, l'évaluation et le maintien en conditions opérationnelles des systèmes d'IA. Le référentiel insiste notamment sur la transparence des processus, listant les éléments qui doivent être documentés et communiqués au client. Par ailleurs, il ne se base pas sur une technologie en particulier mais sur une manière de fonctionner au niveau de l'entreprise. S'il n'a pas une volonté particulière de traiter l'éthique, il témoigne néanmoins du souhait d'instaurer aux sein des entreprises d'IA un ensemble de bonnes pratiques. Il permet, en outre, la délivrance d'un certificat attestant que l'entreprise remplit bien les conditions indiquées. Quelques entreprises françaises ont déjà pu recevoir ce certificat [44]. Au sein du LNE, le laboratoire d'évaluation des systèmes d'IA [45] travaille actuellement sur d'autres normes pour encadrer les pratiques et auditer les systèmes. Le LNE pourrait, dans le cadre de l'*AI Act*, devenir un organisme notifié, chargé de vérifier la conformité. Sa façon d'aborder l'éthique est donc décisive pour les futurs cadres d'évaluation des systèmes d'IA.

Aux États-Unis, le *National Institute of Standards and Technology* (NIST) travaille également sur des normes pour l'IA [48]. Parmi ses travaux majeurs, les *Face Recognition Vendor Test* [46] posent les bases de l'évaluation des systèmes de reconnaissance faciale. Le NIST compare ainsi les performances de dizaines d'algorithmes provenant de différents fabriquants partout dans le monde [47]. Ces comparaisons sont effectuées sur différent critères de précision et, en ce qui concerne l'équité algorithmique, sur des mesures mathématiques d'écarts démographiques [27, 20]. Des métriques comme le taux de divergence ou le ratio de cas d'erreurs entre deux populations permettent ainsi de calculer la gravité d'un biais. Les tests réalisés par le NIST sont très suivis par les industriels et sont déterminants comme argument de vente auprès des clients. Les mesures qu'ils sélectionnent pour évaluer les systèmes ont donc de fortes chances de devenir le standard de référence du domaine. L'ISO pourrait notamment s'en inspirer dans le cadre de son sous-comité 37 sur la biométrie et de sa norme *ISO/IEC WD 19795-10* relative à la quantification de la variation des performances des systèmes biométriques dans les groupes démographiques [48].

Plus récemment, le NIST a publié un cadre de gestion des risques visant à *« améliorer la capacité d'intégrer des considérations de fiabilité dans la conception, le développement, l'utilisation et l'évaluation des produits, services et systèmes d'IA »* [49]. Il présente notamment les *« caractéristiques des systèmes d'IA dignes de confiance »*, ainsi que des actions pour assurer leur mise en pratique [47].

Associations de professionnels En plus de ces laboratoires nationaux, certaines associations de professionnels développent leurs propres référentiels. Parmi les initiatives internationales sur l'éthique de l'IA, la plus importante est sans doute celle de l'*Institute of Electrical and Electronics Engineers* (IEEE). En 2019, l'IEEE publie un document, reprenant une liste de principes éthiques et construisant un cadre pour les rendre opérationnels [12]. Le rapport présente notamment des recommandations et lignes directrices pour les *« normes, la certification, la réglementation ou la législation dans la conception, la fabrication et l'utilisation des systèmes [...] pour le bien-être social »* [50]. Cette étape est la première de la *IEEE Global Initiative* [51] sur l'éthique des systèmes autonomes et intelligents. Elle est suivie d'un développement massif de normes relatives à l'éthique des systèmes. La première à être publiée en 2021 est la norme *IEEE Std 7000™-2021* qui traite la prise en compte des préoccupations éthiques lors de la conception des systèmes [52]. D'autres normes suivront, plus spécialisée sur un aspect éthique, comme la norme *IEEE 7001™-2021* sur la transparence [53]. Certaines sont encore en développement, comme la norme *IEEE P7003™* sur les biais algorithmiques [36]. Au total, le projet prévoit le développement de plus d'une quinzaine de normes relatives à l'éthique de l'IA [54]. En plus de ces normes « éthiques », l'IEEE développe également un programme de certification pour l'éthique de l'IA, nommé *CertifAIEd* [55], basé sur la détermination d'un profil de risques du système d'IA puis d'une évaluation selon une série de critères.

D'autres initiatives méritent également d'être soulignées, comme le *Verband der Elektrotechnik* (VDE) en Allemagne, qui a développé son propre cadre pour la caractérisation de l'IA de confiance [50]. Il reprend des valeurs comme la transparence, la responsabilité, la vie privée, l'équité et la fiabilité, et leur attribue des critères et des indicateurs mesurables. Un score peut ensuite être calculé, représentant le degré d'éthique du système.

Entreprises privées Certaines entreprises développent aussi leurs propres cadres d'évaluation. Ainsi, Microsoft a publié l'année dernière sa norme sur l'IA responsable [39] dans laquelle sont listées diverses exigences relatives à la

44. C'est le cas par exemple de la société Axionable : https://www.axionable.com/ia-performante-et-ethique-axionable-decroche-la-1ere-certification-delivree-par-le-lne-2/

45. https://www.lne.fr/fr/actualites/leia-plateforme-inedite-evaluation-intelligence-artificielle

46. Tous les FRVT sont repertoriés sur le site du NIST : https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

47. Voir leur dernier classement : https://pages.nist.gov/frvt/html/frvt11.html

48. ISO/IEC WD 19795-10 Information technology — Biometric performance testing and reporting — Part 10 : Quantifying biometric system performance variation across demographic groups : https://www.iso.org/standard/81223.html

49. Traduction des auteurs

50. Traduction des auteurs

51. https://standards.ieee.org/industry-connections/ec/autonomous-systems/

52. IEEE Std 7000™-2021 : IEEE Standard Model Process for Addressing Ethical Concerns during System Design : https://standards.ieee.org/standard/7000-2021.html

53. IEEE Std 7001™-2021 : IEEE Standard for Transparency of Autonomous Systems : https://standards.ieee.org/ieee/7001/6929/

54. Ces normes sont recensées sur le site de l'IEEE : https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/

55. https://engagestandards.ieee.org/ieeecertified.html

responsabilité, la transparence, l'équité, la fiabilité, la vie privée et l'inclusivité. Des cadres plus larges sont parfois publiés sous la forme de politique d'entreprise en matière d'IA [56]. Néanmoins, ce ne sont souvent que de simples documents de recommandations et non de réels référentiels d'évaluation.

D'autres plus petites entreprises développent également leurs propres cadres d'évaluation des systèmes, mettant en place des labels d'IA éthiques, basés sur des questionnaires ou des audits algorithmiques. Notons par exemple en France, la société GoodAlgo qui propose un label nommé ADEL, pour évaluer l'éthique des systèmes d'IA basé sur le respect d'un certain nombre de critères « éthiques » [57]. Certains labels se centrent sur des critères plus précis, comme le label de « garantie humaine » pour l'IA en santé portée par le *Digital Medical Hub* et la société Ethik-IA [58], ou le label GEEIS IA pour l'égalité des chances [59]. Ces labels sont plus faciles et plus rapides à mettre en place que les processus de certification basé sur les normes harmonisées et permettent aux consommateurs d'orienter leurs choix en attendant un examen officiel. Néanmoins les critères d'évaluation des systèmes et de délivrance du certificat ne sont souvent pas publics.

La diversité de voix qui s'expriment peut favoriser la démocratisation de l'éthique, mais elle peut également provoquer une collision des différentes visions. L'analyse de ces voix permet de détecter des dynamiques et de visualiser les différentes orientations que peut prendre l'éthique de l'IA : une approche par la gouvernance et le management d'entreprise, mettant en place des bonnes pratiques pour la planification, la conception ou encore la supervision des systèmes, la gestion des risques et l'anticipation des défis sociaux ; ou une approche par la mesure et l'évaluation de la performance technique des systèmes.

4 Les normes et certifications « éthiques » soulèvent des risques

4.1 La frontière entre l'éthique et la technique est difficile à tracer

Le développement de ces normes « éthiques » ne fait pas l'unanimité. Selon [57], l'adjectif « éthique » se rapporte à la morale et ne peut donc s'appliquer qu'à *« une démarche, une délibération, une réflexion, une question, un principe, une valeur »*. Ainsi la notion de « conformité éthique » ou de « norme éthique » est contestable car l'éthique est variable par essence [57]. D'autres acteurs considèrent que la normalisation s'éloigne à tort des procédures scientifiques et techniques pour englober des questions sociales qui nécessitent un consensus politique [46]. Les normes sont en effet souvent utilisées pour la capture réglementaire [8] et certains vont jusqu'à accuser les organismes de normalisation européens de faire de la politique au service des normes [37].

Par ailleurs, la délimitation du caractère éthique d'une norme n'est pas toujours simple à évaluer. Actuellement, peu de normes abordent activement les défis sociaux [41]. Cependant, même les normes les plus techniques ont des implications sociales. Définir des critères d'équité ou de transparence des processus peut déjà être considéré comme un choix éthique. Selon [63], toutes les normes peuvent donc être considérées comme des normes éthiques implicites.

Notre recensement des initiatives de normalisation permet de distinguer deux types de normes. Les premières sont les normes « éthiques », ou « normes de gouvernance », qui discutent d'aspects sociaux ou des mécanismes à mettre en place pour effectivement respecter les principes éthiques. Ce sont souvent plutôt des normes autour des processus et du cycle de vie de l'IA, considérant la façon dont le système s'intègre dans un contexte social. Un exemple d'une telle norme est l'*IEEE Std 7000™-2021* [60]. Elle inclut notamment la prise en compte des différentes parties prenantes et de leurs valeurs durant les phases d'exploration et de développement des systèmes. La deuxième catégorie de normes correspond aux normes « éthiques implicites », ou « normes de mesure », qui répondent à des principes éthiques par des critères techniques. Par exemple, la norme *ISO/IEC TR 24027 :2021* [61] donne de nombreuses définitions mathématiques de l'équité. Même si elle ne prend pas parti quant à la meilleure mesure à adopter, cette norme comporte inévitablement un aspect éthique en raison du sujet qu'elle aborde. Dans les deux cas, les normes ne dictent pas les résultats attendus d'un système : il n'existe pas pour l'instant de « norme de performance ». Les normes actuelles mettent en lumière des bonnes pratiques, que ce soit dans l'organisation des procédures entourant le système, ou les méthodes techniques de conception et d'évaluation. Dans ce sens, elles ne remplacent pas l'avis d'un juge en décidant de ce qui serait ou non acceptable en matière de droits humains. Leur utilisation n'est pourtant pas neutre par rapport au respect de ces droits.

4.2 La difficile définition de critères objectifs

L'identité de l'Europe s'est construite sur la définition d'un ensemble de valeurs démocratiques et économiques qui, dans le cas du numérique, s'allient aux critères techniques et complexifient leur application [34]. Ainsi, les normes techniques ont du mal à s'adapter aux aspects éthiques et sociaux auxquels sont confrontés les systèmes. Puisque l'*AI Act* délègue au processus de normalisation une grande partie des exigences quant à l'IA éthique, savoir quels garde-fous seront mis en place, et par qui, devient un enjeu de gouvernance et de démocratie.

Cet enjeu est renforcé par l'aspect contextuel et culturel de

56. Voir par exemple celle de la société SAP : https://www.sap.com/documents/2022/01/a8431b91-117e-0010-bca6-c68f7e60039b.html

57. https://goodalgo.fr/labels-ethiquement-engages/

58. http://esante.gouv.fr/agenda/lancement-du-premier-label-de-garantie-humaine-de-lintelligence-artificielle

59. https://arborus.org/label/

60. Voir supra note 52

61. Voir supra note 38

l'éthique : il n'y a pas une unique façon de l'aborder ou de la définir. Les décisions prises lors des processus de normalisation et de certification ne sont pas des choix neutres, mais plutôt le reflet de valeurs ayant des répercussion sociales. Les « bons » choix à faire ne sont pas universels. Il existe pourtant une volonté de converger vers des valeurs et principes communs, comme en témoignent les nombreux documents sur l'éthique de l'IA publiés au niveau européen [49] et international [59]. Toutefois, dans le cadre de la normalisation, ces choix sont rarement présentés comme des choix de valeurs, mais comme des choix purement techniques et donc « neutres ». Par exemple, l'adhésion aux normes harmonisées est supposée fournir un moyen « objectivement vérifiable » de se conformer aux exigences essentielles [51]. La certification « éthique » de l'IA se confronte alors aux mêmes obstacles que d'autres domaines comme le commerce équitable ou les normes environnementales, pour lesquels les labels et certifications servent de marque de qualité et déterminent implicitement les valeurs morales et sociales à suivre [42].

Les différentes lignes directrices et chartes en matière d'éthique édictées par les instances internationales ont déjà été critiquées pour le prisme qu'elles adoptent dans leur vision de l'éthique. Toutefois, cette critique revêt une importance accrue dans le cas des normes et des certifications qui non seulement codifient les principes éthiques, mais les figent également dans nos pratiques. La légitimité des organismes de normalisation quant à l'imposition de cette vision de l'éthique est donc remise en question.

De plus, l'éthique de l'IA est dynamique et un système ne saurait être perpétuellement aligné sur des valeurs éthiques [8]. En cela, une certification statique en matière d'éthique de l'IA ne peut saisir aucun enjeu pertinent. Une IA certifiée « éthique », le restera alors même que le système et son contexte évoluent. Bien que certains organismes souhaitent réévaluer régulièrement le système et garder un certificat à jour [62], une telle marque ne pourra jamais être totalement adaptée à un contexte spécifique, sinon elle perdrait son caractère universellement applicable.

En plus d'instituer arbitrairement ces valeurs, de nombreuses entités se contentent d'adopter des normes de façon symbolique, sans changer leurs pratiques [14]. Le respect de ces normes devient alors un simple argument de vente et l'obtention d'une marque ou d'un label « éthique » ne fait que légitimer ces pratiques, et perpétue une culture de l'*ethics washing* [61]. Certains craignent alors que les normes développées au niveau européen ne soient que trop peu contraignantes, laissant ainsi l'industrie agir à sa guise [51].

Les choix effectués pour aboutir à une norme ou un référentiel d'évaluation impactent fortement l'évolution des systèmes d'IA. Par exemple, la diffusion généralisée du cadre de gestion des risques du NIST [47] sur les réseaux et par diverses instances pourrait favoriser son adoption à grande échelle, préemptant de fait l'émergence de référentiels spécifiquement européens. Bien que le référentiel NIST puisse être compatible avec la vision de l'*AI Act* concernant l'IA « de confiance », le fait que cela soit une norme américaine constitue un obstacle pour l'Europe dans sa volonté de définir sa propre vision.

La question de l'équité illustre les difficultés à converger sur une vision commune de l'éthique. En effet, de nombreuses définitions techniques de l'équité existent [43], et sont même souvent contradictoires [35]. Le choix des approches mathématiques aura un impact sur les droits des personnes. Un système équitable selon une définition ne l'est pas forcément selon une autre. C'est le cas du logiciel COMPAS utilisé aux Etats-Unis pour prédire le taux de récidive de criminels, qui a été accusé de pénaliser les personnes afro-américaines selon un certain critère d'équité [5], alors qu'il respectait l'équité selon une autre méthode de mesure [45]. Cette diversité des mesures d'équité risque de mener à des choix de simplification stratégique [3] : les constructeurs affichent simplement la mesure de l'équité qui montre que leur système est exempt de biais et donc « juste » selon eux, et pas les autres mesures.

Instaurer une unique mesure, ou un ensemble limité de mesures dans les normes risquent d'accentuer cette tendance. De plus, ces choix sont intégrés dans les systèmes d'IA sans véritable débat public.

4.3 La certification comme garantie de l'éthique

La certification vis-à-vis des normes techniques est souvent perçue comme une garantie de sécurité [18]. C'est le cas notamment du marquage CE, compris à tort comme un gage de qualité alors même qu'il ne signifie que la conformité à la réglementation. De plus, il est généralement apposé par le fabriquant lui-même. Or un produit marqué CE peut également avoir des failles de sécurité [62].

De la même façon que pour le marquage CE et les normes de sécurité, il est probable qu'une certification vis-à-vis de normes « éthiques » soit également considérée par le consommateur comme une garantie de l'éthique du produit ou procédé. Pourtant, respecter une notion mathématique de l'équité telle que définie dans une norme ne garantit pas que le système ne discriminera pas. De même, respecter une norme de conception éthique qui donnerait pour critère la réalisation d'une étude d'impacts permet de mesurer certaines conséquences du déploiement d'un système, mais cela ne signifie pas que tous les préjudices possibles ont été pris en compte, ou que les mesures de protection prises sont suffisantes. La marque de certification pourrait alors induire chez les utilisateurs des systèmes et des citoyens un faux sentiment de protection.

Les normes doivent alors être co-construites avec un régime de responsabilité [38]. Toutefois, cette responsabilité est plus dure à prouver une fois les produits et processus certifiés : lorsque la conformité est démontrée, il est difficile de s'y opposer. Certains considèrent qu'il y a une « juridification » du processus de normalisation [58]. Depuis l'affaire James Elliott Construction [60], les normes harmonisées peuvent même être considérées comme des dispositions du droit de l'Union européenne.

62. C'est le cas notamment du référentiel du LNE [17]

Utiliser des normes techniques pour assurer le respect de droits fondamentaux est risqué, car les normes rentreraient alors en concurrence avec le travail du législateur et des juges, qui sont les seuls compétents à évaluer, et mettre en équilibre, les ingérences dans différents droits fondamentaux [26]. Une norme technique visant à protéger la sécurité des personnes vise également à préserver un droit fondamental, notamment le droit à la vie. Mais une norme technique sur la résistance au feu, par exemple, rentrera moins en conflit avec le rôle premier des juges. Certes, un juge pourra toujours estimer que l'application d'une norme de sécurité dans un cas précis était insuffisant. Mais la norme technique en matière de sécurité ne sera pas perçue par les juges comme une ingérence dans leur travail, alors qu'une norme en matière de respect des droits fondamentaux pourrait l'être.

5 Conclusion

Notre cartographie des acteurs et initiatives offre une grille de lecture pour mieux comprendre les enjeux de la normalisation et de la certification de l'éthique de l'IA. Nous pensons qu'elle peut servir à la fois aux dirigeants d'entreprises, aux chercheurs et aux développeurs de systèmes d'IA dont les pratiques vont être impactées par ces normes, ainsi qu'aux juristes, sociologues et philosophes s'intéressant à l'évolution des représentations de l'éthique normative.

Nous avons vu que de plus en plus d'acteurs se lancent dans la mise en place de normes ou de processus de certification pour l'éthique de l'IA, alors que le concept même d'une norme technique pour l'éthique fait débat. Nous avons dressé un panorama de leurs activités et avons cherché à identifier les tensions qui peuvent émerger entre les différents organismes. Cet écosystème est notamment caractérisé par une rivalité de gouvernance opposant d'un côté les normes européennes, et de l'autre les normes internationales. L'émergence de cadres normatifs en dehors de ces structures officielles permet de multiplier les garde-fous de façon plus rapide que les processus habituels, mais risquent également de dévoyer l'éthique en diffusant des critères d'évaluation manquant toute légitimité démocratique. Toutefois, il est essentiel de garder à l'esprit que l'éthique est intrinsèquement contextuelle et, dans une certaine mesure, subjective. Une certification ne pourra garantir le respect absolu de principes éthiques et de droits individuels. La question de savoir qui est responsable du développement de ces cadres d'évaluation est cruciale, car ils façonneront le développement des futurs systèmes, avec parfois un impact mondial. Afin de ne pas donner aux utilisateurs et aux citoyens la fausse impression que leurs droits sont nécessairement préservés grâce au marquage CE, les limites de cette certification doivent toujours être mises en avant. Même si le marquage CE témoigne d'une certaine démarche de qualité, pour identifier et réduire les biais par exemple, cette démarche ne change rien en ce qui concerne la responsabilité de l'opérateur pour une décision algorithmique discriminatoire. Ainsi, la certification ne fonctionnera que si elle est accompagnée d'un régime de responsabilité et de procédures de recours en cas de préjudice.

Enfin, notons que si les initiatives visant à normaliser et à certifier l'éthique cherchent à prévenir les dérives potentielles, les mesures de protection précises qu'elles définiront restent encore indéterminées. En effet, la plupart des normes dans ce domaine ne sont pas encore développées, et celles qui le sont, ne sont pas encore opérationnelles [22]. Enfin, l'interconnexion entre les instances de normalisation techniques et les instances étudiant le respect par l'IA des droits fondamentaux [63], semble indispensable.

Remerciements

Merci à Tiphaine Viard pour ses commentaires précieux.
Cette recherche a été financée dans le cadre du projet LIMPID [64] (Projet ANR 20-CE23-0028).

Références

[1] Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation.

[2] AFNOR. Parler normes couramment. L'essentiel, 2014.

[3] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing : the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

[4] Robert H Allen and Ram D Sriram. The Role of Standards in Innovation. *Technological Forecasting and Social Change*, 64(2) :171–181, 2000.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016.

[6] Knut Blind. The impact of standardisation and standards on innovation. In *Handbook of Innovation Policy Impact*, pages 423–449. 2016.

[7] Joanna Bryson and Alan Winfield. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer*, 50(5) :116–119, 2017.

[8] Joanna J Bryson. Belgian and Flemish Policy Makers' Guide to AI Regulation. 2022.

[9] Céline Castets-Renard and Philippe Besse. Ex ante Accountability of the AI Act : Between Certification and Standardization, in Pursuit of Fundamental Rights in the Country of Compliance, 2022.

[10] Tom Cellucci. Developing Operational Requirements - A Guide to the Cost-Effective and Efficient Communication of Needs. Technical report, U.S. Department of Homeland Security, 2008.

[11] CEN-CENELEC. CEN-CENELEC Focus Group Report : Road Map on Artificial Intelligence (AI), 2020.

63. Voir par exemple le Comité sur l'Intelligence Artificiel (CAI) du Conseil de l'Europe : `https://www.coe.int/fr/web/artificial-intelligence/cai`

64. `https://limpid.telecom-paris.fr/`

[12] Raja Chatila and John C. Havens. Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Version 2. Technical report, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019.

[13] Julien Chiaroni and Arno Pons. IA de confiance - Opportunité stratégique pour une souveraineté industrielle et numérique. Technical report, Digital New Deal, 2022.

[14] Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. AI Certification : Advancing Ethical Practice by Reducing Information Asymmetries. *IEEE Transactions on Technology and Society*, 2(4) :200–209, 2021.

[15] European Commission. Commission Notice - The 'Blue Guide' on the implementation of EU product rules 2022. Technical report, Information from European Union Institutions, Bodies, Offices and Agencies, 2022.

[16] Pierluigi Cuccuru. Interest Representation in European Standardisation : The Case of CEN and CENELEC, 2019.

[17] Laboratoire National de Métrologie et d'Essais (LNE). Référentiel de certification de processus pour l'IA - Conception, développement, évaluation et maintien en conditions opérationnelles, 2021.

[18] Av de Tervueren. ANEC Position Paper on CE marking "Caveat Emptor - Buyer Beware", 2012.

[19] Henk J. de Vries. *Standardization : A Business Approach to the Role of National Standardization Organizations*. Springer US, Boston, MA, 1999.

[20] David L Duewer. Face Recognition Vendor Test (FRVT) Part 8 : Summarizing Demographic Differentials. Technical Report NIST IR 8429, National Institute of Standards and Technology (NIST), 2022.

[21] European Trade Union Confederation (ETUC). Feedback on the (roadmap) consultation of citizens and stakeholders on the forthcoming "EU Standardisation strategy", 2021.

[22] European Commission. Joint Research Centre. AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework. Technical report, Publications Office, LU, 2021.

[23] Commission Européenne. Proposition de règlement du Parlement Européenn et du Conseil établissant des règles harmonisées concernant l'Intelligence Artificielle (législation sur l'Intelligence Artificielle) et modifiant certains actes législatifs de l'Union, 2021.

[24] European Committee for Electrotechnical Standardization (CENELEC). CENELEC Guide 13 FAQ. Frequently Asked Questions on the Frankfurt Agreement. Edition 1, 2017.

[25] DIN German Institute for Standardization e. V. *Economic benefits of standardization - Summary of results*. Beuth Verlag, DE, 2000.

[26] Mélanie Gornet and Winston Maxwell. Intelligence artificielle : normes techniques et droits fondamentaux, un mélange risqué. *The Conversation*, 2022.

[27] Patrick J. Grother, Mei L. Ngan, and Kayee K. Hanaoka. Face Recognition Vendor Test (FRVT) Part 3 : Demographic Effects. Technical Report NIST IR 8280, National Institute of Standards and Technology (NIST), 2019.

[28] Laurens Hernalsteen and Constant Kohler. Drafting Harmonized Standards in support of the Artificial Intelligence Act (AIA) - CEN-CENELEC, 2022.

[29] Kashmir Hill. Wrongfully Accused by an Algorithm. *The New York Times*, 2020.

[30] Marion Ho-Dac. La normalisation, clé de voûte de la réglementation européenne de l'intelligence artificielle (AI Act). *Dalloz IP/IT*, pages 228–236, 2023.

[31] Japan Business Council in Europe (JBCE). The Roadmap of European Standardisation Strategy, 2021.

[32] ISO and CEN. Foire aux questions relatives à l'Accord de Vienne, 2016.

[33] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9) :389–399, 2019.

[34] Jonathan Keller and Claire Levallois-Barth. La fragile définition de l'identité européenne par ses valeurs numériques. *Revue générale du droit. Chronique de droit de l'Union*, 2021.

[35] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv :1609.05807*, 2016.

[36] Ansgar Koene, Liz Dowthwaite, and Suchana Seth. IEEE P7003™ standard for algorithmic bias considerations : work in progress paper. In *Proceedings of the International Workshop on Software Fairness*, pages 38–41. ACM, 2018.

[37] Mark McFadden, Kate Jones, Emily Taylor, and Georgia Osborn. Harmonising Artificial Intelligence : The role of standards in the EU AI Regulation. 2021.

[38] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic Impact Assessments and Accountability : The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 735–746, New York, NY, USA, 2021. Association for Computing Machinery.

[39] Microsoft. Microsoft Responsible AI Standard, v2, 2022.

[40] Mari Morikawa and Jason Morrison. A Survey of Participation in ISO's International Standards Development Processes. 2004.

[41] Markus Mueck, Scott Cadzow, Cadzow Communications, and Suno Wood. ETSI White Paper No. #52. ETSI Activities in the field of Artificial Intelligence Preparing the implementation of the European AI Act. 1st Edition. 2022.

[42] Warwick Murray, Overton John, and Howson Kelle. *Ethical Value Networks in International Trade : Social Justice, Sustainability and Provenance in the Global South*. Edward Elgar Publishing, 2022.

[43] Arvind Narayanan. Tutorial : 21 fairness definitions and their politics, 2019.

[44] Executive Board Netherlands Standardization Institute (NEN). Statement on the European Commission's Roadmap for a Standardization Strategy, 2021.

[45] Northpointe. Practitioner's Guide to COMPAS Core. Technical report, 2019.

[46] European Council of Engineers Chambers. Statement on "Roadmap Standardization Strategy", 2021.

[47] National Institute of Standards and Technology (NIST). AI Risk Management Framework : AI RMF (1.0). Technical Report NIST AI 100-1, 2023.

[48] National Institute of Standards and Technology (NIST). NIST AI Program. Artificial Intelligence : The Vitals, 2023.

[49] Independent Hight-Level Expert Group on Artificial Intelligence set up by the European Commission (HLEG). Ethics guidelines for trustworthy AI. Technical report, European Commission, 2019.

[50] Christoph Peylo, Dirk Slama, Sebastian Hallensleben, Andreas Hauschke, and Stephanie Hildebrandt. VCIO based description of systems for AI trustworthiness characterisation. Technical Report VDE SPEC 90012 V1.0 (en), Verband der Elektrotechnik (VDE), 2022.

[51] Hadrien Pouget. The EU's AI Act Is Barreling Toward AI Standards That Do Not Exist. *Lawfare*, 2023.

[52] Thomas Reardon, Jean-Marie Codron, Lawrence Busch, R. James Bingen, and Craig Harris. Global change in agrifood grades and standards : agribusiness strategic responses in developing countries. *International Food and Agribusiness Management Review*, 02(3-4), 1999.

[53] Reuters. Amazon ditched AI recruiting tool that favored men for technical jobs. *The Guardian*, 2018.

[54] Stephen Russell. Roadmap for the standardisation strategy. ANEC response. 2021.

[55] Björn ten Seldam and Alex Brenninkmeijer. The Dutch benefits scandal : a cautionary tale for algorithmic enforcement. *EU Law Enforcement*, 2021.

[56] Catherine Tessier. Ethique et IA : analyse et discussion. In *Conférence Nationale en Intelligence Artificielle (CNIA)*, 2021.

[57] Catherine Tessier. Parler du numérique et de son éthique : un questionnement... éthique. In *Pour une éthique du numérique*, pages 97–105. Puf edition, 2022.

[58] Carlo Tovo. Judicial review of harmonized standards : Changing the paradigms of legality and legitimacy of private rulemaking under EU law. *Common Market Law Review*, 55(4), 2018.

[59] United Nations Educational Scientific and Cultural Organization (UNESCO). Recommendation on the Ethics of Artificial Intelligence. Technical report, 2021.

[60] Arnaud van Waeyenberge and David Restrepo. James Elliot construction : A "New(ish) approach" to judicial review of standardisation. *European Law Review*, 42 :882–893, 2017.

[61] Ben Wagner. Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping ? In *BEING PROFILED :COGITAS ERGO SUM*. Amsterdam University Press, 2018.

[62] I. M. E. Wentholt, J. B. L. Hoekstra, A. Zwart, and J. H. DeVries. Pendra goes Dutch : lessons for the CE mark in Europe. *Diabetologia*, 48(6) :1055–1058, 2005.

[63] Alan Winfield. Ethical standards in robotics and AI. *Nature Electronics*, 2(2) :46–48, 2019.

[64] Wolfgang Ziegler. A Landscape Analysis of Standardisation in the Field of Artificial Intelligence. *Journal of ICT Standardization*, pages 151–184, 2020.

Session 5 : Explicabilité et équité 2

Une revue systématique de la littérature autour du biais, de l'équité et de l'explicabilité

M.L. Ndao[1,2], G. Youness[1,2], N. Niang[2], G. Saporta[2]

[1] Laboratoire LINEACT CESI, Nanterre, IDFC
[2] Laboratoire Cedric-MSDMA, Paris, France

mlndao@cesi.fr ; gyouness@cesi.fr ; ndeye.niang_keita@cnam.fr ; gilbert.saporta@cnam.fr

Résumé

Ce travail fournit une analyse d'une bibliographie autour du biais de l'équité et de l'explicabilité des algorithmes de l'IA entre 2015 et 2022. Par une approche de Traitement Automatique du Langage Naturel, plus précisément la LDA, nous avons extrait 8 sujets traités par cette bibliographie. Une analyse de la popularité de ces sujets a permis de constater une évolution plus rapide du nombre et du pourcentage des publications traitant surtout l'explicabilité et l'équité dans les algorithmes de l'IA. Une comparaison a permis de noter une similarité entre nos résultats et ceux fournis par BERTopic.

Mots-clés

Intelligence Artificielle Explicable (XAI), biais, équité, Traitement Automatique du Langage Naturel (TAL), Latent Dirichlet Allocation (LDA)

Abstract

This work provides an analysis of a bibliography concerning the bias, fairness and explainability of AI algorithms between 2015 and 2022. Using a Natural Language Processing approach, specifically LDA, we extracted 8 topics covered by this bibliography. An analysis of the frequency of these topics showed a faster increase in the number and proportion of publications dealing mainly with explainability and fairness in AI algorithms. A comparison revealed a similarity between our results and those provided by BERTopic.

Keywords

Explainable Artificial Intelligence (XAI), bias, fairness, Natural Language Processing (NLP), Latent Dirichlet Allocation (LDA)

Introduction

Dans un contexte marqué par l'utilisation massive des algorithmes d'apprentissage automatique en Intelligence Artificielle (IA) dans les processus de prise de décision dans presque tous les domaines (finance [31], recommandation [38], santé [26] , etc.) un besoin de confiance en ces algorithmes se pose. Selon Alain Mille et al. (2020) [28], « Nous sommes dans un contexte où les algorithmes de l'IA, initialement destinés à automatiser des tâches mécaniques, s'intéressent à des fonctions cognitives que l'on pensait hors champs de l'automatisation. Ayant eu le statut d'objet de recherche à partir de 1956 (conférence de Dartmouth), l'IA intervient, aujourd'hui, à tous les niveaux de la vie. ». Cependant, de nombreux incidents ont démontré des failles dans ces algorithmes qui sont souvent source de discrimination dans plusieurs domaines comme en reconnaissance faciale, en justice, en recommandation, en recrutement, en banque, en santé, etc. (Google photo [1], COMPAS [2], logiciel de recrutement chez Amazon [3]).

La plupart des algorithmes d'apprentissage automatique (Machine Learning ML) se basent sur des données d'apprentissage susceptibles de contenir un biais : par exemple, une sous représentation d'un groupe d'individus. Ainsi, ce biais pourrait être reconduit dans les prédictions issues de ces algorithmes. Le cas du logiciel de recrutement d'Amazon peut être expliqué par le fait que l'algorithme s'est basé sur les CV collectés depuis plusieurs années et composés essentiellement de CV d'hommes. De surcroît, ces CV ont été sélectionnés par des humains et sont susceptibles d'avoir été choisis de façon biaisée.

Le développement sans précédent des algorithmes de ML dans presque tous les domaines en termes de prise de décisions est conjugué à des failles en termes de biais discriminatoires (non représentativité d'un groupe d'individus comme l'exemple des données du logiciel d'Amazon), d'équité (décision défavorisant un groupe d'individus) et de manque de compréhension des modèles (explicabilité). Cela a provoqué une vague de recommandations de la part de certains organismes tels que la DARPA (Defense Advanced Research Projects Agency) et à l'annonce de l'XAI (eXplainable Artificial Intelligent) en 2016 (D. Gunning et al. 2019) [15].

Depuis cette annonce, on note une forte multiplication des recherches et publications sur l'équité, l'explicabilité et le biais des algorithmes de l'IA. C'est ce qu'on observe en analysant les données de Google Trends sur les tendances de recherches des termes « explainable XAI », « Bias XAI » et « Fairness XAI » (FIGURE 1).

1. https://www.dailymail.co.uk/sciencetech/article
2. ProPublica. 23 mai 2016 ajouter l'article dans ref
3. https://www.assessfirst.com/fr/algorithme-sexiste-amazon/

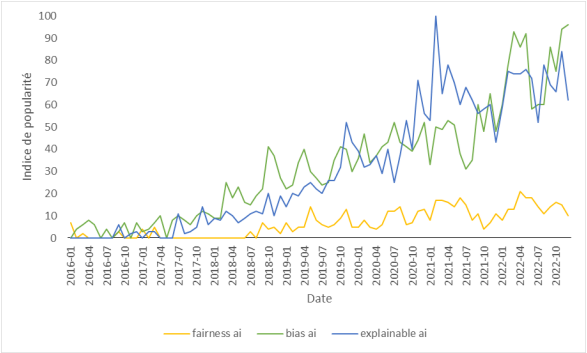


FIGURE 1 – Les tendances de recherches des termes « explainable XAI », « Bias XAI » et « Fairness XAI » dans le monde depuis 2016 selon Google Trends.

Aujourd'hui, une des problématiques autour de la littérature du biais, de l'explicabilité et de l'équité est le nombre important de propositions de modèles d'XAI et de métriques d'équité (plus d'une dizaine de métriques dont certaines sont contradictoires (Mitchell et al., 2021 [30]). Ainsi, un besoin de positionnement des unes par rapport aux autres sur ces différentes propositions se pose. Cela faciliterait l'encadrement de l'utilisation de ces algorithmes afin d'éviter les incidents discriminatoires.

Dans le cadre du biais, nous recensons également un grand nombre de propositions de typologies. Par exemple, Mehrabi et al. (2021)[27] propose une typologie du biais en 3 groupes : des données vers l'algorithme ; de l'algorithme vers les utilisateurs et des utilisateurs vers les données. Tandis que dans Bertail et al. (2019) [2], on retrouve les 3 types de biais suivants : le biais cognitif, le biais statistique et le biais économique. Choisir une typologie du biais parmi les différentes propositions peut être subjectif. Une réorganisation et une recherche de la structure sous-jacente de la bibliographie de l'explicabilité, du biais et de l'équité en IA est nécessaire. C'est l'objectif de ce travail.

Nous proposons une analyse de la structure sous-jacente de la bibliographie autour du biais, de l'équité et de XAI à l'aide d'une approche de Traitement Automatique du Langage Naturel (Natural Language Processing ou NLP), plus précisément le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [3]). D'une part, il s'agit d'identifier les thèmes ou sujets majeurs traités par un ensemble d'articles publiés entre 2015 et 2022. D'autre part, une analyse fine des résultats obtenus aidera à la réorganisation des publications permettant de proposer un état de l'art sur la bibliographie autour du biais de l'équité et de XAI. Enfin, une comparaison de nos résultats et ceux fournis par BERTopic sera effectuée dans le cadre de validation de nos résultats.

La suite du papier est organisée comme suit : la première section est consacrée à une brève présentation des travaux antérieurs qui ont utilisé la LDA pour synthétiser un ensemble d'archives, ainsi qu'à la présentation de l'approche LDA. Ensuite, la section 2 est d'abord dédiée à la présentation de l'ensemble de notre démarche allant de la collecte des données à la modélisation. Par la suite, la deuxième partie de cette section portera sur l'analyse et la discussion des résultats obtenus et leur comparaison avec BERTopic.

1 Topic modeling

1.1 Travaux antérieurs

Le NLP, plus particulièrement le 'topic modeling', est souvent utilisé dans différents domaines selon divers contextes afin de synthétiser, d'organiser ou d'analyser des collections de documents ou d'archives. C'est une approche pertinente dans un contexte de données massives ou big data. Bernadeta et al. 2023 [13] se sont basés sur cette approche pour proposer une analyse synthétique des journaux à propos de la COVID-19 en Suède. Il s'agit d'une description de 6515 articles de journaux publiés entre janvier 2020 et mars 2021. En utilisant l'approche LDA, ils ont pu découvrir les différents sujets traités par ces journaux ainsi que leur évolution dans le temps.

Dans ce même contexte qui est celui la pandémie de la COVID-19, Eren et al. 2020 [9], conscients de la montée rapide du nombre de publications sur la COVID-19, proposent une analyse des archives de la base de données CORD-19 [50]. À ce propos, ils ont utilisé la LDA afin de découvrir la structure en groupes de l'ensemble de ces publications qu'ils visualisent ensuite. Cette étude constitue ainsi une réorganisation des thèmes abordés dans les publications sur la COVID-19 aux USA.

En journalisme, Jacobie at al., 2018 [17] ont également utilisé l'approche LDA pour analyser l'ensemble des publications de The New York Times portant sur la technologie du nucléaire depuis 1945. Cette étude a également prouvé la pertinence de l'approche LDA dans la recherche des sujets sous-jacents à une collection de documents.

Dans le domaine de la maintenance prédictive, Kamal et al. (2021) [33] ont proposé une analyse principalement descriptive de l'ensemble des publications sur l'XAI et la maintenance prédictive entre 2015 et 2021. Il s'agit d'un état de l'art des publications dans ce domaine qui a également permis d'avoir une vue générale, une réorganisation de la bibliographie, mais également une comparaison entre l'explicabilité et la performance des modèles en maintenance prédictive.

Parmi les approches de topic modeling, il y a les approches classiques comme LDA, mais également des approches qui se basent sur des réseaux de neurones comme BERTopic qui est issu de BERT [7] (Bidirectional Encoder Representations from Transformers). BERT est un modèle profond de représentation bidirectionnelle non supervisé du langage développé par Google et qui a donné de bons résultats dans l'extraction de sujets. Grootendorst (2022) [14] s'est basé sur ce modèle pour proposer BERTopic. C'est une extension de BERT en Topic modeling qui se base sur une variation du TF-IDF pour extraire les sujets pertinents. Ce dernier a fourni de bons résultats dans ce domaine [34, 42]. Ainsi, il peut être considéré comme une référence permettant de valider nos résultats. Dans ce travail, nous utiliserons principalement l'approche LDA. BERTopic sera utilisé

dans le cadre de la validation de nos résultats.

1.2 L'approche LDA

Le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [3]) est l'une des techniques de NLP non supervisées les plus connues qui cherchent à découvrir des thématiques ou sujets cachés dans un ensemble de M documents appelé corpus noté D. C'est un modèle probabiliste génératif permettant de trouver la structure sous-jacente d'un ensemble de documents en termes de sujets. La LDA considère le corpus comme un mélange de K sujets décrits chacun par un ensemble de mots auxquels sont associés une probabilité.

L'ensemble des M documents ou encore corpus est représenté par une matrice dite document-mots, souvent creuse, notée $D_{M,N}$ de dimension (M, N) où la cellule (D_i, w_j) correspond à la fréquence du mot w_j dans le document D_i, par exemple :

$$D_{M,N} = \begin{array}{c} \\ D_1 \\ \vdots \\ D_i \\ \vdots \\ D_M \end{array} \begin{array}{c} \begin{array}{cccccc} w_1 & \cdots & w_j & \cdots & w_N \end{array} \\ \begin{bmatrix} 0.3 & & 0 & & 0.2 \\ \cdots & & \cdots & & \\ 0.1 & & 0 & \vdots & 0 \\ & \cdots & & & \vdots \\ 0 & \cdots & 0 & \cdots & 0.01 \end{bmatrix} \end{array}$$

Le nombre de sujets K est choisi *a priori* ou au regard d'un indicateur comme le score de cohérence que l'on définira dans la section suivante.

Partant de $D_{M,N}$, la LDA estime les matrices $\theta_{M,K}$ (documents-sujets) et $\phi_{K,N}$ (sujets-mots) , par une approche itérative.

Dans la matrice $\theta_{M,K}$, $\theta_{m,k}$ correspond à la probabilité que le sujet z_k soit traité dans le document D_m ($\theta_{i\cdot} = \sum_{k=1}^{K} \theta_{ik}$=1).

Ces probabilités sont initialisées par une distribution de Dirichlet de paramètre α ($Dir(\alpha)$). Le résultat est une classification en K clusters où chaque cluster correspond à un sujet. Nous utilisons dans la suite les deux termes sujet ou cluster indifféremment. À partir de $\theta_{M,K}$, on retrouve une partition des documents en K clusters, en affectant chaque document au sujet pour lequel sa probabilité d'appartenance est maximale.

La matrice $\phi_{K,N}$ correspond à la matrice sujets-mots, où ϕ_{kj} correspond à la probabilité que le mot w_j soit dans le sujet z_k. Chaque sujet z_k est décrit par les n mots ayant les plus fortes probabilités ϕ_{kj}, nous les notons $(w_j^k)_{1\leq j\leq n}$. La matrice $\phi_{K,N}$ est initialisée par une distribution de Dirichlet $Dir(\beta)$. Des exemples de matrices $\theta_{M,K}$ et $\phi_{K,N}$ sont données ci-après :

$$\theta_{M,K} = \begin{array}{c} \\ D_1 \\ \vdots \\ D_i \\ \vdots \\ D_M \end{array} \begin{array}{c} \begin{array}{ccc} z_1 & \cdots & z_K \end{array} \\ \begin{bmatrix} 0 & & 0.2 \\ \vdots & & \vdots \\ 0.1 & & 0.5 \\ & \vdots & \\ 0.6 & \cdots & 0.0 \end{bmatrix} \end{array}$$

$$\phi_{K,N} = \begin{array}{c} \\ z_1 \\ \vdots \\ z_K \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & \cdots & w_j & \cdots & w_N \end{array} \\ \begin{bmatrix} 0 & & 0 & & 0.3 \\ \cdots & & \cdots & & \\ 0.1 & & 0 & \vdots & 0 \end{bmatrix} \end{array}$$

1.3 Évaluation

Pour évaluer la cohérence quantitative de nos résultats, nous avons utilisé deux métriques appelées scores de cohérence : $UMASS$ (Unnormalized Measures of Association Strength) (Mimno et al., 2011 [29]) et C_V (Coherence Value) score (Röder et al., 2015 [40]). Il s'agit d'indicateurs qui évaluent le degré de similitude sémantique entre les mots les mieux notés dans les sujets en moyenne. Ces deux scores de cohérence sont de bons indicateurs permettant d'évaluer la qualité sémantique des résultats de topic modeling comme LDA ou BERTopic [40]. Le calcul de ces scores de cohérence se base sur la co-occurrence des mots dans l'ensemble des documents et dans chaque sujet et l'information mutuelle.

Par exemple, considérant le sujet z_k et $\epsilon > 0$, le score $UMASS$ est donné par :

$$C_{UMASS} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{F(w_i^k, w_j^k) + \epsilon}{F(w_j^k)} \quad (1)$$

Ici, $F(w_j^k)$ est le nombre de fois que le mot w_j^k est apparu au moins une fois dans un document et $F(w_i^k, w_j^k)$ le nombre de fois que les w_i^k et w_j^k ont été observés à la fois au sein d'un même document. Le calcul du score C_V se base en plus sur l'information mutuelle ponctuelle normalisée ($NMPI$) donnée par :

$$NMPI(w_i^k, w_j^k) = \frac{log \frac{F(w_i^k, w_j^k)+\epsilon}{F(w_j^k)}}{-log(F(w_i^k, w_j^k + \epsilon))} \quad (2)$$

Plus ces scores sont élevés, meilleure est la qualité des sujets en termes de cohérence. Röder et al., 2015 [40] ont mené une étude comparative des principaux scores de cohérence en les comparant également à l'appréciation humaine. Selon cette étude C_V et $UMASS$ apparaissent comme les meilleures métriques d'évaluation de cohérence.

Dans cette analyse, on utilisera C_V score pour choisir le nombre de sujets. Nous utiliseront les deux métriques pour comparer nos résultats à ceux de BERTopic.

2 Application

Dans cette section, nous commencerons par expliquer notre processus de modélisation depuis la collecte des données. Ensuite, nous présenterons les résultats obtenus à l'issue de cette analyse.

2.1 Processus d'analyse

Dans le cadre de la modélisation, le processus suivant a été suivi (voir FIGURE 2) :

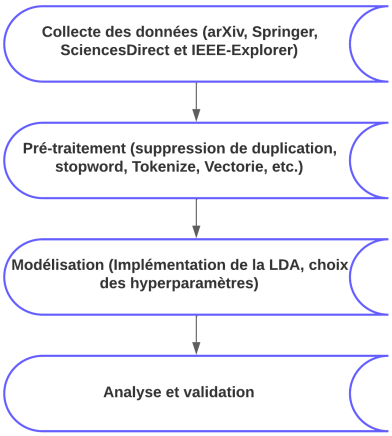


FIGURE 2 – Processus de modélisation.

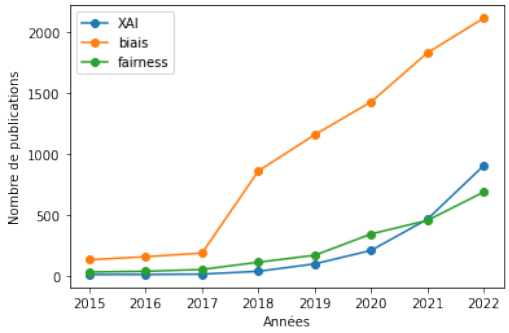


FIGURE 3 – Nombre de publications par thème (XAI, biais, équité) et par année entre 2015 et 2022.

2.1.1 Les données

Cette étude est basée sur les articles publiés sur les quatre plateformes de bases de données suivantes : arXiv, Springer, ScienceDirect et IEEE-Explorer. Sur chaque base de données, nous avons considéré les articles publiés entre 2015 et 2022 avec une recherche séparée sur les méta-données des termes suivants : bias AND (machine learning OR data) ; XAI AND (machine learning OR data) et ; fairness AND (machine learning OR data). Au total, 31 860 articles ont été obtenus. Ensuite, les tâches suivantes ont été réalisées :

- Suppression des duplications : articles ayant les mêmes auteurs, le même titre et le même résumé ;
- Suppression des publications sans résumé ;
- Suppression des articles en d'autres langues que l'anglais.

Par la suite, trois variables binaires ont été créées permettant de vérifier que la publication traite au moins un des trois thèmes : XAI, biais et équité (1 si oui, 0 sinon). Pour chaque thème, les termes suivants ont été considérés :

- Pour XAI : XAI, explainable, explainability, interpretable et interpretability ;
- Pour Biais : bias, harm et disparate ;
- Pour Fairness : fair.

Cette recherche a été faite sur le résumé, le titre et les mots clés de chaque article. Par la suite, seuls les articles ayant traité au moins, un des thèmes a été retenu. Au final, 9 874 publications ont été considérées pour l'étude.

Une analyse du nombre de publications par année montre une augmentation de plus en plus importante de ces dernières, notamment à partir de 2017 (FIGURE 3). Cette situation peut être expliquée par le contexte de cette année (expliquée dans l'introduction) qui a conduit à la création du domaine XAI. Cette analyse montre également que le nombre de publications traitant le biais est plus élevé. En effet, la problématique du biais est présente dans presque tous les domaines. D'où l'importance du nombre de publications qui l'ont traitée.

2.1.2 Pré-traitement

Un pré-traitement a été fait sur les données. Il s'agit de :

- la suppression de mots vides ou 'stopword' qui correspond aux mots qui n'apportent pas d'information tels que : 'the', 'and', etc. Leur suppression accélère l'apprentissage et améliore la précision.
- la tokenisation qui consiste à découper chaque document en une liste de mots appelés tokens. Cette étape conduit à l'obtention d'une matrice documents-termes (matrice d'occurrence).
- la lemmatisation qui consiste à remplacer tous les mots par leur mot-racine. De nombreux mots sont dérivés d'une racine ou d'un mot-racine. (Par exemple, explains, explained ⇒ explain) ;
- la normalisation qui consiste à pondérer chaque terme de la matrice d'occurrence. Dans notre cas, nous avons utilisé l'approche tf-idf (Joachims, T. et al., 1996[19]) qui permet d'évaluer l'importance d'un terme dans un document relativement à tous les autres documents.

Ce processus conduit à l'obtention d'une matrice creuse où chaque ligne correspond à un document et chaque colonne correspond à un mot. Suite à ce processus, une analyse du corpus a été faite pour choisir les paramètres optimaux.

2.1.3 Choix du corpus et des paramètres

Pour le choix du corpus, l'analyse est faite sur la concaténation du résumé et des mots-clés. En effet, il n'était pas possible que celle-ci soit faite sur les articles complets car nous disposions à notre niveau uniquement de leurs méta-données : résumé, titre, mots-clés, etc.

Par ailleurs, une analyse séparée a été faite sur le résumé, sur les mots-clés, sur la concaténation du résumé et des mots-clés et sur la concaténation de résumé, mots-clés et titre. L'analyse des résultats obtenus permet de constater que le corpus "concaténation du résumé et les mots-clés" donne de meilleurs résultats qualitatifs (sens des sujets obtenus) et quantitatifs au sens du score de cohérence (voir TABLE 1) qui évalue le degré de similitude sémantique entre les mots les mieux notés dans les sujets en moyenne.

En effet, après pré-traitement, nous avons procédé à un choix des paramètres du modèle K, α et β au regard du

score de cohérence. Ce processus de choix des paramètres est fait en fixant à chaque fois la valeur de K et en faisant varier les valeurs de α et β. Ainsi, nous avons retenu le triplet ($\alpha = 0.91$, $\beta = 0.91$, $K = 8$) (voir TABLE 1) qui donne le meilleur score de cohérence. Dans ce processus de recherche des paramètres optimaux, nous avons également fixé les valeurs de α et β à $0,91$ dans le corpus abstract-keywords puis nous avons varié la valeur de K comme le montre la FIGURE 4. Ainsi, on peut voir que $K = 8$ a le score de cohérence le plus élevé.

Corpus	K	α	β	C_V
Abstract	7	0,31	0,91	0,57
Keywords	9	asymmetric	0,61	0,52
Abstract-keywords	8	0,91	0,91	**0,58**
Abstract-keyw-title	9	0,61	0,91	0,56

TABLE 1 – Résultats de l'analyse des différents corpus.

Notons que l'option 'asymmetric' est une façon d'initialiser des probabilités d'appartenance d'un document à un sujet. Il consiste à initialiser de façon asymétrique ces dernières avec la formule $\frac{1}{topic\_index+\sqrt{(num\_topics)}}$. Pour de raison de comparabilité, ce même nombre de sujets $K = 8$ sera utilisé avec BERTopic.

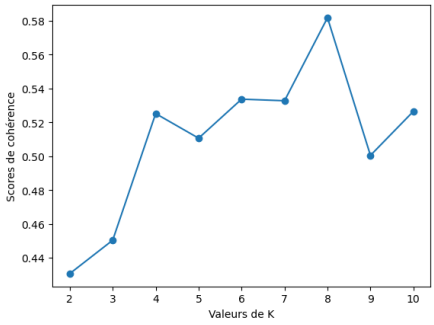


FIGURE 4 – Variation du score de cohérence en fonction du nombre de sujets K lorsque $\alpha = 0.91$ et $\beta = 0.91$ pour le corpus abstract et keywords.

2.2 Résultats et discussions

Dans cette section, nous allons présenter les différents résultats obtenus suite à l'application du modèle LDA dans notre corpus de documents compte tenu des choix de corpus et de hyperparamètres faits plus haut. Dans un premier temps, il s'agira de présenter les 8 sujets obtenus en se basant sur la matrice sujets-termes $\phi_{K,N}$. Dans chaque sujet les mots sont ordonnés suivant l'importance de leurs probabilités. Ensuite, en se basant sur la matrice documents-sujets $\theta_{M,K}$, les documents seront organisés en 8 clusters (sujets).

2.2.1 Analyse des sujets obtenus

Les sujets extraits au moyen de la LDA sont décrits chacun par les 10 mots les plus significatifs en termes de probabilité (TABLE 2). La FIGURE 6 correspond aux pourcentages de documents ayant traité chaque sujet. Par exemple, 13,14% des publications ont traité le sujet1.

L'analyse de ces résultats nous permet de décrire les sujets obtenus de la façon suivante :

- le sujet1, traité par 13,14% des publications, concerne le biais en science cognitive et son impact sur les différents groupes définis par le genre par exemple. La question du biais est très présente dans ce domaine. En effet, les expériences en sciences cognitives sont souvent menées sur des échantillons d'individus. Ainsi, on note de nombreuses publications sur le biais d'échantillonnage qui pourrait limiter la généralisation des résultats issus de ces expériences.
- Le sujet2, traité par le plus faible nombre de publications (2,98%), semble concerner les études de cas autour de l'éthique, la confidentialité des données individuelles (spam, social_medium, véhicule autonome, etc.).
- Le sujet3, traité par 6,77% des publications, porte sur les études de cas en science biologique surtout la génétique et les types de biais rencontrés dans ce domaine.
- Le sujet4, traité par 21,51% des publications, porte sur les données d'images. Il s'agit de d'approches de détection et de classification des images telles que l'apprentissage profond.
- Quant au sujet5, traité par 6,80% des publications, il semble traiter de l'équité, mais dans le cadre du Cloud computing et les objets connectés.
- Quant au sujet6, traité par le plus grand nombre de publications (24,23%), on peut voir qu'il porte sur la confiance en IA à travers l'explicabilité et l'équité des algorithmes de ML. Il s'agit notamment de l'explicabilité des algorithmes dans le cadre de la prise de décisions et des approches (shap, contrefactual) pour garantir l'équité algorithmique.
- Le sujet7, traité par 16,5% des publications, semble porter sur le biais statistique en général. À travers les termes "estimate, forecast, error, parameter" nous pouvons noter qu'il s'agit notamment du biais dans le cadre de l'estimation des paramètres en statistique appliquée.
- Le sujet8, traité par 8,09% des publications, porte sur les études de cas dans le domaine de santé. On note la présence de mots comme "patient, risk, sample, treatment, clinical, etc.".

2.2.2 Analyse des clusters de documents obtenus

L'approche LDA a permis d'organiser les documents en 8 clusters correspondant chacun à un sujet. Cette organisation est faite en se basant sur la matrice de probabilités document-sujet $\theta_{M,K}$. Chaque document est affecté au sujet pour lequel sa probabilité d'appartenance est plus élevée.

Par exemple, le cluster 1 correspond aux publications ayant une plus grande probabilité d'appartenance au sujet1 (FI-

sujet1	sujet2	sujet3	sujet4	sujet5	sujet6	sujet7	sujet8
bias	ethic	cell-coat	feature	fairness	explanation	estimate	patient
cognitive	privacy	gene	image	user	fairness	regression	bias
participant	risk	property	classification	attack	trust	error	risk
attention	policy	structure	detection	cloud_compute	decision	bias	sample
gender	governance	stress	recognition	resource_allocation	human	fault	clinical
stimulus	protection	bias	performance	federate	understand	prediction	vaccine
individual	social_medium	substrate	accuracy	traffic	shap	satellite	climate
group	spam	molecular	semisupervise	agent	counterfactual	forecast	mortality
negative	auto_driving	plasma	task	iot	transparency	performance	disease
social	gdpr	microstructure	cnn	market	interpretable	parameter	treatment

TABLE 2 – Description des sujets par les 10 mots les plus significatifs.

GURE 6).

Visualisation des sujets ou clusters de documents : Pour visualiser les sujets ou clusters de documents dans un espace à deux dimensions, nous avons utilisé l'outil pyLDAviz (Mabey et al., 2021 [25]). Il s'agit d'une approche de visualisation utilisant le positionnement multidimensionnel. L'intérêt de cette visualisation réside dans le fait qu'on arrive à voir :

- la popularité de chaque sujet en termes de nombre de documents l'ayant traité (reflétée par la taille de la surface du cercle) ;
- la similarité entre les différents sujets (reflétée par la proximité entre les cercles).

L'analyse de la distribution des clusters de publications (voir FIGURE 5) montre une bonne séparation de ces premiers (cercles non superposés). Ceci est une garantie de la qualité des résultats obtenus (Mabey et al., 2021 [25]). On peut voir aussi la non-popularité du sujet2 et la proximité entre les sujet7 et sujet4 ainsi que les sujet4 et sujet5.

Nous avons également analysé l'évolution dans le temps du nombre et de la part (en pourcentage) de publications traitant chaque sujet (FIGURE 7 et FIGURE 8). On note que la part de publications qui traitent le sujet6 (en vert) augmente de plus en plus depuis 2018. Cela correspond à une prise de conscience de plus en plus importante sur l'explicabilité des algorithmes de l'IA. Cependant, si la FIGURE 7 semble suivre la même tendance que les données initiales FIGURE 3, la FIGURE 8 a permis d'avoir la part relative de chaque sujet pour chaque année.

On note aussi une montée rapide du nombre de publications ayant traité le sujet4 qui porte sur les modèles de ML en données d'image. En effet, la problématique de l'explicabilité concerne surtout les approches dites "black box" comme les réseaux de neurones. Il s'agit d'approches utilisées surtout dans le cadre des données complexes telles que les données d'image.

Description des clusters de documents : Sur la base de la matrice $\theta_{M,K}$, les documents de chaque cluster ont été ordonnés au regard des probabilités d'appartenance. Chaque

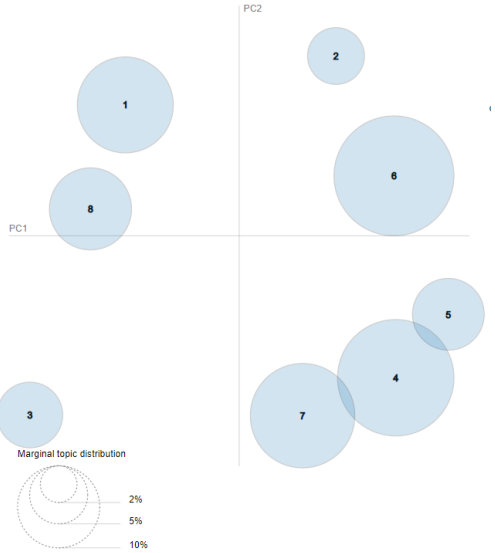


FIGURE 5 – Visualisation de la distribution des clusters par l'outil pyLDAviz (C. Sievert et al, 2014 [46]).

cluster peut être caractérisé par les documents ayant les plus fortes probabilités d'appartenance à ce cluster. Cette description des clusters a permis une organisation de la bibliographie autour du biais, de l'équité et de l'explicabilité en 8 clusters correspondant chacun à un sujet.

En ce qui nous concerne, on s'est surtout intéressé à la description des clusters 4 et 6 parce qu'ils sont plus pertinents par rapport à notre thématique de recherche. De surcroît, une quantification des thèmes recherchés par sujet extrait montre une forte présence de ces deux sujets (voir FIGURE 9).

Cluster 4 : Concernant le cluster 4, les publications les plus significatives sont données dans la TABLE 4 en annexe A. Une analyse approfondie des 21 publications de ce cluster a permis de valider le contenu du sujet correspondant. Il s'agit d'un sujet qui est porté sur l'analyse de données complexes telles que les images et le biais. Parmi les publications les plus significatives de ce cluster, certaines ont

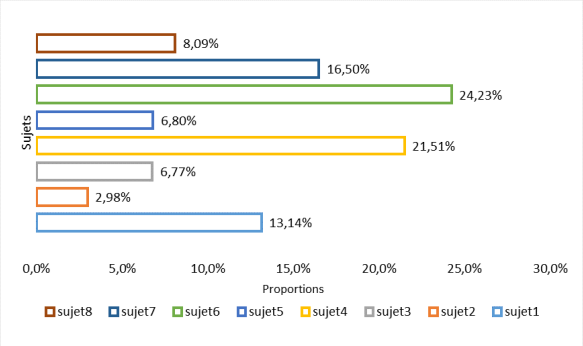


FIGURE 6 – Répartition des publications entre les sujets.

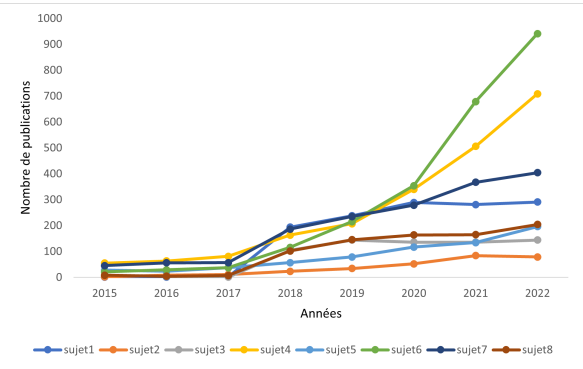


FIGURE 7 – Évolution du nombre de publications par sujet depuis 2015.

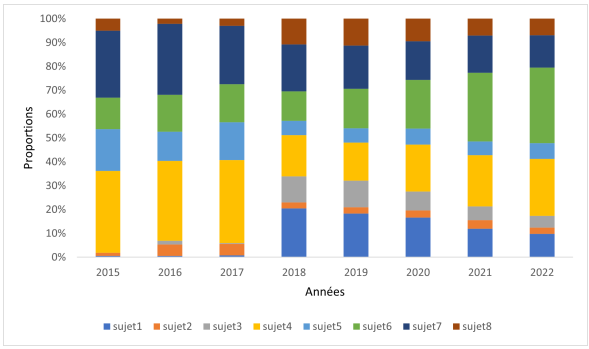


FIGURE 8 – Évolution de la part des publications traitant chaque sujet depuis 2015.

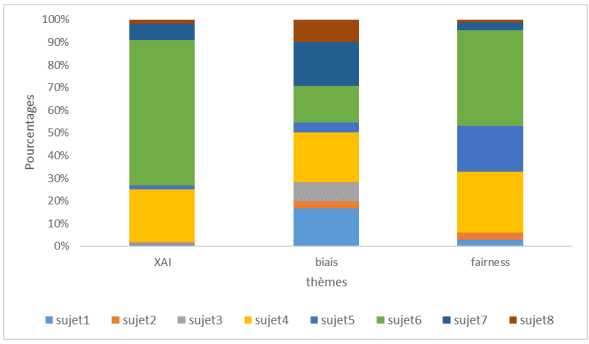


FIGURE 9 – Quantification des thèmes recherchés par sujet. Il s'agit de la répartition des 8 sujets extraits via LDA suivant chaque concept recherché.

analysé le biais selon le type de données (images, graphe, audio, tabulaire). C'est par exemple le cas de Zhengyu Chen et al., 2022 [5]. Constatant que les modèles GNNs (réseaux de neurones en graphes) peuvent être affectés par un potentiel biais lié à une différence de distribution des nœuds dans les données d'entraînement et les données de test, les auteurs ont proposé BA-GNN qui tient compte de cette différence.

Notre analyse a permis également de voir les études sur le biais selon le type de modèle d'analyse utilisé : semi-supervisé (Qiu et al., 2016 [37], Tao et al., 2017 [48])), non supervisé (Dumanvci et al., 2017 [8], Li et al., 2020 [21], Yu et al., 2021[54], Yu et al. [55]) et supervisé (2021[54], Liu et al., 2017 [24]).

Cluster 6 : Une première description a permis de constater que le sujet6 concernait l'équité et l'explicabilité des algorithmes de ML. Lorsqu'on regarde les 11 publications les plus significatives dans le cluster correspondant, on consolide ce constat. Il s'agit d'un sujet traité par des articles qui parlent de l'équité et l'explicabilité (voir TABLE 5, annexe B). Une lecture de ces articles permet de voir une liaison forte de ces deux termes. En effet, ces deux termes sont liés par l'aspect humain, mais également par le fait qu'ils concernent tous les deux directement les décisions prises sur la base des algorithmes de ML. Selon Julie Gerlings et al. (2022) [12], le domaine XAI a été créé dans le but de fournir à l'humain une compréhension des modèles dits de boîtes noires. Cette compréhension conduirait à améliorer la fiabilité des décisions prises sur la base de ces modèles de l'IA.

Si on note, aujourd'hui, un nombre élevé de modèles ou d'approches (appelés modèles XAI) pouvant expliquer les résultats des modèles taxés de "boîtes noires", Jose de Sousa Ribeiro Filho et al., 2022 [39] se demandent si ces explications fournies par l'IA sont conformes aux explications qu'un expert du domaine pourrait fournir. Les résultats de leur analyse ont permis de noter que les explications fournies par les modèles XAI, basées sur les attributs des différentes variables, ne sont pas tout le temps celles fournies par un expert du domaine qui est capable de tenir compte du contexte de l'analyse. Amit Sheth et al., 2021 [45] ont notamment souligné le lien entre la confiance en IA et le niveau d'explicabilité d'un système d'IA. À cet effet, ils soulignent que l'explicabilité ne s'arrête pas aux résultats, mais concerne également d'autres aspects tels que le biais que pourrait contenir les données (par exemple, défaut de représentativité d'un groupe) ou l'équité sur l'acquisition des données.

On note par ailleurs que cette problématique d'explicabilité

concerne tous les types de données et d'approche d'analyse : XAI et Deep Learning (Amit Sheth et al., 2021 [45]), XAI et apprentissage par renforcement (Erika Puiutta et al., 2020 [36]). La limite notée est la non prise en compte de l'aspect humain de façon générale. Cet aspect est essentiel, car pourrait garantir plus d'équité dans les différentes décisions prises sur la base des modèles de l'IA.
Une analyse rapide des autres sujets permet de constater la présence du biais dans de nombreux domaines selon différentes acceptions : (sciences cognitives (sujet1), biologie (sujet3) et santé (sujet8). On a également noté la présence d'un sujet traitant l'éthique et la confidentialité en IA (sujet2), mais également le sujet5 traitant l'équité en Cloud computing.

2.2.3 Comparaison de nos résultats avec ceux de BERTopic

L'application de BERTopic sur nos données montre que nos résultats sont meilleurs que BERTopic au regard de C_V score lorsque HDBSCAN (l'option par défaut) est utilisée comme méthode de clustering(TABLE 3). Cependant, le meilleur C_V score est donné par le modèle BERTopic combiné à un K-moyennes. L'analyse des sujets extraits par ce dernier BERTopic permet de noter une forte similarité entre certains de ces sujets et ceux de notre modèle (TABLE 6 et FIGURE 10) ; on cite : sujet1 BERTopic et sujet4 LDA, sujet2 BERTopic et sujet6 LDA, sujet3 BERTopic et sujet1 LDA).
Cette analyse permet de valider les résultats de notre modèle. Cependant, il faut noter que BERTopic est une approche qualifiée de boîte noire car ayant une structure assez complexe.

	BERT_kmeans	BERT_hdbscan	LDA
UMASS	-0,11	-0,98	-4,31
C_V	0,81	0,57	0,58

TABLE 3 – Scores de cohérence des modèles BERTopic et LDA

Conclusion

Ce travail a permis de classifier la littérature autour du biais, de l'équité et de l'XAI selon 8 sujets : le biais en science cognitive, en santé, en science biologique, en télédétection, l'éthique et la confidentialité des données, les données d'image et les "boîtes noires", l'équité en Cloud computing et enfin l'équité et l'explicabilité des algorithmes de ML. L'analyse de l'évolution des sujets dans le temps permet de noter une évolution plus rapide du nombre et du pourcentage de publications sur les sujets l'équité et l'explicabilité des algorithmes de l'IA.
L'extraction des sujets traités de la bibliographie sur le biais,l'équité et l'XAI a permis de réorganiser de cette bibliographie. Nous avons constaté que cette bibliographie couvre un grand nombre de domaines notamment en ce qui concerne le biais. Il a permis aussi d'extraire un sujet portant exclusivement sur les données complexes comme les images, ainsi que les modèles adaptés à leur analyse comme les CNN qui ont fortement impacté le domaine de l'IA. Ainsi, l'utilisation des approches de traitement de langage naturel pour synthétiser, résumer, et même organiser une bibliographie reste très utile dans un contexte des données massives (big data) où un besoin d'analyse systématique se pose de plus en plus. En effet, cela peut être utile pour organiser une bibliographie en permettant d'aborder de manière directe les principaux sujets d'intérêt. En ce qui nous concerne, la classification obtenue permettra d'organiser notre bibliographie pour une meilleure exploitation de celle-ci.
Cependant, il est important de noter que cette analyse se base sur un échantillon de l'ensemble des publications autour du biais de l'équité et de l'explicabilité. En effet, en dehors des 4 plateformes choisies, il en existe d'autres telles que IJCAI et ECAI. Une analyse d'une plus large base de données pourrait faire ressortir d'autres sujets aussi importants que ceux obtenus. Dans cette analyse, nous avons aussi fait une comparaison rapide entre nos résultats et ceux de BERTopic. Puisque BERTopic fournit de bons résultats dans ce domaine, dans nos travaux futurs, nous souhaitons analyser davantage nos documents à l'aide de ce modèle afin d'améliorer nos résultats.

Références

[1] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, pages 1–17, 2022.

[2] Patrice Bertail, David Bounie, Stéphan Clémençon, and Patrick Waelbroeck. Algorithmes : biais, discrimination et équité. *NA*, 2019.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

[4] Petra Budikova, Michal Batko, and Pavel Zezula. Conceptrank for search-based image annotation. *Multimedia Tools and Applications*, 77 :8847–8882, 2018.

[5] Zhengyu Chen, Teng Xiao, and Kun Kuang. Ba-gnn : On learning bias-aware graph neural network. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 3012–3024. IEEE, 2022.

[6] Ziheng Chen and Jiangtao Ren. Multi-label text classification with latent word-wise label information. *Applied Intelligence*, 51 :966–979, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[8] Sebastijan Dumančić and Hendrik Blockeel. An expressive dissimilarity measure for relational clustering using neighbourhood trees. *Machine learning*, 106 :1523–1545, 2017.

[9] Maksim Ekin Eren, Nick Solovyev, Edward Raff, Charles Nicholas, and Ben Johnson. Covid-19 kaggle literature organization. In *Proceedings of the ACM Symposium on Document Engineering 2020*, pages 1–4, 2020.

[10] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot learning on semantic class prototype graph. *IEEE transactions on pattern analysis and machine intelligence*, 40(8) :2009–2022, 2017.

[11] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 577–590, 2021.

[12] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. Explainable ai, but explainable to whom ? an exploratory case study of xai in healthcare. *Handbook of Artificial Intelligence in Healthcare : Vol 2 : Practicalities and Prospects*, pages 169–198, 2022.

[13] Bernadeta Griciūtė, Lifeng Han, Alexander Koller, and Goran Nenadic. Topic modelling of swedish newspaper articles about coronavirus : a case study using latent dirichlet allocation method. *arXiv preprint arXiv :2301.03029*, 2023.

[14] Maarten Grootendorst. Bertopic : Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv :2203.05794*, 2022.

[15] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2) :44–58, 2019.

[16] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel : Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10) :6767–6781, 2021.

[17] Carina Jacobi, Wouter Van Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. In *Rethinking Research Methods in an Age of Digital Journalism*, pages 89–106. Routledge, 2018.

[18] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. Euca : The end-user-centered explainable ai framework. *arXiv preprint arXiv :2102.02437*, 2021.

[19] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[20] Michail Kaseris, Ioannis Mademlis, and Ioannis Pitas. Adversarial unsupervised video summarization augmented with dictionary loss. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2683–2687. IEEE, 2021.

[21] Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9070–9079, 2020.

[22] Ruihui Li, Jianrui Cai, Hanling Zhang, and Taihong Wang. Aggregating complementary boundary contrast with smoothing for salient region detection. *The Visual Computer*, 33 :1155–1167, 2017.

[23] Tao Lian, Lin Du, Mingfu Zhao, Chaoran Cui, Zhumin Chen, and Jun Ma. Evaluating and improving the interpretability of item embeddings using item-tag relevance information. *Frontiers of Computer Science*, 14 :1–16, 2020.

[24] Meng Liu, Chang Xu, Yong Luo, Chao Xu, Yonggang Wen, and Dacheng Tao. Cost-sensitive feature selection by optimizing f-measures. *IEEE Transactions on Image Processing*, 27(3) :1323–1335, 2017.

[25] Ben Mabey. pyldavis documentation, 2021.

[26] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7) :655–665, 2016.

[27] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6) :1–35, 2021.

[28] Alain Mille, Rémy Chaput, and Amélie Cordier. *Une perspective historique sur l'IA explicable Document préparatoire à un tutorial AFIA juillet 2020*. PhD thesis, LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université . . ., 2020.

[29] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

[30] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness : Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8 :141–163, 2021.

[31] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research*, 9(5) :583–597, 2002.

[32] Oliver Nina, Jamison Moody, and Clarissa Milligan. A decoder-free approach for unsupervised clustering and manifold learning with random triplet mining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[33] Ahmad Kamal Bin Mohd Nor, Srinivasa Rao Pedapait, and Masdi Muhammad. Explainable ai (xai) for phm of industrial asset : A state-of-the-art, prisma-compliant systematic review. *arXiv preprint arXiv :2107.03869*, 2021.

[34] Bayode Ogunleye, Tonderai Maswera, Laurence Hirsch, Jotham Gaudoin, and Teresa Brunsdon. Comparison of topic modelling approaches in the banking context. *Applied Sciences*, 13(2), 2023.

[35] Alvin Poernomo and Dae-Ki Kang. Biased dropout and crossmap dropout : learning towards effective dropout regularization in convolutional neural network. *Neural networks*, 104 :60–67, 2018.

[36] Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning : A survey. In *Machine Learning and Knowledge Extraction : 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 77–95. Springer, 2020.

[37] Zhicong Qiu, David J Miller, and George Kesidis. A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes. *IEEE transactions on neural networks and learning systems*, 28(4) :917–933, 2016.

[38] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing : Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

[39] José Ribeiro, Níkolas Carneiro, and Ronnie Alves. Black box model explanations and the human interpretability expectations–an analysis in the context of homicide prediction. *arXiv preprint arXiv :2210.10849*, 2022.

[40] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[41] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable artificial intelligence for tabular data : A survey. *IEEE Access*, 9 :135392–135422, 2021.

[42] Vasudeva Raju Sangaraju, Bharath Kumar Bolla, Deepak Kumar Nayak, and Jyothsna Kh. Topic modelling on consumer financial protection bureau data : An approach using bert based embeddings. *arXiv preprint arXiv :2205.07259*, 2022.

[43] Teresa Scantamburlo. Non-empirical problems in fair machine learning. *Ethics and Information Technology*, 23(4) :703–712, 2021.

[44] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence : a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.

[45] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. Knowledge-intensive language understanding for explainable ai. *IEEE Internet Computing*, 25(5) :19–24, 2021.

[46] Carson Sievert and Kenneth Shirley. Ldavis : A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[47] S Regina Lourdhu Suganthi, M Hanumanthappa, and S Kavitha. Event image classification using deep learning. In *2018 International Conference on Soft-computing and Network Security (ICSNS)*, pages 1–8. IEEE, 2018.

[48] Hong Tao, Chenping Hou, Feiping Nie, Jubo Zhu, and Dongyun Yi. Scalable multi-view semi-supervised classification via adaptive regression. *IEEE Transactions on Image Processing*, 26(9) :4283–4296, 2017.

[49] Chen Wang, Chengyuan Deng, and Vladimir Ivanov. Sag-vae : End-to-end joint inference of data representations and feature relations. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

[50] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19 : The covid-19 open research dataset. *ArXiv*, 2020.

[51] Yali Wang, Lei Zhou, and Yu Qiao. Temporal hallucinating for action recognition with few still images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5314–5322, 2018.

[52] Yulong Wang, Wei Yang, and Haoxin Zhang. Deep learning single logo recognition with data enhancement by shape context. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[53] Yao Xie, Ge Gao, and Xiang'Anthony' Chen. Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv :1902.06019*, 2019.

[54] Heng Yu, Haoran Luo, Yuqi Yi, and Fan Cheng. A2r2 : robust unsupervised neural machine translation with adversarial attack and regularization on representations. *IEEE Access*, 9 :19990–19998, 2021.

[55] Lingli Yu, Xumei Xia, and Kaijun Zhou. Traffic sign detection based on visual co-saliency in complex scenes. *Applied Intelligence*, 49 :764–790, 2019.

[56] Kai Zhao, Qi Han, Chang-Bin Zhang, Jun Xu, and Ming-Ming Cheng. Deep hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9) :4793–4806, 2021.

A Publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 4

Références	Titres (en anglais)	Probabilités
[37]	A Maximum Entropy Framework for Semisupervised and Active Learning With Unknown and Label-Scarce Classes	0,9944
[35]	Biased Dropout and Crossmap Dropout : Learning towards effective Dropout regularization in convolutional neural network	0,9940
[56]	Deep Hough Transform for Semantic Line Detection	0,9940
[51]	Temporal Hallucinating for Action Recognition with Few Still Images	0,9938
[23]	Evaluating and improving the interpretability of item embeddings using item-tag relevance information	0,9934
[8]	An expressive dissimilarity measure for relational clustering using neighbourhood trees	0,9929
[10]	Zero-Shot Learning on Semantic Class Prototype Graph	0,9925
[22]	Aggregating complementary boundary contrast with smoothing for salient region detection	0,9921
[49]	SAG-VAE : End-to-end Joint Inference of Data Representations and Feature Relations	0,9919
[52]	Deep Learning Single Logo Recognition with Data Enhancement by Shape Context	0,9919
[24]	Cost-Sensitive Feature Selection by Optimizing F-Measures	0,9914
[47]	Event Image Classification using Deep Learning	0,9913
[48]	Scalable Multi-View Semi-Supervised Classification via Adaptive Regression	0,9913
[20]	Adversarial Unsupervised Video Summarization Augmented With Dictionary Loss	0,9912
[16]	AutoNovel : Automatically Discovering and Learning Novel Visual Categories	0,9912
[55]	Traffic sign detection based on visual co-saliency in complex scenes	0,9911
[32]	A Decoder-Free Approach for Unsupervised Clustering and Manifold Learning with Random Triplet Mining	0,9910
[4]	ConceptRank for search-based image annotation	0,9910
[6]	Multi-label text classification with latent word-wise label information	0,9909
[54]	Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification	0,9905
[5]	BA-GNN : On Learning Bias-Aware Graph Neural Network	0,9904

TABLE 4 – Les publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 4

B Publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 6

Références	Titres (en anglais)	Probabilités
[39]	Black Box Model Explanations and the Human Interpretability in the Context of Homicide Prediction	0,9922
[43]	Non-empirical problems in fair machine learning	0,9922
[45]	Knowledge-Intensive Language Understanding for Explainable AI	0,9912
[11]	Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals	0,9910
[1]	A review of predictive policing from the perspective of fairness	0,9910
[18]	EUCA : the End-User-Centered Explainable AI Framework	0,9908
[41]	Explainable Artificial Intelligence for Tabular Data : A Survey	0,9908
[12]	Explainable AI, but explainable to whom ?	0,9906
[36]	Explainable Reinforcement Learning : A Survey	0,9902
[53]	Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis	0,9900
[44]	A Comprehensive Taxonomy for Explainable Artificial Intelligence : A Systematic Survey of Surveys on Methods and Concepts	0,9900

TABLE 5 – Les publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 6

C Description des sujets fournis par BERTopic

sujet1	sujet2	sujet3	sujet4	sujet5	sujet6	sujet7	sujet8
algorithm	bias	bias	bias	bias	bias	algorithm	receptor
model	fairness	attentional	patient	exchange	estimation	resource	cell
feature	algorithm	participant	disease	magnetic	model	traffic	signaling
neural	system	cognitive	risk	exchange bias	error	proposed	gene
method	user	attention	clinical	field	correction	detection	protein
proposed	model	attentional bias	result	film	estimate	fairness	agonist
image	decision	stimulus	study	substrate	result	network	biased
classification	social	result	medical	property	measurement	performance	codon
performance	research	task	model	coating	proposed	attack	ligand
problem	information	negative	method	device	estimator	allocation	pathway

TABLE 6 – Description des 8 sujets fournis par BERTopic_kmeans

D Correspondance entre les 8 sujets LDA et les 8 sujets BERT.

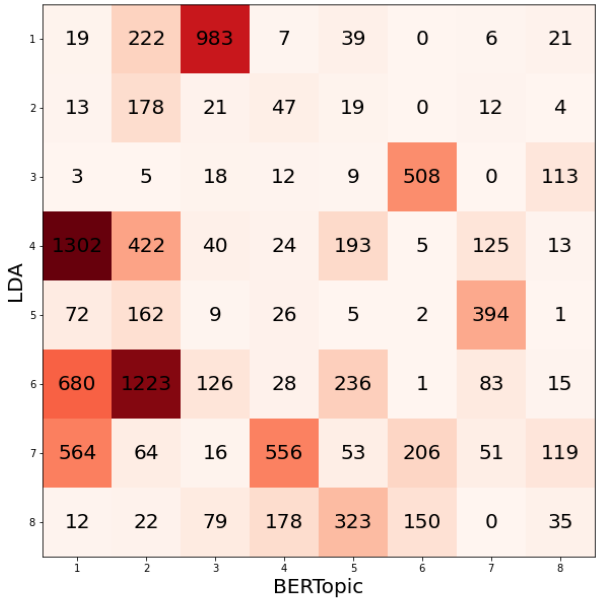


FIGURE 10 – Matrice de confusion entre les clusters de documents des modèles LDA et BERTopic. Cette matrice montre la correspondance quantitative entre les clusters de documents obtenus par les deux approches : LDA et BERTopic.

Panorama des outils de visualisation pour l'explicabilité en apprentissage profond pour le traitement automatique de la langue

A. Delaforge[1,3], J. Azé[1], S. Bringay[1,2], A. Sallaberry[1,2], M. Servajean[1,2]

[1] LIRMM, UMR 5506, Université de Montpellier, CNRS, Montpellier, France
[2] AMIS, Université Paul-Valéry, Montpellier, France
[3] Zortify Labs, Zortify, Luxembourg, Luxembourg

prenom.nom@lirmm.fr

Résumé

L'avènement de l'Intelligence Artificielle (IA) et plus spécifiquement des modèles d'apprentissage profond, a été accompagné de résultats impressionnants dans le domaine du Traitement Automatique de la Langue (TAL). Derrière les performances des réseaux de neurones se cachent de nombreuses problématiques, comme l'interprétabilité. Dans cet article d'état de l'art, nous dressons un panorama des outils de visualisation spécifiques à l'explicabilité des méthodes d'apprentissage profond en TAL qui visent à pallier le caractère boîte noire de ces approches.

Mots-clés

Interprétabilité, transparence, explicabilité, XAI, visualisation, réseaux de neurones.

Abstract

The advent of Artificial Intelligence (AI), and more specifically of deep learning models, has been accompanied by impressive results in the field of Natural Language Processing (ALP). Behind the performance of neural networks lie many issues, such as interpretability. In this state of the art article, we present an overview of visualization tools specific to the explicability of deep learning methods in NLP, which aim to overcome the black box character of these approaches.

Keywords

Interpretability, transparency, explainability, XAI, visualization, neural networks.

1 Introduction

L'avènement de l'Intelligence Artificielle (*IA*), accompagné de résultats impressionnants dans le domaine du Traitement Automatique de la Langue (*TAL*) [88] soulève des problématiques, comme l'interprétabilité [25, 97, 34], l'éthique [62, 26] et la sécurité [29, 34]. Dans cet article, nous nous intéressons aux outils pour interpréter les méthodes d'apprentissage profond en palliant leur caractère boîte noire. Autour de cette problématique, les travaux d'Hohman et al. [32] proposent un état de l'art traitant de l'utilisation de la visualisation de données dans le contexte de l'apprentissage profond. La Rosa et al. [43] proposent un état de l'art s'intéressant spécifiquement aux méthodes d'explication des prédictions. Chatzimparmpas et al. [11] proposent, eux, un état de l'art des états de l'art sur ces sujets. La principale différence avec ces trois travaux est que l'on s'intéresse dans cet article spécifiquement au *TAL* et à l'explication des prédictions en *TAL*. Gilpin et al. [23] étudient les méthodes d'explication en apprentissage profond. Similairement, Madsen et al. [55] s'intéressent à la même problématique mais pour le *TAL*. La principale différence de nos travaux avec ces derniers est que nous nous concentrons ici sur les outils de visualisation servant l'explication des prédictions pour les données textuelles.

L'interprétabilité des méthodes d'apprentissage sert à accroître la confiance des utilisateurs. Pour augmenter l'interprétabilité, il est possible : (1) d'augmenter la transparence du modèle, c'est-à-dire rendre les processus inhérents au modèle facilement compréhensibles ou (2) d'expliquer les prédictions, c'est-à-dire aider à identifier les raisons pour lesquelles un modèle a pris une décision [49].

Les méthodes d'apprentissage profond en *TAL*, nécessitent des méthodes spécifiques pour tendre vers une plus grande interprétabilité. En effet, les explications concernant des images ou des données numériques brutes se construisent différemment de celles représentant un concept contenu dans un mot capturé par un modèle d'apprentissage profond. Il est donc essentiel d'adapter les méthodes d'interprétabilité aux données textuelles et aux tâches à accomplir car les techniques de *TAL* cherchent à représenter les mots, tokens, phrases ou textes dans des espaces de représentation non compréhensibles.

Les techniques de visualisation de données offrent de nombreuses opportunités [32, 43] que nous détaillerons dans cet état de l'art. Son objectif est de dresser un panorama des outils de visualisation spécifiques à l'explicabilité des méthodes d'apprentissage profond et des méthodes de visualisation de données utilisées par ces outils, un sous domaine de l'explicabilité de l'*IA* (eXplainable Artifical Intelligence (*XAI*) en anglais). Dans la section 2, nous redéfinirons tout d'abord les concepts d'interprétabilité, de transparence et d'explicabilité puis dans la section 3, nous présenterons les outils de visualisation utilisées pour l'inter-

prétabilité pour différents modèles d'apprentissage profond avant de conclure dans la section 4. Nous donnerons également quelques perspectives sur les challenges associés à ce type d'approche pour pallier les limitations des méthodes actuelles.

2 Interprétabilité : Explicabilité et Transparence

Les notions d'interprétabilité, d'explicabilité et de transparence ne font pas consensus dans la littérature. Lipton [49] définit deux concepts qui, ensemble, définissent l'interprétabilité : la transparence et les explications post-hoc. Waltl et Vost [92] présentent la transparence et l'interprétabilité comme des sous-catégories de l'explicabilité. Beaudouin et al. [5] utilisent l'interprétabilité et l'explicabilité comme des synonymes. Enfin, Chatzimparmpas et al. [11] utilisent les définitions de Gilpin et al. [23] qui présentent l'explicabilité comme la possibilité pour un modèle de résumer les raisons de son comportement et l'interprétabilité comme la compréhension de ce qu'un modèle a fait. Face à ce manque de consensus, nous précisons, dans la suite de cette section, notre vision de la transparence et de l'explicabilité nécessaire à l'interprétabilité d'un réseau de neurones.

2.1 Transparence d'un modèle

Nous définissons la transparence comme la facilité avec laquelle un humain peut comprendre et reproduire le fonctionnement d'un modèle, indépendamment d'une prédiction. Une régression linéaire est un modèle transparent car son fonctionnement est facilement compréhensible et reproductible par un humain. Au contraire, un réseau de neurones profond dépend de l'activation de millions de neurones. Il est impossible, de comprendre précisément leur fonctionnement ou d'expliquer pourquoi telle coordonnée d'un vecteur de représentation est définie à une valeur et ce que cette valeur représente. D'après [49], la transparence d'un modèle peut être divisée en trois parties :

- **Transp.1** La compréhension globale du fonctionnement du modèle ;
- **Transp.2** La compréhension des différentes parties du modèle ;
- **Transp.3** La compréhension des mécanismes d'apprentissage et de leur convergence vers une solution optimale.

2.2 Explication d'une prédiction

L'explication d'une prédiction, ou explication post-hoc [49] ou encore explication locale [5, 55], est faite lorsque des indicateurs, issus ou non du fonctionnement d'un modèle, sont utilisés pour expliquer sa décision. Si une explication associée à une prédiction sert marginalement à interpréter un réseau, la multiplication des explications peut donner aux utilisateurs des intuitions sur le fonctionnement du modèle. Lipton [49] propose quatre catégories d'explications :

- **Exp.1** Les explications verbales justifiantes ;
- **Exp.2** Les explications locales donnant accès à des explications plus simples ne concernant qu'un sous-ensemble de l'espace de données ;
- **Exp.3** Les explications de complexité modérée présentant des comportements pour des exemples similaires ;
- **Exp.4** Les techniques de visualisation pour explorer l'espace de représentation des données ou afficher des indications sur les parties, dans les données d'entrée, qui contribuent à la prédiction.

Certaines méthodes d'explication appartiennent à plusieurs de ces catégories. Dans cet article, nous nous focalisons sur la quatrième catégorie **Exp.4**.

Indépendamment des catégories auxquelles appartient une explication, il est nécessaire de quantifier sa qualité. Pour cela, on peut imaginer trois types d'évaluations pour mesurer cette qualité : l'utilité pour une application, l'utilité générale et la fidélité [20]. L'utilité pour une application mesure l'utilité de l'explication dans le contexte dans lequel le modèle est utilisé. Dans le cas d'un modèle classifiant des textes servant d'aide à la décision, un utilisateur des explications produites est-il plus efficace qu'un utilisateur n'y ayant pas accès, lorsque l'on mesure la qualité des décisions prises. La seconde mesure possible concerne l'utilité générale évaluant la propension d'un utilisateur à choisir le meilleur modèle parmi un ensemble de modèles, prédire le comportement d'un modèle sur de nouvelles données ou identifier les données anormales dans un jeu de données. Dans le cas d'une classification de textes, un utilisateur pourrait comprendre à l'aide d'une explication, qu'une certaine combinaison de mots produit toujours la même classification. Il pourrait comprendre que cette combinaison est importante pour la prédiction et ainsi prédire le fonctionnement du modèle sur de nouvelles données contenant cette combinaison. Ce concept d'utilité générale est également à mettre en lien avec la plausibilité de l'explication [35, 66] qui détermine à quel point une explication est convaincante pour un utilisateur. Un utilisateur qui n'est pas convaincu par les explications produites pour un modèle pourrait par exemple se reporter sur un autre modèle pour lesquelles les explications seraient plus convaincantes. La fidélité mesure à quel point l'explication reflète réellement le fonctionnement d'un modèle [35]. Ces deux précédents concepts, la plausibilité et la fidélité sont décrits plus précisément dans les travaux de Jacovi et al. [35]. Ces concepts, qui ne s'opposent pas, ne s'adressent pas au même public. La plausibilité s'adresse aux personnes qui ne sont pas expertes en apprentissage automatique mais dans un domaine dans lequel on l'applique (en *TAL* par exemple) quand la fidélité s'adresse aux experts en apprentissage profond.

3 Visualisations appliquées des réseaux de neurones

La visualisation de données cherche à résumer, à mettre en lumière les caractéristiques des données pour assister les utilisateurs dans l'analyse, la recherche des informations précises, le requêtage ou la production de nouvelles données [63]. Dans la classification de Lipton [49], la quatrième catégorie **Exp.4** est dédiée aux techniques de vi-

sualisation explorant l'espace de représentation des données ou affichant des indications sur la partie, dans les données d'entrée, qui contribue à la prédiction. Néanmoins, les techniques de visualisation ne se cantonnent pas à cette catégorie (voir tableau 1). Dans cette section, nous allons donc présenter les différentes applications des techniques de visualisation de données dans le domaine des réseaux de neurones, et plus spécifiquement dans le domaine du *TAL*. Celles-ci couvrent de larges missions, comme l'aide à la transparence, à l'explicabilité ou à la présentation des résultats issus des réseaux de neurones.

Il existe de nombreux objets à visualiser pour expliquer le fonctionnement des réseaux de neurones. Hohman et al. [32] proposent dans leur état de l'art sur la visualisation de données dans l'apprentissage profond, cinq catégories d'objets à visualiser :

- **Obj.1** L'architecture des réseaux de neurones ;
- **Obj.2** Les paramètres des réseaux de neurones ;
- **Obj.3** Les unités de calcul ou couches de neurones ;
- **Obj.4** Les vecteurs de représentation des données dans des espaces à grandes dimensions ;
- **Obj.5** Les informations agrégées issues ou non du fonctionnement du modèle.

Certaines des méthodes présentées peuvent appartenir à plusieurs de ces catégories. La dernière catégorie couvre des travaux n'appartenant pas aux autres catégories.

Dans la suite de cette section, nous nous intéressons à la visualisation de ces différents objets pour différents types de réseaux de neurones (perceptrons multicouche, réseaux de neurones convolutionnels, réseaux de neurones récurrents ou réseaux auto-attentifs).

3.1 Perceptrons multicouches et Réseau neuronal convolutif

La plupart des outils se focalisent sur la visualisation de l'architecture des réseaux de neurones (**Obj.1**) pour les Perceptrons multicouches (Multi Layer Perceptron, *MLP*) et Réseau neuronal convolutif (Convolutional Neural Network, *CNN*). Tensorboard, présent dans la bibliothèque Tensorflow [1], permet de voir les matrices de données traverser les différentes couches et de connaître leurs dimensions. Les outils comme Netron [1] ou Netscope CNN Analyzer [2] présentent des stratégies très similaires à Tensorboard. Les travaux de Harley et al. [28] et de Smilkov et al. [81] (voir TensorFlow Playground [3]) permettent aussi de visualiser les architectures des réseaux de neurones simples dans une démarche pédagogique. DAX [2], présente l'influence des tokens, et à travers quels filtres ou neurones leur influence a été augmentée (qu'elle soit dans le sens de la prédiction ou non). DAX montre également quels sont les tokens qui activent le plus les filtres ou neurones concernés à l'aide de nuages de mots. Chawla et al. [12] proposent, eux, d'expliquer la participation des tokens dans une tâche de classification via un *CNN* à l'aide de fenêtres de convolutions de différentes tailles permettant d'avoir des informations sur l'influence des n-grams de tokens et donc des meilleurs paramètres pour les fenêtres de convolutions.

1. https://github.com/lutzroeder/Netron
2. http://dgschwend.github.io/netscope/quickstart.html
3. http://playground.tensorflow.org/

3.2 Réseaux de neurones récurrents

Les réseaux de neurones récurrents (Recurrent neural networks, *RNN*) ont été pendant longtemps les réseaux incontournables en *TAL* [41]. Les différents travaux en visualisation de données à propos des *RNN* sont parfois seulement adaptés à certaines architectures (*GRU* ou *LSTM*). Dans cette partie, nous traitons donc des outils de visualisation appliqués à tous les *RNN* et précisons les architectures concernées.

Pour la visualisation de l'architecture des *RNN* (**Obj.1**) Tensorboard [1] construit la visualisation d'une cellule de type *RNN* et les processus la composant. Du fait de leur caractéristique de récurrence, la visualisation a peu d'intérêt puisque chaque cellule d'un *RNN* est identique aux autres (dans le cas d'un réseau unidirectionnel avec une unique couche de *RNN*). Les paramètres des *RNN* (**Obj.2**) ne sont eux-aussi que peu visualisés pour la même raison.

Concernant les unités de calcul ou couches de neurones des *RNN* (**Obj.3**), les travaux de Karpathy et al. [42] ou Qian et al. [70] observent l'activation de la fonction mettant à jour le nouvel état caché à l'aide du précédant contexte dans un *LSTM*. Les travaux de Kadar et al. [39, 40] montrent qu'utiliser des tokens plutôt que des caractères permet de constater si les catégories lexicales et les fonctions grammaticales (et donc l'information sémantique que peuvent porter les mots), sont capturées par les *RNN*. Linzen et al. [48] montrent que les *LSTM* capturent des structures grammaticales à l'aide d'apprentissage supervisé. Là encore, ils se basent sur l'étude des états cachés de *LSTM*. LSTMVis [83] et RNNVis [61] analysent les états cachés des *RNN* à l'aide d'outils interactifs. L'utilisateur peut sélectionner une fenêtre de tokens dans la donnée d'entrée de manière à observer l'évolution de l'état caché dans cette fenêtre. Certaines méthodes utilisent des cartes de chaleur pour identifier les mots ayant le plus participé à une prédiction [45, 46, 3, 65, 64]. Il est important de noter que pour les *RNN*, les catégories **Obj.2** et **Obj.3** se confondent car on peut considérer l'état caché comme une représentation de la donnée d'entrée pour un token et non comme une unité de calcul.

En ce qui concerne les vecteurs de représentation des données dans des espaces à grandes dimensions (**Obj.4**), Qian et al. [70] visualisent à l'aide d'une *ACP* l'espace de représentation des mots construit par un *LSTM* et l'activation résultant de leur traitement pour certaines dimensions de portes d'un *LSTM*. Linzen et al. [48] montrent que les représentations des mots construites par un *LSTM* sont capables de connaître le caractère pluriel ou singulier d'un mot alors que cette information était omise pendant l'entraînement (sans le "s" pour l'anglais pas exemple) à l'aide d'une *ACP* de quelques mots sélectionnés. Li et al. [45] utilisent également une *ACP* dans l'espace de représentation des groupes de tokens pour montrer l'influence des comparatifs et des superlatifs sur la représentation d'un groupe de

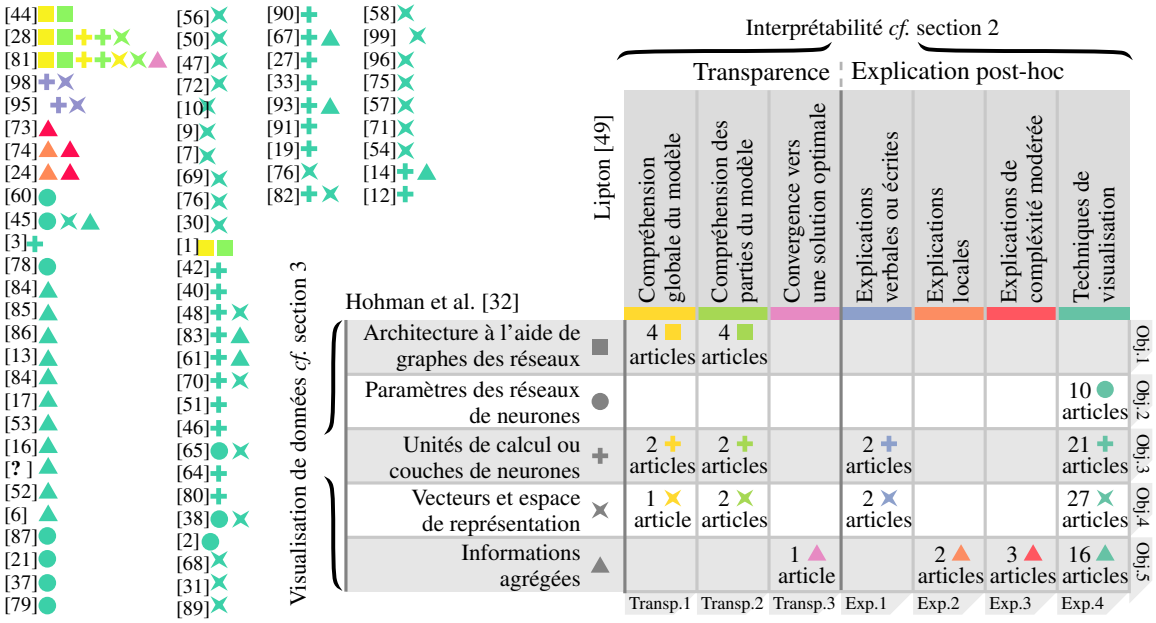


TABLE 1 – Méthodes d'interprétabilité présentées selon les classifications de Hohman et al. [32] et de Lipton [49]. Les catégories ne sont pas exhaustives car, selon le sujet, certaines catégories se confondent (*i.e.*, les unités de calcul et les vecteurs de représentation dans le cas d'utilisation de *RNN* ou réseaux auto-attentifs.)

tokens issu d'un *RNN*.

Pour visualiser des informations agrégées issues ou non du fonctionnement du modèle (**Obj.5**), les outils LSTMVis [83] et RNNVis [61] analysent des *POS* (ou étiquetage morpho-syntaxique) en plus des états cachés, par exemple en complément de la visualisation des états cachés. Li et al. [45], eux, inspectent le rôle de l'intensification et la négation dans la représentation de texte, en inspectant les états cachés des *RNN* à différents instants.

3.3 Réseaux de neurones auto-attentifs

Les réseaux de neurones auto-attentifs (dont les transformeurs) ont supplantés les *RNN* dans la plupart des tâches en *TAL*. Comme pour les *RNN*, il n'y a que peu de travaux concernant la visualisation de l'architecture (**Obj.1**) des réseaux de neurones auto-attentifs ou leurs paramètres (**Obj.2**).

Les matrices d'attention qui peuvent être vues comme des unités de calcul ou des couches de neurones (**Obj.3**), sont issues des nombreuses couches d'attention des réseaux auto-attentifs. Nlize [51] visualise des matrices d'attention pour une tâche d'inférence en langage naturel. Les cases de ces matrices sont colorées avec un encodage séquentiel des couleurs pour identifier les liens faibles ou forts d'attention entre deux tokens en fonction de la couleur foncée ou claire. Nlize visualise aussi à l'aide de graphe biparti des informations similaires. Les visualisations de graphe biparti [51, 90, 67, 14, 33] et les matrices [51, 67] sont très largement utilisés dans l'analyse de l'attention issue de transformeurs. Les visualisations de graphe biparti sont des graphiques utilisés pour comparer deux dimensions, l'une par rapport à l'autre, en utilisant une valeur de mesure comme l'encodage d'une arête entre les valeurs des dimensions. Dans le cas de l'attention, les visualisations de graphe biparti comparent un groupe de mots à un autre (le même le plus souvent) affichant comme liens entre ces deux groupes, les scores d'attention. Le plus souvent, plus ce score est grand, plus l'arête est opaque.

Vig [90] (BERTViz) ou Park et al. [67] (SANVis) proposent des visualisations de l'attention dans chacune des couches dans des transformeurs et permettent une agrégation ou une vision plus fine des têtes d'attention. Clark et al. [14] visualisent dans un espace à deux dimensions les cellules d'attention de BERT et observent les similarités de fonctionnement entre celles-ci. Hao et al. [27] s'intéressent plutôt à l'entraînement de BERT en visualisant la surface d'erreur (error surface) et démontrent que grâce au pré-entraînement et au réglage fin (fine-tuning) de BERT, celui-ci est robuste au sur-apprentissage. Ils démontrent aussi plus globalement l'efficacité d'utiliser BERT pré-entraîné. Wang et al. [93] visualisent l'attention à l'aide d'une disposition radiale des tokens. Ils visualisent aussi les différentes têtes au sein des couches pour montrer leur importance dans la tâche mais aussi la manière dont elles capturent des règles syntaxiques et sémantiques. L'analyse de l'attention ou de l'importance des cellules d'attention peut aussi se faire de manière classique. Voita et al. [91] utilisent par exemple la Layer-Wise Relevance Propagation [4] (*LRP*) pour l'analyse de l'importance des têtes d'attention dans une tâche de traduction. DeRose et al. [19] proposent AttentionFlows, un outil qui explore la façon dont l'attention des modèles auto-attentifs est affinée pendant le réglage fin, et comment l'attention informe les décisions de classification. Pour cela, ils visualisent 12 couches d'attention pour chacun des mots à l'aide

d'une visualisation radiale.
Enfin, pour visualiser les vecteurs de représentation des données (**Obj.4**) les outils s'appuient sur des travaux déjà effectués pour le plongement lexical, comme LMExplorer [76] qui présente les mots dans une réduction en deux dimensions de l'espace de représentation construit par BERT à chaque sortie des couches d'attention. Certains travaux visualisent les matrices d'attention et d'autres informations [51, 67]. Ces travaux entrent aussi dans la catégorie des outils de visualisation avec des informations agrégées (**Obj.5**).
Bien que l'attention soit beaucoup utilisée, il existe un débat autour de son usage comme méthode d'explication [8]. Jain et al. [36] avancent que l'attention n'est pas une méthode d'explication des prédictions en montrant, entre autres, qu'une modification totale des matrices d'attention de manière à ce qu'elle explique la prédiction différemment, n'a pas d'influence sur la sortie du modèle. Wiegreffe et al. [94] estiment dans leurs travaux que les expérimentations de Jain et al. [36] ne permettent pas de soutenir leur théorie. Ethayarajh et al. [22] avancent dans leur étude qu'il existe une grande similarité entre les représentations des mots dans les transformeurs (similarité cosinus). Ceci produit des matrices d'attention pour lesquelles les poids sont uniformément distribués (ceci est d'autant plus vrai pour les couches en entrée du réseau) et elle sont donc peu informatives. C'est ce que montre également les travaux de Clark et al. [14]

3.4 Méthodes agnostiques

Dans cette section, nous présentons des approches agnostiques au type de réseau. En *TAL*, il est parfois nécessaire d'utiliser des auto-encodeurs de manière à pouvoir encoder dans un espace de plus petite dimension, un long texte. Seq2Seq-Vis [82], peut être utilisé pour une tâche de traduction. En effet, l'outil affiche les scores d'attention entre un mot en entrée et un mot en sortie et affiche les espaces de représentation de ces mots mais aussi les prochains mots les plus probables pour chaque nouveau token de la traduction. L'objectif d'un tel outil est d'identifier ce qui a été appris par les auto-encodeurs et de détecter les éventuelles erreurs dans la procédure de décodage de ceux-ci pour ensuite les déboguer ou les corriger.
D'autres méthodes agnostiques permettent de visualiser les parties de la donnée d'entrée qui contribuent à la prédiction à l'aide de carte de chaleur [45, 46, 3]. Dans le cadre de l'utilisation des valeurs de Shapley [77], de SHAP [53, 52] ou de LIME [73], des diagrammes en bâtons sont également utilisés.
Certains travaux montrent l'activation précise de certaines variables des vecteurs de représentation des tokens [45]. Or, ne pouvant pas extraire l'information à propos de ce qu'est censée présenter une dimension, cela ne revêt que peu d'intérêt. Enfin, les explications hiérarchiques sont présentées sous forme d'arbre de décision montrant comment un groupe de tokens influe sur la prédiction à un instant [80, 38, 13] ou comment un modèle de substitution traite l'information à l'aide d'un arbre de décision [24].
Un dernier type de méthode permet de visualiser la frontière de décision. La visualisation de la frontière est dans le domaine de la classification de textes à l'aide de réseaux de neurones intéressante du fait de la grande dimension dans laquelle les classifieurs représentent leurs données. La visualisation de la frontière doit donc adapter ces espaces de représentation en deux ou trois dimensions. Les travaux de Migut et al. [58] ZhiyongYan et Xu [99] proposent des algorithmes pour trouver les données de la frontière de décision. Rodrigues et al. [75] visualisent la frontière de décision des classifieurs en utilisant des techniques de réduction de dimension dans l'espace d'entrée des classifieurs. Parmi les cinq techniques les plus efficaces qu'ils identifient pour visualiser la frontière de décision des *CNN* dans la classification binaire, on retrouve *UMAP* [56] et *t-SNE* [31, 89]. Cependant, l'ensemble de données utilisé dans leur expérience est linéairement séparable, ce qui n'est pas un ensemble de données réaliste et donc cette méthode n'est probablement pas efficace dans un cas réel. Les travaux précédents proposent tous des distances à la frontière qui ne sont pas fidèles à celles dans l'espace de représentation des données, ce qui ne permet pas de comparer les données entre elles en termes de confiance ou force des prédictions. Zhang et al. [96] intègrent des distances fidèles pour comparer deux classifieurs mais toute autre information sur le voisinage des données est perdue.
Dans certain cas, il est pertinent de ne s'intéresser qu'à des sous-parties d'un espace de représentation. Mikolov et al. [59], visualisent, par exemple, le lien sémantique entre pays et ville pour un sous-ensemble de mots. Abadi et al. [1] proposent, dans EmbeddingProjector[4], un échantillon de mots d'un espace de représentation. Melnik et al. [57] analysent la connectivité des données dans l'espace d'entrée. Cela garantit qu'aucune frontière de décision n'existe entre deux données si elles appartiennent aux mêmes régions de décision. Ces régions de décision sont des zones de l'espace de représentation dans lesquelles les prédictions sont toutes identiques. Différentes régions de décision sont calculées et comparées à travers différents classifieurs tels que des classifieurs neuronaux ou des *SVM* [15]. Ramamurthy et al. [71] proposent une méthodologie pour comparer, pour différents modèles, la complexité de la frontière de décision et donc la capacité de généralisation de ces modèles pour un ensemble de données. Enfin, Ma et al. [54] visualisent la décision relative à la frontière de décision en utilisant le *SVM* sur les données proches de la frontière de décision. Ils construisent plusieurs segments linéaires de la frontière de décision avec des mises en lumière de certaines parties de la frontière de décision.

4 Conclusions

Dans cet article, nous avons présenté les problématiques d'interprétabilité, de transparence et d'explicabilité et l'apport des techniques de visualisation de données combinées aux différentes architectures de réseaux de neuronnes. Le tableau 1 présente une synthèse des articles utilisant des

4. https://projector.tensorflow.org/

techniques de visualisation de données pour l'interprétabilité et leur classification dans les classifications de Lipton [49] et de Hohman [32]. Il montre que les techniques de visualisation participent entièrement au processus d'interprétabilité et pas seulement à la visualisation des espaces de représentation ou des parties de la donnée ayant participé à la prédiction.
Parmi les perspectives envisagées, nous pensons que la visualisation de la frontière de décision des réseaux de neurones est essentielle [18] car elle assiste les utilisateurs dans la compréhension des réseaux, leur débogage et la construction d'explications des prédictions. Ce problème est d'autant plus difficile dans le cadre d'une classification multi-classes. De même, la conservation des distances à la frontière dans la visualisation de l'espace de représentation, lui-même de grande dimension, est également un véritable challenge. L'exploration des localités correspondant à des parties de l'espace ajoute à cela la possibilité de comparer entre eux des exemples similaires. Pour une plus grande explicabilité des réseaux de neurones à l'aide de nouvelles visualisations, il est également envisageable de créer de nouvelles métriques d'explicabilité par exemple pour justifier des explications et générer de nouveaux exemples.

Remerciements

Ce travail a été soutenu par des subventions de la Région Occitanie [Programme "Allocation Doctorale 2019"] et du SIRIC Montpellier Cancer [Bourse INCa Inserm DGOS 12553]

Références

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow : Large-scale machine learning on heterogeneous systems, 2015.

[2] Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. Dax : Deep argumentative explanation for neural networks. *arXiv preprint :2012.05766*, 2020.

[3] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark, September 2017. ACL.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klau-schen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7) :1–46, 07 2015.

[5] Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozha-rovskyi, and Jayneel Parekh. Flexible and context-specific ai explainability : a multidisciplinary approach. *SSRN*, 2020.

[6] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. Shaff : Fast and consistent shapley effect estimates via random forests. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5563–5582. PMLR, 28–30 Mar 2022.

[7] Matthew Berger. Visually analyzing contextualized embeddings. In *2020 IEEE Visualization Conference (VIS)*, pages 276–280. IEEE, 2020.

[8] Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation ? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 3889–3900, Dublin, Ireland, May 2022. ACL.

[9] Angie Boggust, Brandon Carter, and Arvind Satyanarayan. Embedding comparator : Visualizing differences in global structure and local neighborhoods via small multiples. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 746–766, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Michael Weigelt. Natural language multitasking : analyzing and improving syntactic saliency of hidden representations. *arXiv preprint :1801.06024*, 2018.

[11] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3) :207–233, 2020.

[12] Piyush Chawla, Subhashis Hazarika, and Han-Wei Shen. Token-wise sentiment decomposition for convnet : Visualizing a sentiment classifier. *Visual Informatics*, 4(2) :132–141, 2020.

[13] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint :2004.02015*, 2020.

[14] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at ? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. ACL.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 :273–297, 1995.

[16] Ian Covert and Su-In Lee. Improving kernelshap : Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR, 2021.

[17] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence : Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016.

[18] A. Delaforge, J. Aze, S. Bringay, C. Mollevi, A. Sallaberry, and M. Servajean. Ebbe-text : Explaining neural networks by exploring text classification decision boundaries. *IEEE Transactions on Visualization & Computer Graphics*, (01) :1–18, jun 5555.

[19] Joseph F DeRose, Jiayao Wang, and Matthew Berger. Attention flows : Analyzing and comparing attention mechanisms in language models. *IEEE TVCG*, 27(2) :1160–1170, 2020.

[20] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint :1702.08608*, 2017.

[21] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7) :620–631, 2021.

[22] Kawin Ethayarajh. How contextual are contextualized word representations ? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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)*, pages 55–65, Hong Kong, China, November 2019. ACL.

[23] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations : An overview of interpretability of machine learning. In *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.

[24] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *ArXiv*, abs/1805.10820, 2018.

[25] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5) :1–42, 2018.

[26] Thilo Hagendorff. The ethics of ai ethics : An evaluation of guidelines. *Minds and Machines*, 30(1) :99–120, 2020.

[27] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of BERT. 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)*, pages 4143–4152, Hong Kong, China, November 2019. ACL.

[28] Adam W Harley. An interactive node-link visualization of convolutional neural networks. In *ISVC*, pages 867–877, 2015.

[29] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. Towards security threats of deep learning systems : A survey. *IEEE Transactions on Software Engineering*, 2020.

[30] Florian Heimerl and Michael Gleicher. Interactive analysis of word vector embeddings. In *Computer Graphics Forum*, volume 37, pages 253–265. Wiley Online Library, 2018.

[31] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, pages 857–864, Cambridge, MA, USA, 2002. MIT Press.

[32] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning : An interrogative survey for the next frontiers. *IEEE TVCG*, 25(8) :2674–2693, 2019.

[33] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT : A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 187–196, Online, July 2020. ACL.

[34] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks : Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37 :100270, 2020.

[35] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems : How should we define and evaluate faithfulness ? In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Online, July 2020. ACL.

[36] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. ACL.

[37] Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations : Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104) :1–54, 2021.

[38] Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. Towards hierarchical importance attribution : Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*, 2020.

[39] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Lingusitic analysis of multi-modal recurrent neural networks. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 8–9, Lisbon, Portugal, September 2015. ACL.

[40] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4) :761–780, 2017.

[41] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456, 2019.

[42] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint :1506.02078*, 2015.

[43] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1) :319–355, 2023.

[44] Alexander LeNail. Nn-svg : Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4(33) :747, 2019.

[45] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 681–691, San Diego, California, June 2016. ACL.

[46] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint :1612.08220*, 2016.

[47] Quan Li, Kristanto Sean Njotoprawiro, Hammad Haleem, Qiaoan Chen, Chris Yi, and Xiaojuan Ma. Embeddingvis : A visual analytics approach to comparative network embedding inspection. In *2018 IEEE VAST*, pages 48–59. IEEE, 2018.

[48] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4 :521–535, 12 2016.

[49] Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3) :31–57, 2018.

[50] Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. Visual exploration of semantic relationships in neural word embeddings. *IEEE TVCG*, 24(1) :553–562, 2018.

[51] Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. Nlize : A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. *IEEE TVCG*, 25(1) :651–660, 2018.

[52] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv :1802.03888*, 2018.

[53] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[54] Yuxin Ma and Ross Maciejewski. Visual analysis of class separations with locally linear segments. *IEEE TVCG*, 27(1) :241–253, 2021.

[55] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp : A survey. *ACM Computing Surveys*, 55(8) :1–42, 2022.

[56] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap : Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29) :861, 2018.

[57] Ofer Melnik. Decision region connectivity analysis : A method for analyzing high-dimensional classifiers. *Machine Learning*, 48(1–3) :321–351, 2002.

[58] M. A. Migut, M. Worring, and C. J. Veenman. Visualizing multi-dimensio-nal decision boundaries in 2d. *21st ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, 29(1) :273–295, 2015.

[59] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[60] Linda Milne. Feature selection using neural networks with contribution measures. In *AI-CONFERENCE-*, pages 571–571. Citeseer, 1995.

[61] Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. Understanding hidden memories of recurrent neural networks. In *2017 IEEE VAST*, pages 13–24. IEEE, 2017.

[62] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11) :501–507, 2019.

[63] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[64] W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance : Contextual decomposition to extract interactions from lstms. *arXiv preprint :1801.05453*, 2018.

[65] W James Murdoch and Arthur Szlam. Automatic rule extraction from long short term memory networks. *arXiv preprint :1702.02540*, 2017.

[66] Duc Hau Nguyen, Guillaume Gravier, and Pascale Sébillot. A study of the plausibility of attention between rnn encoders in natural language inference. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1623–1629, 2021.

[67] Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. Sanvis : Visual analytics for understanding self-attention networks. In *2019 IEEE Visualization Conference (VIS)*, pages 146–150. IEEE, 2019.

[68] Karl Pearson F.R.S. 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.

[69] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings : Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November 2018. ACL.

[70] Peng Qian, Xipeng Qiu, and Xuan-Jing Huang. Analyzing linguistic knowledge in sequential model of sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835, 2016.

[71] Karthikeyan Natesan Ramamurthy, Kush Varshney, and Krishnan Mody. Topological data analysis of decision boundaries with application to model selection. In *36th International Conference on Machine Learning (ICML)*, volume 97, pages 5351–5360. PMLR, 2019.

[72] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32, 2019.

[73] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you ?" explaining the predictions of any classifier. In *22nd ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, pages 1135–1144, 2016.

[74] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors : High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[75] Francisco C. M. Rodrigues, Mateus Espadoto, Roberto Hirata, and Alexandru C. Telea. Constructing and visualizing high-quality classifier decision boundary maps. *Information*, 10(9) :280, 2019.

[76] Rita Sevastjanova, Aikaterini-Lida Kalouli, Christin Beck, Hanna Schäfer, and Mennatallah El-Assady. Explaining contextualization in language models using visual analytics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 464–476, 2021.

[77] Lloyd S Shapley. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953.

[78] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[79] Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients : Feature interaction attribution for neural nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 865–878, 2021.

[80] Chandan Singh, W. James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2019.

[81] Daniel Smilkov, Shan Carter, D. Sculley, Fernanda B. Viégas, and Martin Wattenberg. Direct-manipulation visualization of deep networks. *ArXiv*, abs/1708.03788, 2017.

[82] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. Seq2seq-vis : A visual debugging tool for sequence-to-sequence models. *IEEE TVCG*, 25(1) :353–363, 2018.

[83] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis : A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE TVCG*, 24(1) :667–676, 2017.

[84] Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11 :1–18, 2010.

[85] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3) :647–665, 2014.

[86] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR, 2020.

[87] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[88] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. Natural language processing advancements by deep learning : A survey. *arXiv preprint :2003.01200*, 2020.

[89] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.

[90] Jesse Vig. A multiscale visualization of attention in the transformer model. In *57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations (ACL)*, pages 37–42. ACL, 2019.

[91] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention : Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019. ACL.

[92] Bernhard Waltl and Roland Vogl. Explainable artificial intelligence the new frontier in legal informatics. *Jusletter IT*, 4 :1–10, 2018.

[93] Zijie J. Wang, Robert Turko, and Duen Horng Chau. Dodrio : Exploring Transformer Models with Interactive Visualization. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations (ACL-IJCNLP)*, pages 132–141. ACL, 2021.

[94] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. 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)*, pages 11–20, Hong Kong, China, November 2019. ACL.

[95] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell : Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[96] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S. Ebert. Manifold : A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE TVCG*, 25(1) :364–373, 2019.

[97] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[98] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet : A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

[99] Y. Zhiyong and Xu. Congfu. Using decision boundary to analyze classifiers. In *3rd International Conference on Intelligent System and Knowledge Engineering (ISKE)*, volume 1, pages 302–307, 2008.

Métriques d'équité en Apprentissage Automatique et droit de l'Union Européenne en matière de non-discrimination

M. Legast[*1], Y. Yousefi[*2, 3], L. Koutsoviti Koumeri[*4], A. Legay[1], C. Schommer[3], K. Vanhoof[4]

[1] Université catholique de Louvain, ICTEAM
[2] Università di Bologna, CIRSFID Alma AI
[3] Université du Luxembourg, DP CSCE
[4] Universiteit Hasselt, BINF

Juillet 2023

Résumé

Les modèles d'apprentissage automatique (AA) peuvent présenter des biais discriminatoires envers certains groupes sociaux. Nous étudions à quel point les techniques et définitions d'équité utilisées en AA peuvent garantir le respect du droit de l'UE en matière de non discrimination. À travers des modèles de classification entraînés avec différentes contraintes d'équité, nous évaluons l'efficacité des méthodes de correction de biais et discutons les résultats sous l'angle de l'AA et de l'informatique juridique.

Mots-clés

Apprentissage automatique, Droit de l'Union Européenne, Décision algorithmique, Équité, Non-discrimination

Abstract

Machine Learning (ML) models have been shown to present biases leading to discrimination against certain social groups. Our research studies the extend to which ML techniques and fairness definitions can ensure compliance with the EU non-discrimination legal framework. Using classification models trained with different fairness constraints, we evaluate how effective the bias mitigation process is and discuss the results using both an ML approach and legal informatics methodology.

Keywords

Machine learning, European Union law, Algorithmic decision-making, Fairness, Non-discrimination

[*] All these authors have contributed equally.

1 Introduction

Les décisions algorithmiques prises par des modèles d'Apprentissage Automatique (AA) ont un impact important et croissant sur nos vies. L'implication de l'intelligence artificielle (IA) dans des domaines conséquents amène de nombreuses questions éthiques et légales, telle que la question de l'équité. Il y a notamment un certain nombre d'études qui font état des biais qui peuvent se trouver dans de tels systèmes de décisions et des résultats discriminants qui peuvent en découler [16]. En réaction, un certain nombre de recherches ont été menées dans le domaine de l'équité pour l'IA et l'AA.

Néanmoins, établir une définition de l'équité et garantir des résultats qui ne contiennent pas de biais discriminatoires reste une question ouverte et difficile, que ce soit pour les décisions algorithmiques [20] ou humaines. En informatique, de nombreuses définitions d'équité et de discrimination existent, faisant appel à des notions mathématiques et des concepts éthiques différents. Les différentes métriques d'équité qui en découlent sont utilisées pour donner une évaluation chiffrée du niveau d'équité ou de discrimination d'une BDD ou des prédictions d'un modèle. Plusieurs méthodes pour éviter et réduire les biais ont également été proposées [16]. En droit, la notion d'équité est généralement liée à celles d'égalité et de (non-)discrimination. C'est notamment le cas au niveau de l'Union Européenne (UE) dont le cadre légal est celui qui est pris en considération dans cet article.

Plusieurs analyses de la littérature existante, à la fois au niveau juridique [12] et technique [7], font état d'un écart entre le droit en matière de non-discrimination et la manière dont la recherche en informatique répond à cette question. Ces articles encouragent donc des recherches interdisciplinaires, notamment sur la compatibilité entre les définitions mathématiques et juridiques de l'équité.

Notre recherche répond à ce constat en analysant la concordance entre les méthodes et définitions liées à l'équité dans le domaine de l'AA d'une part et le cadre réglementaire de l'UE en matière de non-discrimination d'autre part. Nous évaluons à quel point les définitions mathématiques d'équité et les métriques et méthodes de correction de biais associées peuvent permettre le respect de ce cadre réglementaire. Nous étudions aussi leur pertinence pour détecter et prouver une discrimination.

Notre méthode de recherche combine des expérimentations avec des modèles de prédiction en classification et l'usage de méthodologie d'informatique juridique, qui interprète et adapte le concept légal d'équité aux paradigmes des nouvelles technologies et inversement [18]. Pour comparer plusieurs scénarios, nous entraînons des modèles avec diffé-

rentes contraintes d'équité. Nous analysons ensuite, à l'aide de plusieurs métriques, le niveau de biais dans les bases de données (BDD) d'entraînement et dans les prédictions émises par ces différents modèles et leurs équivalents non contraints. Nous discutons les résultats obtenus d'un point de vue technique et au regard du cadre réglementaire de l'UE sur la non-discrimination.

2 L'équité, ses approches et enjeux

L'équité est un concept pour lequel il est difficile de donner une définition unique, claire et sans ambiguïté. Cette notion peut être interprétée différemment suivant notamment les contextes, les cultures et les individus [11].

Plusieurs approches co-existent, tant au niveau éthique ou légal qu'en AA. Certaines définitions se complètent et peuvent être appliquées simultanément tandis que d'autres sont incompatibles entre elles [3], [15]. Le choix de la définition d'équité considérée ainsi que la manière de l'implémenter est donc particulièrement important, à la fois dans le développement d'IA et au niveau du cadre réglementaire.

2.1 L'équité dans le droit de l'UE

L'équité dans le droit de l'UE est généralement envisagée via un cadre réglementaire de non-discrimination qui promeut l'égalité. La notion juridique d'équité dans l'UE découle de l'article 21 de la Charte des droits fondamentaux de l'UE et de l'article 14 de la Convention européenne des droits de l'homme. Ces deux textes interdisent les discriminations fondées sur certains attributs sensibles comme l'origine sociale, la religion ou le sexe.

Sous ce cadre réglementaire, les **discriminations directes** et **indirectes** fondées sur des critères protégés sont illégales [1], à moins qu'un objectif légitime, jugé approprié et nécessaire, puisse objectivement et raisonnablement le justifier [17], [22]. Nous utiliserons les termes ***discrimination explicable*** et *attribut explicatif* pour parler d'une différence de traitement avec une telle justification et de la caractéristique qui la justifie, selon la terminologie de [14].

Si les discriminations directes et indirectes sont explicitement abordées, les discriminations fondées sur d'autres critères ou sur la combinaison de plusieurs de ces critères (discriminations intersectionelles [5]) sont moins protégées et sont peu présentes dans la jurisprudence. Ces différentes notions sont autant d'aspects qui doivent être pris en compte pour la réalisation d'IA équitables.

À un autre niveau, on retrouve le principe d'**égalité formelle** qui considère qu'il faut traiter tous les individus de la même manière pour éviter les discriminations. Cette notion est à distinguer de l'**égalité matérielle** (*substantive equality*) qui implique de tenir compte du contexte social et des inégalités historiques pré-existantes pour les corriger et atteindre une égalité effective. D'après la jurisprudence de la Cour de justice de l'Union européenne (CJUE), l'objectif du droit anti-discrimination n'est pas seulement de garantir une égalité formelle, mais aussi d'atteindre une égalité matérielle, ce qui nécessite de tenir compte des différences entre groupes de population [6].

Pour développer des IA qui participent à la réduction des inégalités sociales et soient en cohérence avec le droit de l'UE, ces notions doivent être prises en compte dans l'AA, en particulier dans le choix des métriques d'équité. Dans cette optique, Wachter et al. [21] catégorise ces métriques en deux groupes. Les **métriques conservatrices de biais** (*bias preserving*) reproduisent les performances historiques avec un taux d'erreur par chaque groupe semblable à celui des données d'entraînement (qui contiennent généralement des biais [16]). Les **métriques transformatrices de biais** (*bias transforming*) comparent les taux de résultats favorables entre les différents groupes et prennent en compte les biais sociaux en nécessitant une décision explicite quand aux biais qui devraient être présents dans le système. Ce deuxième type de métrique est d'avantage en adéquation avec le principe d'égalité substantive et donc le droit anti-discrimination de l'UE.

Enfin, nous abordons également le caractère **contextuel** de l'équité dans les décisions juridiques. Cet aspect contextuel implique que chaque situation peut être traitée différemment dans différents contextes et en considérant des éléments différents. Ainsi, la manière dont les tribunaux considèrent l'égalité et l'équité peut varier d'un cas à l'autre. En effet, prendre une décision équitable demande généralement d'effectuer un jugement sur base de multiples facteurs spécifiques à une situation précise [11]. Cette approche contextuelle se retrouve dans la jurisprudence de l'UE [22].

2.2 Enjeux particuliers liés à l'AA

Plusieurs publications, notamment dans le domaine juridique, pointent de nouvelles difficultés pour détecter et comprendre les discriminations algorithmiques par rapport aux cas de discriminations humaines plus évidents [22]. Les discriminations algorithmiques ont souvent un caractère opaque ou intangible [19] et il peut être plus difficile pour les victimes d'avoir un élément de comparaison qui permet de repérer une discrimination [22]. À ceci s'ajoute souvent un manque d'accès aux données et algorithmes utilisés. Détecter ces discriminations et prouver leur occurrence en justice est donc une tâche ardue [10], [23].

En plus de cela, les décisions algorithmiques peuvent être fondées sur des nouveaux attributs ou des catégorisations moins évidentes. Ceci peut mener à des discriminations reposant sur de nouveaux éléments et pas seulement sur les critères légalement protégés. Les IA peuvent donc avoir des comportements discriminatoires injustes, mais pas illégaux, ce qui constitue un frein à l'accomplissement de l'objectif juridique d'égalité contre lequel les législations actuelles ne sont pas nécessairement suffisantes [23].

Enfin, notons que la caractéristique contextuelle de l'équité ajoute une difficulté supplémentaire au développement d'IA équitables et compatibles avec le cadre réglementaire de l'UE. D'après [22], la prise de décisions algorithmiques au niveau légal et éthique doit être conditionnée au fait que les systèmes soient capables de reproduire l'approche juridique d'égalité contextuelle.

2.3 Traitement de l'équité en classification

Dans le domaine de l'AA, de nombreux travaux ont proposé des manières de détecter, mesurer et corriger les discriminations [16]. Dans notre recherche, nous considérons en particulier le problème canonique de la classification. Ce problème consiste en la prédiction de la classe d'une nouvelle observation en utilisant les connaissances apprises à partir d'autres observations pour lesquelles la classe était connue. Le problème de la classification équitable considère un ou plusieurs attribut(s) protégé(s) duquel (ou desquels) les prédictions ne peuvent pas dépendre.

Pour ce problème uniquement, plus de 90 définitions d'équité et méthodes de correction des biais ont été recensées [13]. Il n'existe cependant pas de méthode ni de définition ou métrique qui réponde à tous les contextes et toutes les contraintes d'équité à prendre en considération. Certaines méthodes et métriques peuvent être combinées pour améliorer les résultats, mais d'autres sont incompatibles entre elles. Il est donc important de considérer les spécificités des définitions et des méthodes correspondantes afin de faire le choix le plus approprié selon le contexte.

À ce stade de notre recherche, nous analysons deux définitions d'équité en particulier, qui sont les deux suivantes :

Demographic Parity (DP) correspond à une équivalence entre la probabilité d'obtenir une prédiction $\hat{y}$ favorable pour une personne appartenant au groupe privilégié ($G = priv$) ou au groupe défavorisé ($G = def$) [9].

$$P(\hat{y} = +|G = def) = P(\hat{y} = +|G = priv)$$

Cette définition d'équité est la plus couramment utilisée dans la recherche pour l'équité en classification [13]. Elle est fondée sur le principe d'équité de groupe qui requiert un traitement égal à l'échelle des différents groupes [1] [9].

La différence ou le ratio entre les deux probabilités constituent des métriques d'équité associées à cette définition. Dans le cas de DP, ces métriques permettent de détecter les discriminations directes et indirectes. Il s'agit également de métriques transformatrices de biais [21] qui peuvent donc être utilisées pour mesurer le niveau d'égalité matérielle.

Cette approche qui évalue le traitement différencié global a néanmoins été critiquée, notamment parce qu'elle ne permet pas de différencier entre discrimination illégale et explicable et que son utilisation peut également mener à des discriminations inverses [14] [22].

Conditional Demographic Disparity (CDD) considère que l'équité est atteinte lorsque la proportion de personnes défavorisées ($G = def$) parmi celles qui obtiennent une prédiction $\hat{y}$ favorable équivaut à leur proportion parmi celles qui en obtiennent une défavorable, en considérant un attribut explicatif R [22].

$$P(G = def|\hat{y} = +, R = r) = P(G = def|\hat{y} = -, R = r)$$

La prise en compte d'un tel attribut a pour but de corriger les limitations de DP précitées [14] et de se rapprocher de l'approche contextuelle du droit [22]. Il s'agit également d'une définition fondée sur l'équité de groupe, qui permet de mesurer des discriminations directes et indirectes et dont les métriques associées sont considérées comme transformatrices de biais [21].

1. Ceci le distingue de l'équité individuelle qui requiert un traitement similaire entre individus semblables. [16]

3 Méthodologie

Nous considérons le problème de la classification binaire avec un unique attribut protégé binaire. Notre objectif est d'analyser l'impact de la définitions d'équité considérée sur les résultats, à la fois au niveau de la mesure des biais et de leur correction.

Pour ce faire, nous entraînons et comparons plusieurs modèles de classification entraînés avec une correction de biais imposée via une contrainte sur l'équité. Cette contrainte prend la forme d'une fonction qui mesure le niveau d'équité du modèle et dont la valeur ne peut pas descendre sous un certain seuil. Nous utilisons pour cela le méta-algorithme présenté dans [4] qui permet d'imposer une contrainte appartenant à un large choix de métriques d'équité. Cela nous permet de comparer des modèles de prédiction entraînés avec des contraintes correspondant à différentes définitions, tout en gardant le reste de l'algorithme inchangé. Nous employons l'implémentation disponible via AIF360 [2] qui utilise l'algorithme du gradient (*gradient descent*) pour l'apprentissage.

Nous considérons différents scénarios qui correspondent à la combinaison d'une contrainte d'équité, d'une BDD d'apprentissage et du choix d'un attribut protégé. Cela nous donne par exemple le scénario de la prédiction du risque de récidivisme pour des justiciables (BDD COMPAS [8]) avec une correction du biais racial selon la définition d'équité DP. Pour chacun de ces scénarios, nous créons différents modèles en faisant varier la force de la contrainte (c'est à dire le seuil minimal d'équité imposé) de 0 (pas de correction des biais) à 1 (contrainte correspondant à une équité "parfaite").

Nous mesurons ensuite le niveau de biais avec plusieurs métriques, notamment DP et CDD, pour chacun des modèles ainsi créés et pour les données considérées comme la vérité terrain (*ground truth*). A ce stade, nous avons utilisé DP comme contrainte d'équité et allons inclure des contraintes basées sur d'autres définitions dans le futur.

4 Résultats et contribution attendues

Cette recherche apporte plusieurs contributions. Tout d'abord, nous comparons l'impact de différentes contraintes d'équité et niveau de contraintes sur les performances des modèles, et ce, en considérant plusieurs approches de l'équité.

Nous apportons également une analyse juridique aux résultats d'expérimentations sur la mesure et réduction des biais, qui sont généralement abordés principalement d'un point de vue technique. Nous considérons notamment le choix du niveau de contrainte pour obtenir un bon compromis entre la précision et l'équité du modèle. Nous explorons une approche juridique pour le choix de valeurs seuils

impliquant de tels compromis. Nous tenons compte pour cela de l'approche contextuelle du droit et de l'objectif du cadre réglementaire anti-discrimination de l'UE d'atteindre une égalité matérielle. Cette approche nécessite de ne pas se limiter à l'égalité formelle et à la correction technique des biais, mais également de reconnaître et lutter contre les différences sociales historiques et actuelles entre différents groupes, comme le souligne la jurisprudence de la CJUE.
Enfin, nous suggérons l'introduction de marges concernant les valeurs d'équité et de précision minimales autorisées dans la législation à venir sur l'intelligence artificielle (*Artificial Intelligence Act*). Celles-ci permettraient d'avoir des lignes directrices claires pour le développement des IA. Cela participerait également à la protection des personnes impactées par les discriminations algorithmiques, à la fois en prévenant l'occurrence de telles discriminations et en facilitant le recours à la justice lorsqu'elles se produisent néanmoins.

Références

[1] M. BELL, "The right to equality and non-discrimination," *Economic and Social Rights under the EU Chatter of Fundamental Rights : A Legal Perspective*, p. 91-110, 2003.

[2] R. K. E. BELLAMY, K. DEY, M. HIND et al., *AI Fairness 360 : An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*, 2018.

[3] R. BERK, H. HEIDARI, S. JABBARI, M. KEARNS et A. ROTH, "Fairness in criminal justice risk assessments : The state of the art," *Sociological Methods & Research*, t. 50, n° 1, p. 3-44, 2021.

[4] L. E. CELIS, L. HUANG, V. KESWANI et N. K. VISHNOI, "Classification with fairness constraints : A meta-algorithm with provable guarantees," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, p. 319-328.

[5] K. W. CRENSHAW, "Mapping the margins : Intersectionality, identity politics, and violence against women of color," in *The public nature of private violence*, Routledge, 2013, p. 93-118.

[6] M. DE VOS, "The European Court of Justice and the march towards substantive equality in European Union anti-discrimination law," *International Journal of Discrimination and the Law*, t. 20, n° 1, p. 62-87, 2020.

[7] M. DOLATA, S. FEUERRIEGEL et G. SCHWABE, "A sociotechnical view of algorithmic fairness," *Information Systems Journal*, t. 32, p. 754-818, 2021.

[8] J. DRESSEL et H. FARID, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, t. 4, 2018.

[9] C. DWORK, M. HARDT, T. PITASSI, O. REINGOLD et R. ZEMEL, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, sér. ITCS '12, New York, NY, USA : ACM, 2012, p. 214-226.

[10] P. HACKER, "Teaching fairness to artificial intelligence : existing and novel strategies against algorithmic discrimination under EU law," *Common Market Law Review*, t. 55, n° 4, 2018.

[11] N. HELBERGER, T. ARAUJO et C. H. de VREESE, "Who is the fairest of them all ? Public attitudes and expectations regarding automated decision-making," *Computer Law & Security Review*, t. 39, 2020.

[12] D. HELLMAN, "MEASURING ALGORITHMIC FAIRNESS," *Virginia Law Review*, t. 106, n° 4, p. 811-866, 2020.

[13] M. HORT, Z. CHEN, J. ZHANG, F. SARRO et M. HARMAN, "Bias Mitigation for Machine Learning Classifiers : A Comprehensive Survey," *ArXiv*, 2022.

[14] F. KAMIRAN, I. ŽLIOBAITĖ et T. CALDERS, "Quantifying explainable discrimination and removing illegal discrimination in automated decision making," *Knowledge and information systems*, t. 35, n° 3, p. 613-644, 2013.

[15] J. KLEINBERG, S. MULLAINATHAN et M. RAGHAVAN, "Inherent trade-offs in the fair determination of risk scores," *arXiv*, 2016.

[16] N. MEHRABI, F. MORSTATTER, N. SAXENA, K. LERMAN et A. GALSTYAN, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, t. 54, n° 6, 115 :1-115 :35, 2021.

[17] D. MOECKLI et al., "Equality and non-discrimination," *International human rights law*, p. 189-208, 2010.

[18] G. SARTOR, "Informatica giuridica," *Il diritto nella società dell'informazione*, 2006.

[19] A. SIAPKA, "The Ethical and Legal Challenges of Artificial Intelligence : The EU response to biased and discriminatory AI," *SSRN*, 2018.

[20] S. VERMA et J. RUBIN, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness*, IEEE, ACM, 2018, p. 1-7.

[21] S. WACHTER, B. MITTELSTADT et C. RUSSELL, "Bias preservation in machine learning : the legality of fairness metrics under EU non-discrimination law," *W. Va. L. Rev.*, t. 123, p. 735, 2020.

[22] S. WACHTER, B. MITTELSTADT et C. RUSSELL, "Why fairness cannot be automated : Bridging the gap between EU non-discrimination law and AI," *Computer Law & Security Review*, t. 41, 2021.

[23] Y. YOUSEFI, "Notions of Fairness in Automated Decision Making : An Interdisciplinary Approach to Open Issues," in *EGOVIS 2022, Proceedings*, Vienna, Austria : Springer, 2022, p. 3-17.

Invités

Automating Unsupervised Learning.

Leman Akoglu[1]

[1] Carnegie Mellon University, États-Unis

12 juin 2023

Mots-clés

Automating Unsupervised Learning, Artificial Intelligence

1 Abstract

Learning models are equipped with hyperparameters (HPs) that control their bias-variance trade-off and consequently generalization performance. Thus, carefully tuning these HPs is of utmost importance to learn "good" models. The supervised ML community has focused on Auto-ML toward effective algorithm selection and hyper-parameter optimization (HPO) especially in high dimensions. Yet, automating unsupervised learning remains significantly understudied. In this talk, I will present vignettes of our recent research toward unsupervised model selection, specifically in the context of anomaly detection. Especially with the advent of end-to-end trainable deep learning based models that exhibit a long list of HPs, and the attractiveness of self-supervised learning objectives for unsupervised anomaly detection, I will demonstrate that effective model selection becomes ever so critical, opening up challenges as well as opportunities. Biographie :

2 Biography

Leman Akoglu is the Heinz College Dean's Associate Professor of Information Systems at Carnegie Mellon University. She holds courtesy appointments in the Computer Science Department (CSD) and Machine Learning Department (MLD) of School of Computer Science (SCS). She has also received her Ph.D. from CSD/SCS of Carnegie Mellon University in 2012. Dr. Akoglu's research interests broadly span machine learning and data mining, and specifically graph mining, pattern discovery and anomaly detection, with applications to fraud and event detection in diverse real-world domains. At Heinz, Dr. Akoglu directs the Data Analytics Techniques Algorithms (DATA) Lab.

Éthique computationnelle et Causalité.

Gauvain Bourgne[1]

[1] Sorbonne Université, LIP6

12 juin 2023

Mots-clés

Ethique computationnelle, Intelligence artificielle

1 Résumé

L'éthique computationnelle est un domaine de l'IA qui s'intéresse à la modélisation des raisonnements éthiques et la conception d'agents artificiels adoptant des comportements respectueux des principes éthiques, en se fondant sur les fondements philosophiques explorés dans le champ de l'éthique normative.

La question centrale du raisonnement éthique est la déterminer du Juste : dans un contexte donné, quelles décisions devraient-elles être préférées ou jugées inadmissibles, et sur quelles bases fonder ce jugement. Cette question est abordée par l'éthique normative selon différentes approches, dont les principales sont les approches déontologiques, considérant qu'une action doit être évaluée par rapport à son respect de certaines règles de conduite (définissant une notion de devoir) et les approches conséquentialisme, qui mettent l'accent sur les conséquences réelles de l'action et cherchent à évaluer ce qu'il en ressort afin de maximiser le bien général.

Différents modèles ont été proposés pour modéliser ces différents principes dans divers formalismes et il apparaît que pour traiter de ces approches de façon générale, il est nécessaire de s'appuyer sur une notion de causalité. Que ce soit pour déterminer les conséquences d'une action ou prendre en compte des questions de fins et de moyens, la question de la causalité réelle (actual causality) occupe une place importante dans ces considérations. La causalité réelle s'attache à déterminer les causes effectives d'un événement spécifique (par opposition à la causalité générale qui s'intéresse à la découverte de lois causales). La définition précise de ce qui peut être considéré comme une cause fait encore l'objet de débats, opposant des approches contractuelles (B ce serait-il produit si A n'avait pas eu lieu), mises en difficulté par les cas de surdétermination, et des approches par régularité (considérant des ensembles minimaux de faits permettant différer la conséquence) parfois critiquée sur le fait que la causalité ne se réduit pas à la conséquence logique. Nous proposons dans le contexte de prise de décision une approche fondée sur des langages d'action, différenciant états du monde et évènements, qui permet d'éviter certains de ces écueils.

Cet exposé présente le cadre ACE (Action-Causalité-Ethique) de modélisation de raisonnements éthiques et se concentre ensuite sur le rôle de la causalité dans ces raisonnements et sur la définition d'une notion de causalité adaptée à nos problématiques sur la base d'un langage d'action suffisamment expressif pour représenter les situations complexes qui interviennent dans les différents exemples et dilemmes mis en avant dans ces domaines.

Ethique computationnelle et Causalité [invité CNIA].