8ème édition de l'atelier "Fouille de données complexes" : complexité liée aux données multiples

Guillaume Cleuziou, Cyril de Runz, Mustapha Lebbah, Cédric Wemmert

Extraction et Gestion des Connaissances

Brest, 25 janvier 2011

8ème Atelier sur la Fouille de Données Complexes

Complexité liée aux données multiples

Responsables

Guillaume Cleuziou (LIFO, Université d'Orléans)

Cyril de Runz (CreSTIC, Université de Reims)

Mustapha Lebbah (LIPN, Université Paris 13)

Cédric Wemmert (LSIIT, Université de Strasbourg)

8ème Atelier sur la Fouille de Données Complexes Complexité liée aux données multiples

Guillaume Cleuziou*, Cyril de Runz**, Mustapha Lebbah***, Cédric Wemmert****

*LIFO, Université d'Orléans
guillaume.cleuziou@univ-orleans.fr
**CReSTIC, Université de Reims
cyril.de-runz@univ-reims.fr
***LIPN, Université Paris 13
mustapha.lebbah@univ-paris13.fr
****LSIIT, Université de Strasbourg
wemmert@unistra.fr

Résumé. L'atelier sur la fouille de données complexes est proposé à l'instigation du groupe de travail EGC "Fouille de données complexes". Chaque année les organisateurs proposent une thématique de recherche qui suscite l'intérêt des chercheurs et des industriels.

La précédente édition de l'atelier ayant permis de fédérer avec succès des recherches portant sur la problématique de complexité liée à la multiplicité des données, nous avons choisit de renouveler l'affichage de cette priorité thématique pour cette huitième édition.

1 Le groupe de travail "Fouille de Données Complexes"

La huitième édition de l'atelier sur la fouille de données complexes est organisée par le groupe de travail EGC "Fouille de Données Complexes". Ce groupe de travail rassemble une communauté de chercheurs et d'industriels désireux de partager leurs expériences et problématiques dans le domaine de la fouille de données complexes telles que le sont les données non-structurées (ou faiblement), les données obtenues à partir de plusieurs sources d'information ou plus généralement les données spécifiques à certains domaines d'application et nécessitant un processus d'extraction de connaissance sortant des itinéraires usuels de traitement.

Les activités du groupe de travail s'articulent autour de trois champs d'action progressifs :

- l'organisation de **journées scientifiques** une fois par an (vers le mois de juin) où sont présentés des travaux en cours ou plus simplement des problématiques ouvertes et pendant lesquelles une large place est faite aux doctorants,
- l'organisation de l'**atelier "Fouille de Données Complexes"** associé à la conférence EGC qui offre une tribune d'expression pour des travaux plus avancés et sélectionnés sur la base d'articles scientifiques par un comité scientifique constitué pour l'occasion,

– la préparation de **numéros spéciaux** de revue nationale, dans lesquels pourront être publiés les études abouties présentées dans un format long et évaluées plus en profondeur par un comité scientifique. Le second et prochain numéro spécial est actuellement finalisé, il devrait être publié début 2011.

2 Contenu scientifique de l'atelier

Nous avons reçu cette année 16 propositions, chacune d'elle a été relue par au moins deux rapporteurs. Pour la grande majorité des propositions nous avons été en mesure de proposer trois rapports d'experts afin d'offrir un processus scientifique constructif aux auteurs. Nous avons retenu 12 propositions en fonction de leur degré d'avancement ainsi que de leur cohérence avec la thématique de l'atelier.

Les articles qui vous sont proposés explorent :

– de nouvelles **problématiques** de recherche : représentation d'objets complexes, classification faiblement supervisée, évaluation des la qualité des sources d'information.
– de nouvelles **méthodologies** d'extraction de connaissance : nouvelles méthodes de classification, nouvelles mesures de similarité, nouveaux formalismes de fusion d'information.
– des **domaines d'application** variés : bioinformatique, analyse d'images sonar ou satellite, études de spécimens biologiques, vision par ordinateur, études socio-économiques, domaine médical, etc.

La complexité est abordée dans ces travaux principalement au niveau de la nature des données. La masse de données à traiter, la multiplicité des sources d'information, l'incertitude ou l'imprécision sur les données, l'absence partielle de descriptions ou encore la temporalité dans les données sont autant de niveaux de complexité qu'il peut être utile d'apprivoiser pour répondre aux besoins actuels en matière d'extraction de connaissances.

3 Comité de programme

- Hanane Azzag (LIPN, Université Paris 13)
- Boutheina Ben Yaghlane (LARODEC, IHEC Carthage)
- Khalid Benabdeslem (LIESP, Université de Lyon 1)
- Frédéric Blanchard (CreSTIC, Université de Reims Champagne-Ardenne)
- Omar Boussaïd (ERIC, Université de Lyon 2)
- Hend Bouziri (LARODEC-ISG, Tunisie)
- Martine Cadot (LORIA, Nancy)
- Guillaume Cleuziou (LIFO, Université d'Orléans)
- Sylvie Despres (LIM&BIO, Université Paris 13)
- Cyril De Runz (CReSTIC, Université de Reims Champagne-Ardenne)
- Gaël Harry Dias (University of Beira Interior, Portugal)
- Mounir Dhibi (ISSAT, Gafsa)
- Zied Elouedi (ISG, Université de Tunis)
- Rim Faiz (IHEC, Université de 7 Novembre de Carthage)
- Sami Faiz (INSAT, Université de 7 Novembre de Carthage)
- Pierre Gançarski (LSIIT-AFD, Université de Strasbourg)
- Lamia Hadrich (Laboratoire MIRACL, Université de Sfax)
- Mustapha Lebbah (LIPN, Université Paris 13)
- Eric Lefèvre (LGI2A, Université d'Artois)
- Arnaud Martin (IRISA, Université Rennes 1)
- Florent Masséglia (AxIS-Inria Sophia Antipolis)
- Christophe Osswald (ENSIETA, Brest)
- Sébastien Régis (LAMIA, Université des Antilles et de la Guyane)
- Brigitte Trousse (AxIS-Inria Sophia Antipolis)
- Cédric Wemmert (LSIIT-AFD, Univesité de Strasbourg)
- Djamel Zighed (ERIC, Université de Lyon 2)

4 Remerciements

Nous tenons à remercier les auteurs pour la qualité de leurs contributions, les membres du comité de programme et plus généralement tous les relecteurs de cet atelier pour le travail accompli et pour la qualité de leurs prestations.

Nous remercions également les responsables des ateliers pour EGC 2011, A. Martin, A.O. Boudraa et M. Benbouzid.

Enfin nous remercions vivement les présidents : Pascal Poncelet, président du comité de programme et Ali Khenchaf, président du comité d organisation d EGC 2011.

5 Programme

8h45 Accueil

Classification

9h00 Fouille vidéo orientée objet, une approche générique
J. Weber, S. Lefèvre et P. Gancarski

9h30 Classification non supervisée de données satellites multirésolution
C. Kurtz

10h Intérêt des distributions alpha-stables pour la classification de données complexes
A. Fiche, J-C. Cexus, A. Martin et A. Khenchaf

10h30-11h Pause

Fusion

11h Apprentissage itératif pour une connaissance a priori des labels
R. Lefort, R. Fablet et J-M. Boucher

11h30 Utilisation de méthodes d'évaluation de sources d'information dans le cadre de la théorie des fonctions de croyance pour une application réelle
S. Régis, J. Frominville, A. Doncescu et M. Collard

12h Stratégie de fusion d'informations exploitant le réseau des sources
T. Bärecke, M-J. Lesot, H. Akdag et B. Bouchon-Meunier

12h30-14h Pause déjeuner

Similarités, métriques et applications

14h	Représentativité et graphe de représentants : une approche inspirée de la théorie du choix social pour la fouille de données relationnelles *F. Blanchard, C. de Runz, M. Herbin et H. Akdag*
14h30	Représentation et classification d'objets biologiques complexes *H. Ralambondrainy, D. Grosser et N. Conruyt*
15h00	Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets : Application to cancer expression data *S. Benabderrahmane, M-D. Devignes, M. Smail-Tabbone, O. Poch, A. Napoli, W. Raffelsberger, N.H. Nguyen, D. Guenot and E. Guerin*
15h30	Un modèle génératif pour la comparaison de métriques en classification de profils d'expression de gènes *A. Diallo, A. Douzal et F. Giroud*
16h00	Extraction de connaissances agronomiques par fouille de données de l'évolution temporelle des voisinages entre occupations du sol *L. El Ghali, N. Schaller et J-F. Mari*

16h30-16h45 Pause

Entrepôt

16h45	Complexité liée à la variabilité sémantique des statistiques socio-économiques *C. Plumejeaud et J. Gensel*
17h15	Discussion sur l'avenir de l'atelier
17h30	Clôture

Summary

The workshop on mining complex is done at the incitement of the work group EGC "Complex data mining". Each year the organizers propose a topic that interests the researchers and companies.

Last year the workshop focused with success on the complexity associated with multiple data. Indeed we decided to keep this research field as a priority for the present edition of the workshop.

Fouille vidéo orientée objet, une approche générique

Jonathan Weber*, Sébastien Lefèvre**
Pierre Gançarski*

*Université de Strasbourg
j.weber@unistra.fr,gancarski@unistra.fr
https://lsiit.u-strasbg.fr/
**Université Bretagne-Sud
sebastien.lefevre@univ-ubs.fr
http://www-valoria.univ-ubs.fr/

Résumé. Actuellement, le média vidéo est une des premières sources d'information, mais aussi une des plus volumineuses. Pour traiter cette masse d'informations, les systèmes actuels de fouille vidéo font face à un problème de fossé sémantique : il existe une différence entre la signification sémantique du contenu des séquences vidéos et l'information numérique codée dans les fichiers associés. Ce fossé peut être en partie comblé par l'utilisation des objets réels (du point de vue de l'utilisateur) présents dans les séquences. Cependant la fouille vidéo orientée objet nécessite l'introduction d'informations sémantiques, que ce soit pour l'extraction des objets ou pour la fouille de ces objets. Nous proposons d'introduire de telles informations par le biais d'une interaction avec l'utilisateur. Cette interaction consiste en un mécanisme de retour de pertinence. Le système propose à l'utilisateur un échantillon des résultats obtenus, puis l'utilisateur valide, invalide ou corrige ces résultats. Ces informations de validation/invalidation/correction sont alors utilisées pour guider le système et lui permettre d'améliorer les résultats qu'il produit. Cet article ne propose pas un système complètement opérationnel mais explore certaines pistes pour arriver à un tel système.

1 Introduction

Après l'augmentation massive des données textuelles, et plus récemment des images, disponibles dans des bases de données et sur le Web, nous observons aujourd'hui une augmentation dans le domaine de la vidéo. La vidéo est en train de devenir une des principales sources d'informations. L'aspect temporel des vidéos empêche un parcours rapide et efficace de telles bases de données. Cependant, l'aspect temporel est peu utilisé dans les algorithmes existants liés à la fouille vidéo, sauf dans le cas de la segmentation en plans où l'information temporelle joue un rôle presque exclusif (Koprinska et Carrato, 2001). La fouille de données (Cios et al., 2007) est le processus d'extraction d'informations et de connaissances d'une masse de données. La fouille vidéo (Rosenfeld et al., 2002) est donc l'application de ce processus à des données vidéo, c'est-à-dire des séquences temporelles d'images, éventuellement couplées à

des données audio. Cependant, dans cet article, nous considérerons uniquement les données visuelles. Selon ces définitions, un Système de Fouille Vidéo est un système capable d'extraire de l'information à partir d'une grande base de séquences vidéo.

Plusieurs auteurs se sont essayés à dresser un état de l'art relatif à la fouille vidéo. Cependant, aucun de ces travaux ne présente l'ensemble du domaine de la fouille vidéo, chaque contribution concernant plutôt un sous-domaine (par exemple l'indexation). Parmi les premiers travaux, Idris et Panchanathan (1997) présentent quelques méthodes d'indexation vidéo et précisent que le niveau normal pour analyser le contenu visuel devrait être l'objet. Brunelli et al. (1999) présentent également des systèmes d'indexation vidéo et notent, en 1999, que la détection des objets génériques ne peut pas être réalisée avec les méthodes actuelles. Dix ans plus tard, Brezeale et Cook (2008) étudient le niveau auquel les informations font l'objet d'une classification dans le domaine vidéo : la plupart des méthodes étudiées proposent de travailler au niveau global, quelques-unes utilisent le plan ou la scène et, surtout, aucune n'utilise l'objet. Money et Agius (2008) ont étudié les systèmes de résumé de vidéo. Ils proposent d'utiliser des informations propres à l'utilisateur pour produire des résumés personnalisés et plus riches sémantiquement. Ren et al. (2009) se sont concentrés sur l'utilisation d'informations spatio-temporelles pour la recherche de vidéos. Ils remarquent l'efficacité de l'utilisation des relations spatio-temporelles entre les objets pour résoudre le problème de la recherche de vidéos. Snoek et Worring (2009) présente la recherche de vidéo basée sur des concepts via une étude de 300 articles. Les auteurs insistent sur l'importance de l'efficacité computationnelle des méthodes et de disposer d'un très large panel de détecteurs de concepts permettant d'aborder la diversité des contenus vidéos. Contrairement à ces études, nous ne nous concentrons pas sur un objectif spécifique de fouille vidéo, mais nous souhaitons prendre en considération tous les objectifs possibles. Comme Idris et Panchanathan (1997), nous pensons que le niveau de l'objet est le plus adapté pour la fouille vidéo. Dans ce document, nous nous concentrons sur le rôle de l'objet dans le processus de fouille de données vidéo et discutons de la position de l'utilisateur comme élément fondamental du processus visuel d'exploitation.

Dans cet article, nous introduisons en premier lieu une nouvelle taxonomie pour caractériser les différents systèmes de fouille de données vidéo (SFV). Puis, nous étudions des SFV récents en utilisant la taxonomie introduite précédement. Nous déterminons ensuite les caractéristiques d'un SFV orienté objet, présentons le problème de l'extraction d'objets au sein de vidéo et étudions l'introduction de la sémantique au sein d'un tel système. Enfin, nous proposons un cadre générique qui permettra la création de SFV orienté objet.

2 Caractéristiques des SFV

De nombreux aspects sont à prendre en compte lorsque l'on conçoit un nouveau SFV. Dans cette section, nous allons identifier ces aspects et introduire certains termes qui peuvent être utilisés pour caractériser les SFV.

2.1 Objectifs des SFV

Les objectifs accomplis par un SFV sont variés et dépendent des besoins de ses utilisateurs. Les bases de vidéo nécessitent de grandes capacités de stockage et la fouille manuelle de ces bases est fastidieuse. Des SFV ont donc été développés pour accomplir de façon automatique

les tâches jusqu'alors accomplies par des êtres humains. Le *résumé de vidéo* (Res) vise à produire de courts et représentatifs extraits de vidéo dans le but de permettre aux utilisateurs de retrouver leurs thèmes et leurs contenus sans avoir à regarder la vidéo en entier. L'*indexation de vidéo* (Ind) est la caractérisation d'une vidéo afin d'être capable de la retrouver rapidement ultérieurement en utilisant des requêtes spécifiques. La *classification de vidéo* (Cla) vise à regrouper les vidéos dans des catégories prédéfinies afin d'identifier leur contenu. La *recherche basée sur le contenu* (Rec) permet aux utilisateurs de retrouver des vidéos similaires à une autre vidéo donnée en requête.

2.2 Propriétés des SFV

Les SFV sont caractérisées par différentes propriétés relatives à la nature des données et des informations à traiter, aux descripteurs à extraire et à l'échelle à laquelle les calculer, et au rôle de l'utilisateur. Dans cette section, nous introduisons et décrivons ces différentes propriétés.

Données

Un SFV peut avoir à traiter différents types de données vidéo. Une vidéo peut être compressée (C) par différents algorithmes ou disponible sous forme brute (B). La compression permet un stockage moins couteux en mémoire mais requiert un processus de décompression avant une visualisation et peut induire une perte d'information. Traiter des vidéos compressées est plus rapide car le volume de données à traiter est moindre, cependant l'extraction de concepts visuels est plus complexe, alors que l'analyse du contenu visuel peut être effectué directement dans le cadre de vidéos brutes (avec un cout de calcul plus élevé). Les SFV peuvent également être dédiés au traitement de types de vidéos spécifiques (S) ou génériques (G). Traiter des vidéos spécifiques permet d'obtenir de meilleurs résultats car l'on peut utiliser les connaissances du domaine dans le processus de fouille. Par contre, la prise en compte de vidéos génériques par un SFV facilite la réutilisation et l'adaptation de ce dernier à des contextes variés.

Éléments

Quel que soit l'objectif d'un SFV, il peut être suivi en considérant différentes éléments, de la vidéo dans son intégralité au simple pixel. La première étape en fouille vidéo consiste généralement à extraire l'élément à traiter. La *vidéo intégrale* (Vid) est l'élément classique et le plus simple à traiter car il ne nécessite aucune extraction. Néanmoins, si la vidéo contient des scènes très différentes, cet élément peut ne pas être très significatif. Une *scène* (Sce) est composée de plusieurs plans dans un contexte spatio-temporel identique et peut être difficile à extraire. Un *plan* (Pla) est un segment de vidéo délimité par deux transitions, le problème de son extraction est un sujet très étudié et de nombreuses méthodes ont été proposées pour le résoudre (Lefèvre et al., 2003). La *trame* (Tra) est l'unité temporelle d'une vidéo, une vidéo étant une séquence temporelle de trames. Un *objet* (Obj) est, selon nous, l'élément qui comporte le plus de sémantique mais son utilisation est limitée par la dificulté d'extraire un objet réel à un instant donné, ou au cours du temps. Une *région* (Reg) est un ensemble de pixels connexes qui (au contraire de l'objet) ne repose pas sur un concept sémantique. Enfin, le *pixel (Pix)* est le

plus petit élément et, pris isolément, il n'apporte que peu voire pas d'information. La figure 1 représente les différents éléments.

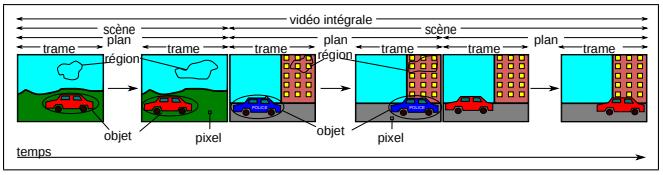


FIG. 1 – *Les différents éléments d'un SFV.*

Descripteurs

Il existe de nombreux descripteurs pour décrire le contenu d'une vidéo. Il est possible de décrire le *mouvement* (Mou) mais également la *couleur* (Coul). Il existe des moyens de caractériser une *texture* (Tex) présente dans les données visuelles. Nous pouvons abstraire les formes et les contours (For) afin de décrire la morphologie des éléments présents dans la vidéo. Outre ces descripteurs classiques, il existe de nombreux autres descripteurs spécifiques régulièrement proposés dans la littérature (Rui et al., 1999). Choisir le descripteur le plus adapté à un SFV précis n'est pas trivial car chaque descripteur vise à caractériser un contenu vidéo selon un point de vue particulier.

Échelles

Afin d'analyser et caractériser les éléments par les différents descripteurs, il est nécessaire de choisir l'échelle à laquelle les descripteurs vont être calculés sur les éléments. L'échelle est liée en partie à l'élément utilisé et au descripteur choisi. À une échelle *globale* (Glo), les descripteurs vidéos sont appliqués sur l'intégralité de la vidéo. Considérant l'échelle du *bloc* (Blo), la vidéo est divisée en blocs suivant une grille spatiale, les descripteurs sont alors calculés dans chaque bloc indépendamment. A la différence de l'échelle bloc, l'échelle *region* ne divise pas la vidéo en blocs mais en régions de tailles et formes variées par une étape de segmentation. Les descripteurs sont ensuite associés à chaque région indépendamment. L'échelle *objet* (Obj) consiste à définir les descripteurs pour les objets réels présents dans les éléments de la vidéo. L'échelle *point d'intérêts* consiste à calculer les descripteurs sur des points (et leur voisinage) dont le voisinage est particulier. Enfin l'échelle *pixel* (Pix) est la plus petite possible : les descripteurs ne servant alors qu'à décrire un pixel, cette échelle ne semble pas très utile. La figure 2 représente ces notions.

2.3 Implication de l'utilisateur

L'implication de l'utilisateur est un point critique dans un SFV. Il y a quatre niveaux d'implication possibles pour un utilisateur. Celle-ci peut être *Nulle* (Nul) si le système est totalement automatique et que l'utilisateur n'intervient pas dans le processus. Elle est *Supervisée* (Sup) lorsque l'utilisateur doit fournir un ensemble complet de données étiquetées afin

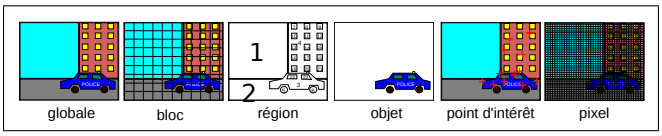


FIG. 2 – *Les différentes échelles de descripteurs possibles pour un SFV.*

de configurer le système pour traiter un jeu de données spécifiques. L'implication est dite *Semi-supervisée* (S-sup) quand l'utilisateur doit fournir moins de données étiquetées et/ou doit valider/invalider certains résultats pour guider le processus de fouille. Enfin, l'implication est appelée *Paramétrique* (Param) lorsque l'utilisateur doit fixer différents paramètres du système.

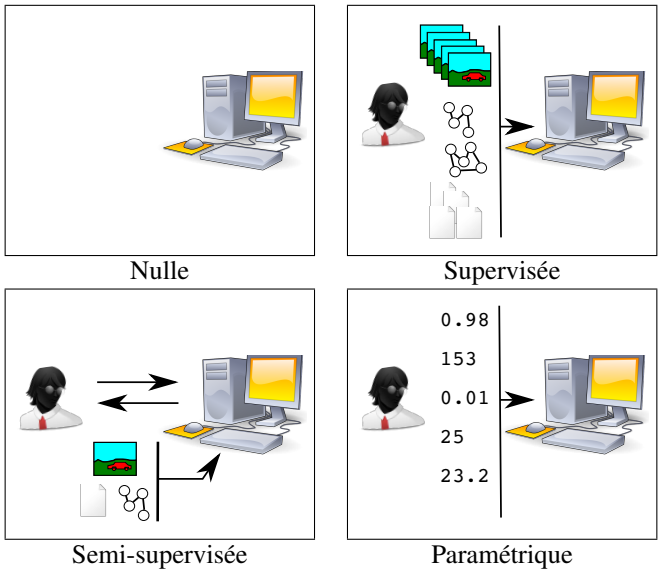


FIG. 3 – *Les différentes implications possibles de l'utilisateur dans un SFV.*

La taxonomie que nous avons introduite dans cette section permet de décrire et caractériser un SFV. Nous allons l'illustrer dans la section suivante en caractérisant les principaux SFV de la littérature.

3 Utilisation de l'objet dans la fouille vidéo

Dans cette section, nous caractérisons des travaux récents concernant de près ou de loin la fouille vidéo orientée-objet, afin de mettre en lumière les tendances actuelles dans ce domaine. Le tableau 1 résume les différentes caractéristiques des systèmes étudiés, selon la taxonomie introduite dans la section 2.

Méthodes	Tâches	Données	Elément	Descripteur	Echelle	Implication
Anjulan et Canagarajah (2007b)	Rec	B,G	Obj	LIR,SIFT	Reg	Param
Anjulan et Canagarajah (2007a)	Cla	B,G	Obj	LIR,SIFT	Reg	Param
de Avila et al. (2008)	Res	B,G	Vid	Col,LP	Glo	Param
Basharat et al. (2008)	Rec	B,G	Vid	SIFT,Col,Tex,Mot	Reg	Nul
Chevalier et al. (2007)	Rec	C,G	Obj	RAG	Reg	Param
Gao et al. (2009)	Rec	B,G	Pla	OFT	Blo	Param
Liu et Chen (2009)	Rec	B,G	Vid	Diverse	Obj	Param
Ren et Zhu (2008)	Res	B,G	Vid	PLR,ECR,HCC	Glo	Param
Sivic et Zisserman (2008)	Rec	B,G	Obj	SIFT	Reg	Param
Teixeira et Corte-Real (2009)	Cla	B,S	Obj	SIFT	Obj	Sup
Zhai et al. (2007)	Res	B,G	Vid	KNNG	Glo	Param

TAB. 1 – *Caractérisation des approches récentes en fouille vidéo.*

Ces articles récents traitent majoritairement des vidéos génériques non-compressées. La recherche est l'objectif le plus fréquent. En effet, la demande principale d'un utilisateur de SFV est certainement de retrouver les vidéos dont il a besoin, surtout dans le cas où celles-ci sont noyées dans une grande quantité de données. Le résumé de vidéo suscite également l'intérêt de la communauté, puisqu'il a pour but de permettre à l'utilisateur de connaître le contenu d'une vidéo sans avoir à la regarder dans son intégralité, ce qui se traduit par un gain de temps important. Mis à part dans le cas du résumé de vidéo, l'élément le plus courant semble être l'objet. Cependant, l'objet est loin d'être l'échelle la plus utilisée, les échelles globale et région sont les plus communes : en effet, produire une segmentation sémantique (en objets) de façon automatique reste aujourd'hui encore un problème ouvert. Les descripteurs utilisés sont variés et souvent combinés afin d'obtenir de meilleurs résultats. Enfin, la grande majorité des SFV demande à l'utilisateur de fixer des paramètres, ce qui est une tâche relativement peu intuitive. Notons qu'un SFV est supervisé tandis qu'un autre est complètement automatique. Aucun des SFV étudiés n'implique l'utilisateur de façon semi-supervisée, ce qui nous semble pourtant un moyen fiable et léger pour guider le système.

4 Vers une fouille vidéo orientée-objet

L'étude des tendances récentes dans la fouille vidéo montre que si les échelles objet et région semblent être adoptées, la vidéo intégrale et les plans sont toujours les éléments les plus couramment traités (excepté pour la recherche où de nombreuses méthodes utilisent l'objet comme élément de base). Pourtant, dans le contexte de l'analyse vidéo, les informations sont apportées principalement au travers des objets et de leur évolution temporelle. En exploitant l'environnement des objets, comme le fond ou les objets adjacents, il est également possible d'introduire une certaine sémantique. De la même façon, les relations spatio-temporelles entre les objets peuvent être utilisées pour enrichir le processus de fouille. Si la fouille vidéo orientée-objet semble pertinente, elle pose cependant le problème de l'extraction des objets qui n'est pas possible sans introduction de sémantique. Notons que cette démarche peut également s'appliquer au processus de fouille afin d'exploiter au mieux les objets. Dans cette section, nous expliquons dans quelle mesure les caractéristiques d'un Système de Fouille Vidéo Orienté-Objet (SFV orienté objet) sont différentes d'un SFV n'étant pas centré sur l'objet.

4.1 Caractéristiques d'un SFV orienté objet

Choisir l'objet comme élément d'un SFV a une influence importante sur les autre caractéristiques (sauf pour les objectifs car a priori ils peuvent tous être accomplis en s'appuyant sur l'objet comme élément).

Données

L'impact sur le type de données est relativement faible. Même s'il n'est pas trivial d'extraire des objets depuis un flux vidéo compressé, des solutions existent, citons notamment les travaux de Babu et al. (2004), Toreyin et al. (2005) et Hsu et al. (2006). De plus, l'approche orientée-objet est adaptée à n'importe quel type de vidéos. Mais, intuitivement, il semble plus simple de traiter des vidéos spécifiques puisque la variété d'objets considérés sera plus limitée. Au contraire, utiliser l'objet comme élément rend plus difficile le traitement de vidéos génériques en l'absence de méthode d'extraction adaptée à tout type d'objet.

Échelle

L'approche orientée-objet entraîne un changement d'échelle. En effet, l'utilisation de l'objet comme élément nous amène à considérer deux types d'échelles complémentaires, une échelle pour l'objet et une échelle pour le contexte, tel qu'illustré en figure 4. Les deux trames présentées comportent le même objet, la navette spatiale *Discovery*. Dans le cadre d'une base contenant de nombreuses vidéos de cette navette spatiale, l'utilisateur pourrait vouloir différencier les différentes situations dans laquelle se trouve cette navette (par exemple celles présentées dans les deux vidéos). Décrire seulement l'objet, qui est ici identique dans les deux vidéos, ne permettrait pas de les distinguer. Il est donc nécessaire de pouvoir également décrire l'environnement propre à l'objet afin de pouvoir distinguer la navette sur sa rampe de lancement de la navette en pleine ascension.

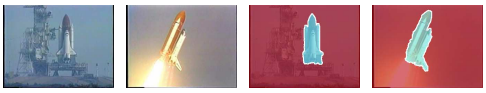


FIG. 4 – *Deux trames extraites de la séquence vidéo STS-53 Launch and Landing, segment 02 of 5 de The Open Video Project (2010) (gauche) et leurs segmentations respectives (droite) de la navette Discovery (bleu) et de son environnement (rouge).*

Descripteurs

Tous les descripteurs peuvent être considérés dans un SFV orienté objet, mais leur usage est différent de celui suivi par les autres types de SFV. En effet, ils peuvent être utilisés pour décrire l'objet et/ou son environnement, et non plus seulement pour décrire l'élément. En outre, les descripteurs de mouvement peuvent être exploités pour décrire le mouvement général de l'objet mais aussi son mouvement interne dans le cas d'objet complexes. De façon plus générale, les descripteurs utilisés dans une approche objet doivent être sémantiquement discriminants, c'est-à-dire que la différence ou la similarité qu'ils mettent en valeur doit avoir une signification sémantique (par exemple la couleur).

Implication de l'utilisateur

Dans un SFV orienté objet, le rôle de l'utilisateur est prédominant de par la sémantique associée au concept d'objet. Un système totalement automatique ne sera pas capable de fouiller sémantiquement les objets, et aura besoin des connaissances de l'utilisateur. De plus, la perception d'un objet est subjective et peut donc être différente d'un utilisateur à l'autre. En effet, chaque utilisateur désire obtenir un résultat personnalisé. L'utilisateur doit donc être particulièrement impliqué dans le processus de fouille afin de pouvoir guider ce dernier. Cependant, même si cette intervention est fondamentale pour le SFV orienté objet, elle doit rester intuitive et légère afin d'être efficace et peu couteuse en temps. Ces propriétés peuvent être assurées au travers de la mise en place d'un retour de pertinence, tel que présenté dans la section 4.3.

4.2 Extraction des objets

Afin d'extraire les objets d'une vidéo, la plupart des méthodes intègrent une étape de segmentation. Seules font exception les méthodes dédiées aux vidéos compressées selon un schéma orientée-objet, voire celles basées sur des points d'intérêt. Pour les SFV, la segmentation vidéo consiste la plupart du temps en un découpage en plans (Lefèvre et al., 2003). Au contraire, pour les SFV orienté objet, l'étape d'extraction doit produire des objets et décrire leur évolution temporelle. Comme nous l'avons souligné dans la section 1, la principale difficulté rencontrée ici est de combler le fossé sémantique séparant les données brutes des objets. Cette extraction peut être effectuée pendant la phase d'encodage de la vidéo dans le cas des données compressés, ou plus généralement avec une segmentation.

Nous représentons une vidéo dans un espace tri-dimensionnel (X,Y,T). une segmentation spatio-temporelle est une partition de cet espace en volumes, chacun représentant un objet spatio-temporel (ou, autrement dit, la définition spatiale de cet objet couplée à son évolution temporelle). Puisqu'un objet est supposé posséder une sémantique, la segmentation nécessite des méthodes intégrant de telles informations. Plus généralement, l'introduction de sémantique est un point important des SFV orienté objet que nous détaillons dans la section suivante.

4.3 Introduire de la sémantique

Un SFV orienté objet nécessite d'introduire de la sémantique dans le processus de segmentation ainsi que dans le processus de fouille. Les descripteurs bas-niveau présentés dans la section 2.2 fournissent des représentations numériques mais ne sont pas capables de donner une perception sémantique de l'objet comme le font les êtres humains. Le fossé sémantique n'est donc toujours pas comblé. Nous pensons néanmoins qu'il puisse l'être en introduisant des connaissances humaines dans un SFV orienté objet.

Pour ce faire, il serait possible de fournir des exemples pour chaque objet potentiellement présent dans une vidéo, mais cette approche ne serait évidemment pas réaliste. Exploiter un mécanisme tel que le retour de pertinence (Ruthven et Lalmas, 2003) semble être une solution plus pertinente. À l'issue du processus de fouille, l'utilisateur évalue un échantillon du résultat qu'il peut également corriger si nécessaire (à l'instar d'un apprentissage par renforcement). En fonction de cette évaluation, le processus peut être relancé pour tenir compte de la connaissance introduite par l'utilisateur (via l'évaluation et la correction de l'échantillon). Ce processus itératif est moins couteux en temps que la production d'exemples complets nécessaire dans

une approche supervisée. Cela garantit également la personnalisation du résultat. De plus, le retour de pertinence peut également être appliqué à l'étape de segmentation (dans le but, ici aussi, de l'améliorer) selon le principe suivant : meilleure sera la segmentation, plus facile la fouille sera. Finalement, créer des descripteurs dédiés aux objets considérés pourrait être également une solution intéressante mais ce problème reste ouvert dans le contexte de vidéos génériques.

4.4 VOMF : Video Object Mining Framework

Nous proposons dans cette section un cadre générique pour les SFV orienté objet, VOMF (Video Object Mining Framework).

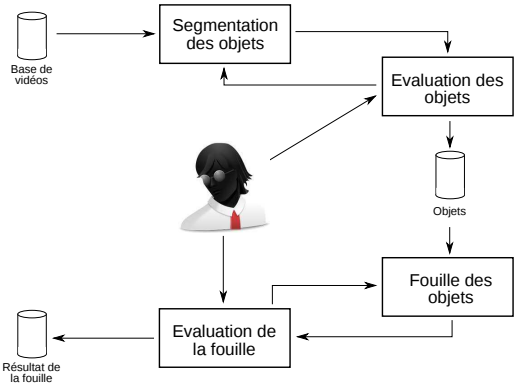


FIG. 5 – *VOMF : Video Object Mining Framework*

Le cadre proposé par VOMF est présenté dans la figure 5 et commence par l'extraction des objets présents dans les vidéos d'une base. Un échantillon des objets obtenus est évalué par l'utilisateur via un système de retour de pertinence. Si les segmentations de l'échantillon sont approuvées par l'utilisateur, l'ensemble des objets est transmis à l'étape de fouille. Dans le cas contraire, les erreurs de segmentation sont identifiées par l'utilisateur, qui introduit de la sémantique par ce biais. Une nouvelle segmentation est alors construite en s'appuyant sur les segmentations courantes et la sémantique apportée par l'utilisateur. Ce cycle est répété jusqu'à ce que l'utilisateur soit satisfait par les objets obtenus. Cependant, il faut veiller à ce que le cycle ne soit pas répété de trop nombreuses fois pour que le temps nécessaire à l'utilisateur soit acceptable. Les résultats de la fouille vidéo sont également évalués par l'utilisateur, via un retour de pertinence sur un échantillon du résultat. Si l'échantillon évalué est satisfaisant, le traitement est terminé. Sinon, à l'instar de la segmentation, l'utilisateur peut corriger l'échantillon. Dans ce cas, la fouille vidéo est relancée et exploite les corrections de l'utilisateur pour améliorer le résultat. L'utilisateur est placé au centre du système. Il supervise le processus de fouille à travers le retour de pertinence et introduit de la sémantique en corrigeant les résultats inappropriés. Pour être efficace, le retour de pertinence ne doit pas être exhaustif. Il faut au contraire que quelques évaluations/corrections suffisent pour influencer profondément les processus de segmentation et de fouille. Ce point est développé dans la prochaine section.

4.5 Un retour de pertinence pour évaluer et guider la fouille vidéo

VOMF est composé de deux étapes, l'extraction des objets et la fouille vidéo. Chacune d'elle dispose de son propre retour de pertinence pour évaluer et guider les processus.

L'extraction des objets est, comme indiqué précédemment, un point critique de VOMF. En fait, sans une extraction d'objets de qualité, il sera particulièrement délicat d'effectuer le processus de fouille. Extraire les objets dans tous les types de vidéos n'est pas une tâche triviale. Le retour de pertinence permet une évaluation directe : le système montre les objets à l'utilisateur et lui demande si ceux-ci correspondent à des objets réels (ou, en d'autres termes, à ceux recherchés par l'utilisateur). Ce retour de pertinence est simple mais très couteux en temps si l'utilisateur doit valider tous les objets extraits. De plus, quel doit être le comportement du système en cas d'insatisfaction de l'utilisateur ? Nous tenons compte de ce problème et proposons la solution suivante. Le système présente un échantillon des objets extraits. L'utilisateur a alors trois possibilités. Il peut valider les objets s'ils représentent ce qu'il recherche. Il peut les corriger. Il peut également les rejeter s'ils ne répondent absolument pas à ses attentes. Les décisions de validation/correction/rejet sont réinjectées dans le système pour guider et améliorer l'extraction des autres objets.

La fouille d'objets est basée sur leurs descriptions mais nécessite également un retour de pertinence. Celui-ci consiste à présenter un échantillon des résultats à l'utilisateur. Par exemple, si l'objectif est la classification, le système présente quelques objets et leur classification. Pour corriger ces objets, l'utilisateur doit changer la classe à laquelle ils appartiennent. Ainsi, l'utilisateur guide le processus de fouille et, à l'itération suivante, cette information est utilisée pour améliorer le résultat.

5 Conclusion

Les systèmes de fouille vidéo (SFV) récents s'appuient sur une description des vidéos réalisée à l'échelle des objets ou des régions, mais sont appliqués sur des éléments tels que les plans ou les vidéos intégrales. Dans cet article, nous avons introduit une nouvelle taxonomie pour caractériser les SFV et l'avons utilisée pour étudier et comparer les SFV actuels. Nous avons également montré que l'objet devrait être l'élément à considérer par les SFV, et nous avons justifié notre proposition en présentant les avantages des systèmes de fouille vidéo orienté objet. Nous avons discuté les répercussions du choix de l'objet comme élément sur les autres caractéristiques définies dans notre taxonomie. L'importance de la segmentation a été soulignée, et nous avons suggéré comment pouvait être introduites des informations de nature sémantique dans les SFV orienté objet. Enfin, nous avons proposé VOMF, un cadre générique pour la fouille vidéo orientée-objet. VOMF offre de nouvelles perspectives, la fouille vidéo étant plus pertinente si les objets considérés sont les objets réels (du point de vue de l'utilisateur) présents dans les vidéos.

Nos futurs travaux incluent l'utilisation de VOMF pour construire des SFV orienté objet pour différents objectifs. Dans cette optique, nous travaillons actuellement sur le problème du clustering des objets. Le but de ces travaux est d'obtenir des groupes d'objets similaires depuis une base de vidéo. Nous disposons actuellement de prototypes pour la fouille guidée par l'util-

isateur d'une part, et pour l'amélioration de segmentation vidéo guidée par l'utilisateur d'autre part. Ces prototypes ont donné des premiers résultats prometteurs mais ne sont aujourd'hui pas totalement aboutis, et nécessitent encore certains travaux de recherche.

Remerciements

Ce travail a été soutenu par Ready Business System et l'Association Nationale de la Recherche et de la Technologie (ANRT). Nous remercions particulièrement Christian Dhinaut de RBS pour sa contribution.

Références

Anjulan, A. et N. Canagarajah (2007a). A novel video mining system. In *14th IEEE International Conference on Image Processing*, pp. 185–188. IEEE.

Anjulan, A. et N. Canagarajah (2007b). Object based video retrieval with local region tracking. *Signal Processing : Image Communication 22*(7-8), 607–621.

Babu, R., K. Ramakrishnan, et S. Srinivasan (2004). Video object segmentation : A compressed domain approach. *IEEE Transactions on Circuits and Systems for Video Technology 14*(4), 462–474.

Basharat, A., Y. Zhai, et M. Shah (2008). Content based video matching using spatiotemporal volumes. *Computer Vision and Image Understanding 110*(3), 360–377.

Brezeale, D. et D. Cook (2008). Automatic video classification : A survey of the literature. *IEEE Transactions on Systems, Man and Cybernetics-part C : Applications and Reviews 38*(3), 416–430.

Brunelli, R., O. Mich, et C. Modena (1999). A survey on automatic indexing of video data. *J. of Visual Communication and Representation 10*(2), 78–112.

Chevalier, F., J.-P. Domenger, J. Benois-Pineau, et M. Delest (2007). Retrieval of objects in video by similarity based on graph matching,. *Pattern Recognition Letters 28*(8), 939–949.

Cios, K., W. Pedrycz, R. Swiniarski, et L. Kurgan (2007). *Data Mining A Knowledge Discovery Approach*. Springer.

de Avila, S., A. da Luz, et A. de Araujo (2008). Vsumm : A simple and efficient approach for automatic video summarization. In *15th International Conference on Systems, Signals and Image Processing*, pp. 449–452.

Gao, X., X. Li, J. Feng, et D. Tao (2009). Shot-based video retrieval with optical flow tensor and HMMs. *Pattern Recognition Letters 30*(2), 140–147.

Hsu, C.-C., H. Chang, et T.-C. Chang (2006). Efficient moving object extraction in compressed low-bit-rate video. In *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*, Washington, DC, USA, pp. 411–414. IEEE Comp. Soc.

Idris, F. et S. Panchanathan (1997). Review of Image and Video Indexing Techniques. *Journal of Visual Communication and Image Representation 8*(2), 146–166.

Koprinska, I. et S. Carrato (2001). Temporal video segmentation : A survey. *Signal Processing : Image Communication 16*(5), 477–500.

Lefèvre, S., J. Holler, et N. Vincent (2003). A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging 9*(1), 73–98.

Liu, D. et T. Chen (2009). Video retrieval based on object discovery. *Computer Vision and Image Understanding 113*(3), 397–404.

Money, A. et H. Agius (2008). Video summarisation : A conceptual framework and survey of the state of the art. *J. of Visual Communication and Image Representation 19*(2), 121–143.

Ren, W., S. Singh, M. Singh, et Y. Zhu (2009). State-of-the-art on spatio-temporal information based video retrieval. *Pattern Recognition 42*(2), 267–282.

Ren, W. et Y. Zhu (2008). A video summarization approach based on machine learning. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Los Alamitos, CA, USA, pp. 450–453. IEEE Comp. Soc.

Rosenfeld, A., D. Doermann, et D. DeMenthon (Eds.) (2002). *Video Mining*. Springer.

Rui, Y., T. Huang, et S. Chang (1999). Image retrieval : current techniques, promising directions, and open issues. *J. of Visual Communication and Image Representation 10*(4), 39–62.

Ruthven, I. et M. Lalmas (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review 18*(2), 95–145.

Sivic, J. et A. Zisserman (2008). Efficient visual search for objects in videos. *Proceedings of the IEEE 96*(4), 548–566.

Snoek, C. G. M. et M. Worring (2009). Concept-based video retrieval. *Foundations and Trends in Information Retrieval 4*(2), 215–322.

Teixeira, L. F. et L. Corte-Real (2009). Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters 30*(2), 157–167.

The Open Video Project (2010). http ://www.open-video.org/.

Toreyin, B., A. Cetin, A. Aksay, et M. Akhan (2005). Moving object detection in wavelet compressed video. *Signal Processing : Image Communication 20*(3), 255–264.

Zhai, S., B. Luo, J. Tang, et C.-Y. Zhang (2007). Video abstraction based on relational graphs. In *Proc. of the Fourth Int. Conf. on Image and Graphics*, pp. 827–832. IEEE Comp. Soc.

Summary

Today, video is becoming one of the primary sources of information. Current video mining systems face the problem of the semantic gap (i.e., the difference between the semantic meaning of video contents and the digital information encoded within the video files). This gap can be bridged by relying on the real objects present in videos because of their semantic meaning. But video object mining needs some semantics, both in the object extraction and in the object mining steps. We think that the introduction of semantics during these steps can be ensured by user interaction. We then propose a generic framework to deal with video object mining.

Classification non supervisée de données satellites multirésolution

Camille Kurtz*

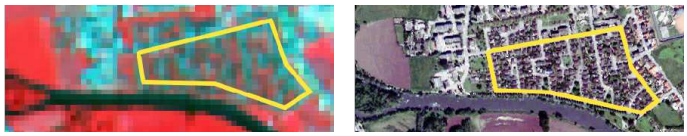
*Université de Strasbourg, LSIIT, UMR CNRS 7005, Strasbourg, France
ckurtz@unistra.fr

Résumé. Depuis quelques années les données issues de capteurs satellitaires deviennent de plus en plus accessibles. Différents systèmes satellitaires sont maintenant disponibles et produisent une importante masse de données utilisée pour l'observation de la Terre. Pour mieux comprendre la complexité de la surface terrestre, il devient courant d'utiliser plusieurs données provenant de capteurs différents. Cependant, la résolution spatiale de ces données n'est pas forcément équivalente, ce qui induit que le contenu sémantique de ces images peut varier. Ainsi, il est souvent difficile d'analyser automatiquement, et de manière conjointe, ces données complexes. Dans cet article nous présentons une approche permettant de tirer partie de l'aspect multirésolution de ces données au sein du processus de classification. Les expériences menées permettent de mettre en avant l'intérêt de cette méthodologie dans le cadre de l'analyse automatique de données satellitaires multirésolution.

1 Introduction

Depuis quelques années les données issues de capteurs satellitaires deviennent de plus en plus accessibles. Différents systèmes satellitaires sont maintenant disponibles et produisent une masse de données importante utilisée pour l'observation de la Terre. Un moyen d'analyser automatiquement le contenu de ces images consiste à classifier ces données (d'une manière supervisée ou non) en fonction des valeurs radiométriques associées à chacun de ces pixels. Cependant, dû à la grande complexité de ces données, les résultats fournis par ces méthodes deviennent de moins en moins intéressants pour les experts (effet poivre et sel, trop grands nombres de classes, etc.). Parallèlement, pour mieux comprendre la complexité de la surface terrestre, il devient courant d'utiliser plusieurs données provenant de capteurs différents. Toutefois, la résolution spatiale de ces données (*e.g.* la surface au sol couverte par chacun des pixels) n'est pas forcément équivalente, induisant que le contenu sémantique peut ne pas être le même d'une image à l'autre. Ces différences de résolution et de sémantique rendent de plus en plus complexe l'analyse automatique de ces données via les méthodologies classiques.

Un exemple concret de ce problème apparaît dans le domaine de la cartographie urbaine, dans lequel plusieurs types d'image, correspondant à des besoins différents, sont utilisés. Pour cartographier le territoire au niveau des quartiers urbains, les images à moyenne résolution (MSR – Medium Spatial Resolution, 30-5m) sont utilisées. Pour cartographier le territoire au niveau des objets urbains (*i.e.* maisons individuelles, jardins, routes, ombres, etc.), les images



(a) Blocs urbains sur une image MSR. (b) Blocs urbains sur une image HSR.

FIG. 1 – *Blocs urbains représentés à différentes résolutions spatiales.*

à haute résolution (HSR – High Spatial Resolution, 3-1m) sont utilisées. Cependant, pour cartographier le territoire en fonction des blocs urbains le composant, aucune résolution d'image adéquate n'est disponible. Les images à moyenne résolution comportent un trop faible niveau de détails pour discriminer efficacement les blocs urbains (Fig. 1(a)) tandis que les images à haute résolution comportent trop de détails pour permettre d'extraire directement ce type d'objets (Fig. 1(b)). Ainsi les classes induites par ces blocs urbains ne peuvent pas être directement obtenues par un processus de classification classique appliqué à l'une ou l'autre de ces images.

Ce problème apparaît donc comme un problème de classification multirésolution nécessitant le développement de méthodes adaptées à la classification conjointe et automatique de ces données complexes, dans le but d'obtenir des résultats de niveaux sémantiques intermédiaires. Nous proposons dans ce papier une méthodologie basée segmentation et classification non supervisée, permettant d'analyser les données sans recourir à une phase de fusion. L'originalité de cette méthode réside dans le fait de classifier des données image à une résolution r en utilisant la décomposition de ces données en termes de clusters [1] dans une résolution $r+1$.

Le reste de cet article est organisé de la manière suivante. En Section 2, un état de l'art (non exhaustif) sur différentes méthodes relatives à la classification multirésolution est proposé. La Section 3 décrit ensuite la méthodologie proposée. La Section 4 présente les différentes expérimentations réalisées sur un jeu de données multirésolution ainsi que les résultats obtenus. Finalement, la Section 5 propose des perspectives de recherches ouvertes par cette étude.

2 État de l'art

Dans le contexte de la classification de données multirésolution, différentes méthodes ont été proposées. Une première approche (approche par fusion) consiste à combiner toutes les descriptions des objets associés aux différentes résolutions en une unique vue (Chang et al., 2007). Cependant, dû à la forte augmentation de la dimension des données (Bellman, 1961), la plupart des algorithmes basés sur des distances ne sont plus assez efficaces pour analyser ces données (Hughes, 1968).

Une solution alternative consiste à trouver un consensus entre les résultats des classifications indépendantes de ces images. Dans (Forestier et al., 2008), une approche est proposée pour produire un résultat unifié qui représente un consensus entre les résultats des classifications non supervisées des images. Cependant, cette approche génère des résultats de classifi-

1. Pour éviter toute confusion, on désignera par « cluster », un ensemble d'objets regroupés automatiquement par un procédé, sans connaissance *a priori*. En revanche, on désignera par « classe », un ensemble thématique d'objets regroupés par l'expert d'une manière supervisée.

cation comportant, pour chaque image, le même nombre de clusters ce qui n'est pas toujours pertinent quand les images ne portent pas le même contenu sémantique. Dans (Wemmert et al., 2009), une approche, utilisant simultanément deux images à différentes résolutions, est décrite. Elle permet, pour chaque résultat de classification, de ne pas générer forcément le même nombre de clusters. Cette méthode commence par réaliser une classification non supervisée (au niveau pixels) sur chacune des deux images. Ensuite, pour chacune d'elles, des régions sont construites. Les régions de l'image à la plus haute résolution sont ensuite caractérisées en fonction du résultat de classification de l'image à la plus basse résolution (chaque région de l'image classifiée est caractérisée suivant sa décomposition en termes de clusters dans l'autre image). Finalement, ces régions sont classifiées en utilisant ces proportions de décompositions.

Cette méthode a produit des résultats prometteurs. Cependant, elle fonctionne directement dans une optique de classification pixels, ce qui est un point faible pour traiter les problèmes de sauts sémantiques (*i.e.*, le manque de concordance entre les informations bas-niveau automatiquement extraites de ces images et les informations haut-niveau attendues par les experts (Smeulders et al., 2000)).

Pour réduire les problèmes des approches pixels, de nouvelles méthodes utilisant des stratégies objets/régions ont été proposées (Baatz et al., 2008). Ces méthodes utilisent une première étape de segmentation pour partitionner l'image en un ensemble de régions homogènes. Dans une seconde étape, ces régions sont regroupées (par un processus de classification) suivant des caractéristiques élémentaires comme des propriétés spectrales et/ou géométriques calculées sur ces régions (Carleer et Wolff, 2006). Un état de l'art sur l'utilisation de ces méthodes dans le cadre de la classification de données satellites est présenté dans (Blaschke, 2010).

Pour conclure sur ce bref état de l'art, il existe des méthodes basées pixels destinées à l'analyse automatique d'images multirésolution. Ces méthodes ont déjà été appliquées à une analyse au niveau sémantique des blocs urbains. Cependant les résultats fournis par ces méthodes souffrent des problèmes liés aux approches pixels. Par ailleurs, des méthodes basées régions ont récemment été proposées et répondent aux problèmes des approches pixels dans le cadre de la classification mono-image. Basé sur ces considérations, le but des travaux présentés dans cet article est de proposer une méthode non supervisée combinant (1) les avantages offerts par les méthodes pixels d'analyse muti-images et (2) l'efficacité des méthodes objets, dans le cadre de l'analyse automatique d'images multirésolution et plus particulièrement pour l'analyse au niveau sémantique des blocs urbains.

3 Méthodologie

L'idée principale sous-jacente à cette méthode est de fusionner les informations fournies par l'analyse des régions de l'image à haute résolution avec le résultat de classification non supervisée de l'image moyenne résolution pour obtenir un résultat final de classification correspondant à un niveau sémantique intermédiaire (dans le cas présent, le niveau sémantique des blocs urbains). Pour ce faire, le contexte spatial des objets d'intérêts ainsi que leurs relations sémantiques à travers les différentes résolutions disponibles sont utilisés pour améliorer l'analyse simultanée des images étudiées. Ces travaux étendent ceux proposés dans (Wemmert et al., 2009), cependant la méthodologie multirésolution proposée dans cet article n'opère pas de la même façon. Nous employons ici une stratégie opposée, qui étudie la composition des régions à moyenne résolution en fonction des clusters formés dans l'image à haute résolu-

tion. Cette nouvelle manière d'appliquer ce processus multirésolution permet, en particulier, la découverte de nouveaux clusters pouvant correspondre à des classes de couverture des sols.

La méthode proposée est divisée en quatre étapes principales décrites ci-dessous (voir schéma, Fig. 2) ; dans le cas standard, les deux images utilisées en entrée sont respectivement de types HSR et MSR.

Étape 1 : Dans un premier temps, les deux images sont segmentées indépendamment par un processus de fusions de zones plates (Fig. 2, Étape 1 ; Sous-section 3.2).

Étape 2 : Les régions de l'image HSR segmentée sont alors classifiées d'une manière non supervisée en utilisant les moyennes des valeurs radiométriques des pixels composant ces régions, fournissant ainsi un résultat de classification pour l'image HSR (Fig. 2, Étape 2 - partie droite ; Sous-section 3.3). La composition de chacune des régions de l'image MSR segmentée est alors calculée en fonction des clusters composant l'image HSR (Fig. 2, Étape 2 - partie gauche ; Sous-section 3.3).

Étape 3 : Basée sur ces compositions, une classification non supervisée des régions de l'image MSR segmentée est réalisée (Fig. 2, Étape 3 ; Sous-section 3.4) : par opposition à une classification « classique », cette dernière a pour but de former des clusters plus « sémantiques » que « radiométriques ». Ces clusters portent ainsi un niveau sémantique intermédiaire qui peut correspondre à celui des blocs urbains.

Étape 4 : Pour chacun des ces clusters ainsi formé, sa composition moyenne en fonction des clusters formés dans le résultat de classification HSR, est ensuite calculée (Fig. 2, Étape 4 - partie en haut à droite ; Sous-section 3.5). Finalement, les régions de l'image HSR segmentée sont projetées dans l'espace des données de l'image MSR dans le but de leurs assigner à chacune un cluster de niveau intermédiaire (Fig. 2, Étape 4 - voir ①, ② et ③ ; Sous-section 3.5).

3.1 Entrée / Sortie

Soit $E = [\![0, d_x - 1]\!] \times [\![0, d_y - 1]\!] \subset \mathbb{N}^2$, la partition discrète (la « grille de pixels ») de la scène représentée. Soit $V_b = [\![0, v_m - 1]\!] \subset \mathbb{N}$, l'ensemble formé par la discrétisation des intensités observées pour la bande spectrale considérée. Une image mono-valuée est une fonction $\mathcal{I}_b : E \to V_b$ qui à chaque point $\mathbf{x} = (x, y) \in E$ de la scène, associe une intensité spectrale $\mathcal{I}_b(\mathbf{x}) = v$.

Posons maintenant $V = \prod_{b=1}^{s} V_b$ avec $V_b = [\![0, v_{b,m} - 1]\!] \subset \mathbb{N}$ pour tout $b \in [\![1, s]\!]$, comme une agglomération de plusieurs bandes spectrales. Une image multi-valuée est une fonction $\mathcal{I} : E \to V$ qui à chaque point $\mathbf{x} = (x, y) \subset E$ de la scène, associe $\mathcal{I}(\mathbf{x}) = \mathbf{v} = \prod_{b=1}^{s} \mathcal{I}_b(\mathbf{x})$. Une image multi-valuée est ainsi définie comme une agglomération d'images mono-valuées.

La méthodologie proposée prend en entrée deux images multi-valuées : une image MSR $\mathcal{I}^1 : E^1 \to V^1$ et une image HSR $\mathcal{I}^2 : E^2 \to V^2$ représentant la même scène (avec $E^1 = [\![0, d_x^1 - 1]\!] \times [\![0, d_y^1 - 1]\!]$, $E^2 = [\![0, d_x^2 - 1]\!] \times [\![0, d_y^2 - 1]\!]$). On rappelle que plus la valeur de $d_x^\star$, $d_y^\star$ est élevée, plus la résolution de l'image est grande. Nous avons ici $d_x^1, d_y^1 < d_x^2, d_y^2$.

Posons $\alpha = d_x^2/d_x^1 = d_y^2/d_y^1 \in \mathbb{N}^*$. Ce coefficient caractérise la « différence » de résolution entre les images MSR et HSR. Il est à noter qu'un point $\mathbf{x} \in E^1$ correspond « physiquement » à un ensemble composé de $\alpha \times \alpha$ points dans E^2. Dans le but de modéliser la correspondance entre les points des deux images $\mathcal{I}^1$ et $\mathcal{I}^2$, on définit la fonction de correspondance $\lambda_{2\to1}$: $E^2 \to E^1$ qui à un point $\mathbf{x} = (x, y) \in E^2$ associe un point $(x/\alpha, y/\alpha)$ dans E^1 ainsi que la fonction $\lambda_{1\to2} : E^1 \to \mathcal{P}(E^2)$ qui à un point $\mathbf{x} = (x, y) \in E^1$ associe un ensemble de points $\alpha \times (x, y) + [\![0, \alpha - 1]\!]^2$ dans E^2.

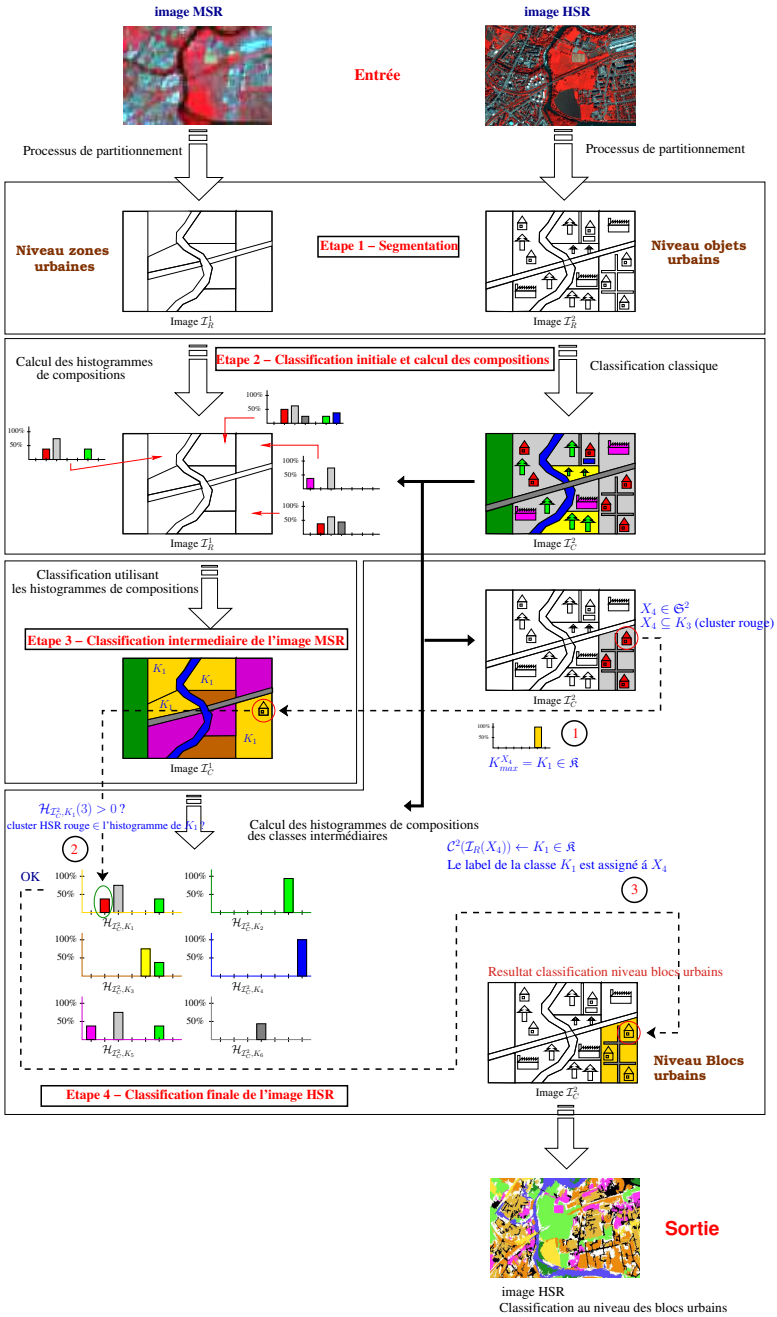


FIG. 2 – *Schéma de la méthode proposée : la méthode prend en entrée deux images, une MSR et une HSR, et fournit en sortie une classification de l'image HSR segmentée au niveau sémantique des blocs urbains.*

La méthode fournit un résultat de classification de la scène à un niveau sémantique intermédiaire (*i.e.*, un niveau correspondant à une résolution située entre $\mathcal{I}^1$ et $\mathcal{I}^2$) qui correspond, dans ce contexte, au niveau sémantique des blocs urbains. Ce résultat de classification est modélisé par une image de label $\mathcal{R} : E^2 \rightarrow [\![1, k]\!] \cup \{\perp\}$ qui, à chaque point $\mathbf{x}$ de la scène (à la résolution la plus haute), associe une valeur de cluster $\mathcal{R}(\mathbf{x})$ parmi les k possibles, ou une valeur inconnue $\perp$ (dans le cas où aucun cluster intermédiaire n'est été associé à ce point).

3.2 Étape 1 - Segmentation des données

La segmentation d'une image $\mathcal{I} : E \rightarrow V$ est une partition $\mathfrak{S} = \{S_i\}_{i=1}^n$ de E ; plus généralement, la scène visualisée à travers $\mathcal{I}$ est décomposée en n parties distinctes S_i, qui sont supposées présenter des propriétés sémantiques spécifiques. A chaque image $\mathcal{I}$ segmentée, on peut ensuite associer une image de régions $\mathcal{I}_R : E \rightarrow [\![1, n]\!]$ qui à chaque point $\mathbf{x} \in E$ associe le label de la région dans laquelle ce point est inclus $S_{\mathcal{I}_R(\mathbf{x})}$.

Deux approches sont généralement considérées pour la segmentation d'images satellites : les approches par lignes de partage des eaux (Vincent et Soille, 1991) et les approches par croissances de régions (Cross et al., 1988). Des études récentes (Carleer et al., 2005) ont montré que les techniques par croissances de régions produisaient de meilleurs résultats, en particulier sur les images HSR. Ainsi, il a été choisi d'utiliser une approche par croissances de régions pour segmenter chacune des images analysées.

La méthode par croissances de régions choisie (Baatz et Schape, 2000) est initialisée avec une partition triviale de l'image, puis fusionne itérativement les éléments de cette partition, en choisissant des paires de régions adjacentes minimisant une fonction d'évaluation donnée (guidée par deux critères principaux : la couleur et la forme). Cette fonction correspond à l'augmentation d'hétérogénéité (*i.e.*, la différence) entre la région $X_{1,2}$ susceptible d'être formée et les deux régions adjacentes (X_1 et X_2) candidates à cette fusion. Durant le processus itératif de fusion, la fonction d'hétérogénéité est calculée pour chaque couple de régions adjacentes de $\mathfrak{S}^{current}$. Ensuite, les couples de régions minimisant la fonction d'hétérogénéité sont fusionnés afin de former de plus larges régions. Quand la valeur courante de cette fonction dépasse un certain seuil τ (appelé le paramètre d'échelle) déterminé par l'utilisateur, le processus de fusion s'arrête et la segmentation est alors terminée.

Ce processus de segmentation est appliqué indépendamment aux deux images $\mathcal{I}^1$ et $\mathcal{I}^2$ pour obtenir les images de régions $\mathcal{I}_R^1$ et $\mathcal{I}_R^2$ (Fig. 2, Étape 1). La partition correspondant à $\mathcal{I}_R^1$ (resp. $\mathcal{I}_R^2$) est notée $\mathfrak{S}^1$ (resp. $\mathfrak{S}^2$) ; le cardinal de cet ensemble est noté n^1 (resp. n^2).

3.3 Étape 2 - Classification initiale des données HSR et calculs des compositions des données MSR

Classification initiale de l'image HSR segmentée Soit $\mathfrak{S} = \{S_i\}_{i=1}^n$ une segmentation de l'image $\mathcal{I} : E \rightarrow V$, et $\mathcal{I}_R : E \rightarrow [\![1, n]\!]$ son image de régions associée. Une classification de $\mathcal{I}$ en k clusters est définie comme une fonction $\mathcal{C} : [\![1, n]\!] \rightarrow [\![1, k]\!]$ qui, à chacune des n régions S_i, associe un des k clusters de $\mathcal{C}(i)$. Un cluster K_i induit par une telle classification est ensuite définie par $K_i = \bigcup_{j \in \mathcal{C}^{-1}(\{i\})} S_j$, *i.e.*, en groupant toutes les régions S_j qui correspondent au même cluster. L'ensemble des k clusters de $\mathcal{I}$ est noté $\mathfrak{K} = \{K_i\}_{i=1}^k$. De la même manière que pour la segmentation, à chaque image classifiée $\mathcal{I}$, on peut associer une image de clusters

$\mathcal{I}_C : E \rightarrow [\![1,k]\!]$ qui à chaque point $\mathbf{x} \in E$, associe le label du cluster dans lequel ce point est inclus $K_{\mathcal{I}_C(\mathbf{x})}$. Durant cette étape, une classification des régions de l'image HSR $\mathcal{I}_R^2$ est réalisée, en utilisant les valeurs radiométriques des pixels de E^2. Ceci conduit à la génération d'une image de clusters $\mathcal{I}_C^2 : E^2 \rightarrow [\![1,k^2]\!]$ (Fig. 2, Étape 2 - partie droite).

Calcul de la composition des régions de l'image MSR segmentée Soit $\mathcal{I}_C : E \rightarrow [\![1,k]\!]$ une image de clusters. L'*histogramme de composition* relatif à cette image, noté $\mathcal{H}_{\mathcal{I}_C}$ est défini comme une fonction $\mathcal{H}_{\mathcal{I}_C} : [\![1,k]\!] \rightarrow \mathbb{N}$ qui associe à chaque cluster i sa valeur $\mathcal{I}_C^{-1}(\{i\})$. L'*histogramme de composition* relatif à l'image $\mathcal{I}_C$, associé à un sous-ensemble $X \subseteq E$, noté $\mathcal{H}_{\mathcal{I}_C,X}$ est défini comme une fonction $\mathcal{H}_{\mathcal{I}_C,X} : [\![1,k]\!] \rightarrow \mathbb{N}$ qui associe à chaque cluster i sa valeur $\mathcal{I}_C^{-1}(\{i\})$. Il correspond à l'*histogramme de composition* relatif à l'image $\mathcal{I}_C$, restreint à l'ensemble X ; en particulier, on a $\mathcal{H}_{\mathcal{I}_C,E} = \mathcal{H}_{\mathcal{I}_C}$. Une fois que l'image $\mathcal{I}_R^2$ a été classifiée, il devient possible de calculer la composition de chaque région $X \in \mathfrak{S}^1$ de l'image MSR segmentée $\mathcal{I}_R^1$ en fonction de l'image de clusters $\mathcal{I}_C^2$. Cette composition est définie par l'histogramme $\mathcal{H}_{\mathcal{I}_C^2,\lambda_{1\rightarrow 2}(X)}$ (Fig. 2, Étape 2 - partie gauche). Celui-ci peut être défini comme une fonction $\mathcal{H}_{\mathcal{I}_C^2,\lambda_{1\rightarrow 2}(X)} : [\![1,k^2]\!] \rightarrow \mathbb{N}$ qui à chaque $i \in [\![1,k^2]\!]$ associe $|\bigcup_{\mathbf{x}\in X} \lambda_{1\rightarrow 2}(\mathbf{x}) \cap (\mathcal{I}_C^2)^{-1}(\{i\})|$. Ainsi, cet histogramme associe pour chaque label i de l'image HSR classifiée, le nombre de pixels qui ont le label i et qui correspondent à un pixel de la région X de l'image MSR.

3.4 Étape 3 - Classification intermédiaire des données MSR

L'étape précédente fournit, pour chaque région $X \in \mathfrak{S}^1$ de l'image MSR $\mathcal{I}^1$, sa composition en fonction des clusters présents dans l'image HSR classifiée $\mathcal{I}^2$, sous la forme d'un histogramme de composition (Fig. 2, Étape 2 - partie gauche). Il devient alors possible de classifier les régions de $\mathfrak{S}^1$ en utilisant la valeur de ces régions dans l'espace de ces histogrammes de composition. Cette classification $\mathcal{C} : [\![1,n^1]\!] \rightarrow [\![1,k]\!]$ permet de regrouper des régions MSR présentant des caractéristiques similaires en termes des objets qui les composent. Cela conduit, en particulier, à l'identification d'associations locales et fréquentes de structures identifiées dans l'image HSR, formant des meta-structures à une résolution plus faible (dans la MSR). Ce processus fournit une image de clusters $\mathcal{I}_C^1 : E^1 \rightarrow [\![1,k]\!]$ associée à l'image MSR $\mathcal{I}^1$ (Fig. 2, Étape 3) qui est indirectement basée sur les valeurs radiométriques de $\mathcal{I}^1$ (grâce à la segmentation initiale $\mathcal{I}_R^1$) et directement basée sur les sémantiques implicites de l'image HSR $\mathcal{I}^2$ (grâce à sa classification $\mathcal{I}_C^2$). Cette image de clusters regroupe des informations relatives à l'image MSR et à l'image HSR, mais à un niveau intermédiaire. Ainsi, les clusters résultants peuvent potentiellement être proches des classes de blocs urbains définies par l'expert.

3.5 Étape 4 - Classification finale des données HSR

De la même manière que pour le calcul des histogrammes de composition pour les régions de $\mathfrak{S}^1$ de l'image MSR $\mathcal{I}^1$, il est possible de calculer les histogrammes de composition pour les k clusters obtenus par le biais de la classification $\mathcal{C}$, en fonction des clusters issus de la classification $\mathcal{I}_C^2$ de l'image HSR $\mathcal{I}^2$ (Fig. 2, Étape 4 - partie en haut à droite). Ces histogrammes sont définis, pour $i \in [\![1,k]\!]$ comme $\mathcal{H}_{\mathcal{I}_C^2,\lambda_{1\rightarrow 2}((\mathcal{I}_C^1)^{-1}(\{i\}))}$, (noté $\mathcal{H}_{\mathcal{I}_C^2,i}$). Ensuite, les régions de l'image HSR segmentée sont projetées dans l'espace des données de

l'image MSR dans le but de leur assigner à chacune un cluster intermédiaire (Fig. 2, Étape 4 - voir ①, ② et ③). L'idée consiste à assigner à chaque région de la segmentation $\mathfrak{S}^2$ de l'image HSR $\mathcal{I}^2$, un label de cluster de $\mathfrak{K}$ fourni par la classification $\mathcal{C}$. Pour ce faire, pour chaque région HSR $X_i \in \mathfrak{S}^2$, un histogramme de composition est calculé en fonction des clusters intermédiaires de $\mathcal{I}^1_C$. Ces n^2 histogrammes sont ensuite définis, pour $i \in [\![1, n^2]\!]$ comme $\mathcal{H}_{\mathcal{I}^1_C, \lambda_{2\to1}(X^2_i)}$. À partir de ces histogrammes, il est possible, pour chaque région HSR $X_i \in \mathfrak{S}^2$, de connaître sa décomposition principale en fonction des clusters de $\mathcal{I}^1_C$ –la classification intermédiaire de l'image MSR (Fig. 2, Étape 4 - voir ①). On note $K^{X_i}_{max} \in \mathfrak{K}$ ce cluster. Ici, t_{max} est le cluster principal qui compose X_i dans $\mathcal{I}^1_C$. On nomme p son pourcentage de majorité ($p = \mathcal{H}_{\mathcal{I}^1_C, \lambda_{2\to1}(X_i)}(t_{max})/|X_i|$). Étant donné un seuil de majorité $s_{maj} \in [0, 1]$, deux cas peuvent apparaître :

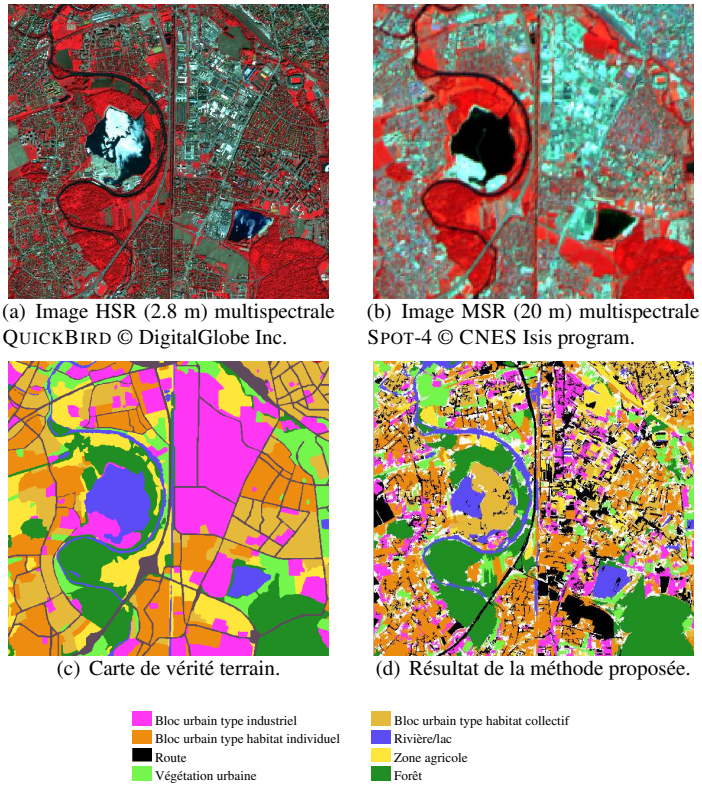
- $p < s_{maj}$ — X_i n'est pas **projetable** : Cela signifie que le cluster t_{max} n'est pas suffisamment majoritaire, *i.e.*, que X_i n'est pas correctement associable (d'un point de vue spatial) à une zone sémantique à la résolution MSR. On considère que cela est dû à un problème de segmentation dans l'une des images. On assigne le label de cluster $\perp$ à X_i.
- $p \geq s_{maj}$ — X_i est **projetable** : À partir de l'histogramme de composition $\mathcal{H}_{\mathcal{I}^2_C, K_{t_{max}}}$ du cluster $K_{t_{max}}$, on récupère $v = \mathcal{H}_{\mathcal{I}^2_C, K_{t_{max}}}(\mathcal{C}^2(i))$, la contribution, dans cet histogramme de composition, du cluster correspondant à la région X_i à la résolution HSR. Ensuite deux cas peuvent apparaître :
  - ⋆ $v = 0$ — X_i est **inclassifiable** au niveau sémantique intermédiaire : cela signifie que le cluster de la région HSR X_i n'est pas cohérent avec les clusters HSR qui composent le cluster MSR $K^{X_i}_{max}$. On considère que cela est dû à un problème de classification dans l'une des images. On assigne le label de cluster $\perp$ à X_i.
  - ⋆ $v > 0$ — X_i est **classifiable** au niveau sémantique intermédiaire : (Fig. 2, Étape 4 - voir ②). On assigne le label de cluster de $K^{X_i}_{t_{max}}$ à X_i (Fig. 2, Étape 4 - voir ③).

4 Expérimentations

4.1 Protocole expérimental

Données Les expériences ont été menées sur des extraits de deux images multispectrales ayant des résolutions spatiales de 2,8m (Fig. 3(a)) et 20m (Fig. 3(b)), représentant une partie de la zone urbaine de Strasbourg (France). Cette zone est un exemple typique de zone péri-urbaine avec des surfaces d'eau (au centre), une zone forestière dans le sud, des zones industrielles, des zones agricoles et des zones de bâti pavillonnaire et/ou collectif. L'image HSR provient du capteur QUICKBIRD (DigitalGlobe Inc.) et propose quatre bandes spectrales (bleu, vert, rouge et proche infrarouge), tandis que l'image MSR provient du capteur SPOT4 (CNES) et propose trois bandes spectrales (vert, rouge, proche infrarouge).

Expériences Pour évaluer l'approche de classification multirésolution proposée (dans le cadre de l'analyse des blocs urbains), les résultats obtenus via cette méthode ont été comparés avec une carte de vérité terrain fournie par l'expert. Pour cela, nous avons étudié l'impact des paramètres de la méthode sur les résultats. Des expérimentations préliminaires ont montré que l'influence des paramètres de segmentation est moins significative sur les résultats de la méthode que ceux relatifs à la classification des données. Ainsi, nous avons décidé de n'étudier



(a) Image HSR (2.8 m) multispectrale QUICKBIRD © DigitalGlobe Inc.

(b) Image MSR (20 m) multispectrale SPOT-4 © CNES Isis program.

(c) Carte de vérité terrain.

(d) Résultat de la méthode proposée.

FIG. 3 – *Extraits du jeu de données utilisé (a,b). Image de vérité terrain au niveau sémantique des blocs urbains (c). Résultat de la méthode proposée (d). Les couleurs des clusters ont été choisies pour correspondre à celles des classes de la carte de Vérité Terrain. La couleur blanche est utilisée pour les régions inclassifiables (ayant le label $\perp$, Étape 4 de la méthode).*

que l'impact des paramètres de classification. Pour ce faire, la méthode a été appliquée plusieurs fois sur le jeu de données en utilisant des nombres de clusters différents (Tab. 1). Dans ces expériences, nous avons appliqué les 4 étapes décrites en section 3 de la manière suivante :

Étape 1 - Segmentation des données Pour trouver les partitions les plus adaptées à chacune des deux images, le processus de segmentation a été appliqué plusieurs fois par l'expert en utilisant des paramètres différents (*i.e.* en faisant varier τ, le paramètre d'échelle). Pour l'image MSR la meilleur valeur trouvée de τ est 15 (produisant ainsi une partition composée de 1688 régions) tandis que pour l'image HSR la meilleur valeur trouvée de τ est 25 (produisant ainsi une partition composée de 19752 régions).

Étape 2 - Classification initiale des données HSR et calculs des compositions des données MSR Les classifications initiales de l'image HSR segmentée ont été produites en utilisant l'algorithme K-MEANS. Il est à noter que n'importe quel algorithme de classification non supervisée aurait pu être utilisé. Pour ce type d'images, le nombre de classes dépend directement

MSR Image \ HSR Image		HSR 2.8 m			
		$C_{HSR} = 15$	$C_{HSR} = 20$	$C_{HSR} = 22$	$C_{HSR} = 25$
MSR 20 m	$C_{inter} = 9$	0.7717	0.7825	0.7839	0.7648
	$C_{inter} = 11$	0.7921	0.7957	0.7972	0.7787
	$C_{inter} = 13$	0.8079	0.8080	**0.8203**	0.7913
	$C_{inter} = 15$	0.7972	0.8047	0.8171	0.7894

TAB. 1 – *Impact du nombre de clusters utilisés sur le résultat final de classification. La meilleur valeur de Kappa obtenue est en rouge.*

du nombre de matériaux différents composant les objets urbains qui peuvent apparaître dans la zone étudiée. En accord avec les experts, nous avons expérimenté la méthode avec 15, 20, 22 et 25 clusters (noté C_{HSR}). Ensuite, pour chaque expérimentation, les histogrammes de composition des régions de l'image MSR segmentée (en termes de clusters dans l'image HSR classifiée) ont été calculés.

Étape 3 - Classification intermédiaire des données MSR Pendant cette étape, l'algorithme K-MEANS a été appliqué sur l'image MSR segmentée en utilisant les histogrammes de composition. Les expérimentations ont montré que la méthode ne pouvait pas directement trouver tous les clusters relatifs aux blocs urbains. Pour remédier à ce problème, l'algorithme K-MEANS a été appliqué en utilisant un plus grand nombre de clusters que le nombre de classes attendues (9, 11, 13 et 15 clusters) noté C_{inter}. Une étape de post-traitement a ensuite consisté à appliquer un algorithme de classification non-supervisée hiérarchique ascendant dans le but de réduire le nombre de clusters (8 classes de blocs urbains dans le contexte courant).

Étape 4 - Classification finale des données HSR Les régions de l'image HSR segmentée sont ensuite projetées dans l'espace des données de l'image MSR ainsi classifiée afin de leurs assigner à chacune un cluster de niveau intermédiaire. Le seuil de majorité s_{maj} a été expérimentalement fixé à $0,75$.

Validations Les résultats obtenus ont été ensuite évalués par une comparaison avec une vérité terrain issue d'une base de données (BDOCS 2000 Cigal 2003) utilisée pour une cartographie au 1/10.000e. Cette carte contient 8 classes thématiques liées aux blocs urbains. Nous avons choisi d'évaluer nos résultats en utilisant l'indice Kappa. Cet indice permet d'évaluer quantitativement la qualité d'un résultat de classification par rapport à carte de vérité terrain. Cet indice peut être vu comme un indicateur de concordance entre deux classifications : il indique le pourcentage des bonnes correspondances qui sont dues à la réalité du terrain et non uniquement au hasard. Une valeur entre 1.00 et 0.81 indique une concordance parfaite, entre 0.80 et 0.61 indique une bonne concordance, etc.

4.2 Résultats et Discussion

Les résultats obtenus (en termes de Kappa) sont récapitulés dans le tableau Tab. 1. Ce dernier montre que les paramètres de classification C_{HSR} et C_{inter} ont chacun une influence différente sur le résultat de la méthode :

$\boldsymbol{C_{HSR}}$: Si le nombre de clusters dans l'image segmentée HSR est trop faible ($C_{HSR} = 15$), certains clusters ne sont pas pertinents ou regroupe des objets trop éloignés sémantiquement (ceci est probablement dû a un trop faible nombre de clusters). Ceci conduit ensuite à

la construction de clusters intermédiaires incorrects. Par exemple, des zones agricoles ont été regroupées dans le même cluster que des surfaces d'eau. Cependant si le nombre de clusters HSR est trop important ($C_{HSR} = 25$), les associations locales et fréquentes de structures HSR (Étape 3 de la méthode) sont difficiles à identifier (probablement à cause du trop faible nombre de régions MSR ayant une composition similaire en termes de clusters HSR). Les meilleurs résultats ont été trouvés avec $C_{HSR} = 20$ et $C_{HSR} = 22$.

$\boldsymbol{C_{inter}}$: Si le nombre de clusters intermédiaires est trop faible ($C_{inter} = 9$), certains clusters finaux ne coïncident pas avec les classes sémantiques attendues (les clusters obtenus regroupent probablement trop d'objets HSR différents). Avec $C_{inter} = 15$, le nombre de clusters intermédiaires est trop important. Ces derniers ne sont pas assez spécialisés et ne coïncident pas avec les classes attendues. Les meilleurs résultats ont été obtenus avec $C_{inter} = 13$.

Ainsi, le meilleur résultat global a été obtenu avec $C_{HSR} = 22$ et $C_{inter} = 13$ et est présenté dans la figure Fig. 3(d). En conclusion, cette étude paramétrique a montré que la qualité du résultat final de classification au niveau des blocs urbains est directement liée aux choix des paramètres de classification. De plus, des tests réalisés sur un autre jeu de données ont montré que les valeurs paramétriques $C_{HSR} = 22$ et $C_{inter} = 13$ sont bien adaptées pour obtenir un résultat de classification au niveau sémantique des blocs urbains ce qui confirme la validité de la méthode. Avec cette configuration, les clusters obtenus peuvent être utilisés par l'expert pour une analyse du territoire au niveau sémantique des blocs urbains.

5 Conclusion et perspectives

Cet article présente des travaux sur la classification automatique de données multirésolution provenant de sources différentes. La méthode proposée s'appuie respectivement sur la segmentation et sur la classification non supervisée et conjointe des données. Elle permet ainsi d'utiliser les informations fournies par les différentes résolutions sans recourir à une phase de fusion de données. Les différentes expériences réalisées sur des jeux de données multirésolution ont montré que la méthode peut s'avérer efficace dans le cadre de la classification de blocs urbains. Ces résultats justifient ainsi de futurs développements qui pourraient permettre d'améliorer la méthode. La principale amélioration à considérer serait d'automatiser la méthode afin de déterminer itérativement les configurations de paramètres les plus adaptées. En effet, la méthode requiert de fixer certains paramètres manuellement (relatifs à la segmentation et à la classification). Alors que certains d'entre eux sont facilement automatisables (*e.g.* le nombre de classes, qui dépend de l'application), d'autres pourraient être fixés, d'une manière plus interactive (*e.g.* les paramètres de segmentation). Ces améliorations permettraient ainsi de rendre la méthode plus ergonomique pour ses utilisateurs potentiels.

Références

Baatz, M., C. Hoffmann, et G. Willhauck (2008). *Object-Based Image Analysis*, Chapter Progressing from object-based to object-oriented image analysis, pp. 29–42. Thomas Blaschke and Stefan Lang and Geoffrey J. Hay.

Baatz, M. et A. Schape (2000). Multiresolution segmentation–An optimization approach for high quality multi-scale image segmentation. In *Angewandte Geographische Informationsverarbeitung Symposium*, pp. 12–23.

Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing 65*(1), 2 – 16.

Carleer, A., O. Debeir, et E. Wolff (2005). Assessment of vhsr satellite image segmentations. *Photogrammetric Engineering and Remote Sensing 71*(11), 1285–1294.

Carleer, A. et E. Wolff (2006). Urban land cover multilevel region-based classification of VHR data by selecting relevant features. *International Journal of Remote Sensing 27*(6), 1035–1051.

Chang, Y.-L., L.-S. Liang, C.-C. Han, J.-P. Fang, W.-Y. Liang, et K.-S. Chen (2007). Multisource data fusion for landslide classification using generalized positive Boolean functions. *IEEE Transactions on Geoscience and Remote Sensing 45*(6), 1697–1708.

Cross, A., D. Mason, et S. Dury (1988). Segmentation of remotely-sensed images by a split-and-merge process. *International Journal of Remote Sensing 9*(8), 1329–1345.

Forestier, G., C. Wemmert, et P. Gançarski (2008). Multi-source images analysis using collaborative clustering. *EURASIP Journal on Advances in Signal Processing 2008*, 1–11.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory 14*(1), 55–63.

Smeulders, A., M. Worring, S. Santini, A. Gupta, et R. Jain (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(12), 1349–1380.

Vincent, L. et P. Soille (1991). Watersheds in digital spaces : An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*(6), 583–598.

Wemmert, C., A. Puissant, G. Forestier, et P. Gançarski (2009). Multiresolution remote sensing image clustering. *IEEE Geoscience and Remote Sensing Letters 6*(3), 533–537.

Summary

For few years, data provided by satellite devices are being widely available. Different satellite systems produce an important mass of heterogeneous data used for Earth observation. To better understand the complexity of the Earth's surface, it becomes more common to use data from different sensors. However, the spatial resolutions of these data are not necessarily equivalent, which indicates that the semantic contents of the images may differ. Thus, it is often difficult to automatically analyze, in a joint fashion, these complex data. In this article we present an approach to take advantage of the multiresolution aspect of data within the classification process. Experiments allow to highlight the relevance of this methodology in the context of the analysis of multiresolution satellite images.

Influence de l'estimation des paramètres de texture pour la classification de données complexes

Anthony Fiche*, Jean-Christophe Cexus*
Arnaud Martin**, Ali Khenchaf*

*ENSTA Bretagne
2 rue François Verny
29806 Brest Cedex 9
{Anthony.Fiche,Jean-Christophe.Cexus,Ali.Khenchaf}@ensieta.fr
**Université de Rennes 1 / UMR 6074 IRISA
Rue Edouard Branly BP 30219
22302 Lannion Cedex
Arnaud.Martin@univ-rennes1.fr

Résumé. Ce papier présente une classification de données basée sur la théorie des fonctions de croyance. La complexité de ce problème peut être vue de deux façons. Tout d'abord, ces données peuvent être imprécises et/ou incertaines. Ensuite, il est difficile de trouver le juste modèle permettant de représenter les données. Le modèle Gaussien est souvent utilisé mais reste limité lorsque les données sont complexes. Ce modèle n'est qu'un cas particulier des distributions α-stables qui permettent une plus grande souplesse dans la modélisation des données. La classification est divisée en deux étapes. La phase d'apprentissage permet de modéliser les données par un mélange de distributions α-stables et de Gaussiennes. La phase de test permet de classifier les données à partir de la théorie des fonctions de croyance et de comparer les deux modèles. La classification est d'abord réalisée sur des données générées puis réelles type images sonar.

1 Introduction

La classification de données réelles à partir d'images radar ou sonar est un problème complexe (Kernéis, 2007; Laanaya, 2007). Par exemple, des paramètres de texture calculés à partir des travaux d'Haralick (Haralick et al., 1973; Haralick, 1979) peuvent être extraits de ces images. La complexité se traduit par le fait que ces paramètres peuvent présenter une queue lourde, c'est à dire que la queue de la distribution décroît plus lentement que la queue de la Gaussienne, être asymétrique ou bien comporter plusieurs modes. Ces différentes contraintes entraînent des difficultés pour choisir un modèle permettant de représenter ces paramètres de texture sans perdre d'information. Le modèle Gaussien est très souvent utilisé du fait de sa simplicité d'utilisation. Cependant, ce modèle devient obsolète dès que les données sont complexes. Il est alors impossible de les représenter à partir d'une seule distribution mais plutôt avec un mélange de distributions. Le modèle Gaussien n'est qu'un cas particulier des distributions α-stables. Les distributions α-stables ont la particularité de modéliser des données non

symétriques ainsi que d'être pourvues d'une queue lourde. Ces distributions ont vu leur utilité s'accroître du fait qu'elles peuvent modéliser des bruits impulsifs en radar et en télécommunications.
L'objectif de cette contribution est de montrer l'intérêt de modéliser des distributions de données à partir d'un mélange de distributions α-stables par rapport à un mélange de Gaussiennes lors de classification de données. Les données issues de capteurs vont être modéliser à partir d'un mélange de distributions α-stables. Ces données sont supposées incertaines et/ou imprécises. La théorie des fonctions de croyance permet de prendre en compte ces considérations. Le papier se divise en trois parties. Tout d'abord, nous présentons les distributions α-stables. Ensuite, nous développons la théorie des fonctions de croyance. Enfin, nous effectuons une classification de données générées et réelles.

2 Les distributions α-stables

Les distributions α-stables ont été introduites par Paul Lévy (Lévy, 1924). Il existe plusieurs définitions permettant de caractériser une distribution α-stable. Dans cette partie, nous présentons tout d'abord la notion de stabilité, ensuite la fonction caractéristique et enfin la densité de probabilité.

2.1 Notion de stabilité

Paul Lévy définit la notion de stabilité par le fait que la somme de deux variables aléatoires indépendantes, chacune suivant une loi stable, suit aussi une loi stable. Cette définition se traduit mathématiquement par : Une variable aléatoire X est dite stable si $\forall(a,b) \in (\mathbb{R}^+)^2$, il existe $c \in \mathbb{R}^+$ et $d \in \mathbb{R}$ tel que :

$$aX_1 + bX_2 = cX + d \tag{1}$$

avec X_1 et X_2 2 variables aléatoires stables indépendantes.
Dans la suite de l'article, nous travaillons avec la densité de probabilité. Or, cette définition ne nous permet pas de la représenter. Par la suite, nous présentons la définition de la fonction caractéristique.

2.2 Fonction caractéristique d'une α-stable

Il n'existe pas qu'une seule définition pour la fonction caractéristique d'une distribution α-stable, notée $S_\alpha(\beta,\gamma,\delta)$. La définition usuelle est celle proposée par Samorodnitsky et Taqqu (1994). Cependant, la fonction caractéristique n'est pas continue pour les valeurs de x où $\alpha = 1$ et $\beta = 0$. On préfère alors celle définie par Zolotarev (1986). Une variable aléatoire est dite stable si sa fonction caractéristique $\phi(t)$ vérifie :

$$\phi(t) = \begin{cases} \exp(it\delta - |\gamma t|^\alpha[1 + i\beta\tan(\frac{\pi\alpha}{2})sign(t)(|t|^{1-\alpha} - 1)]) & \text{si } \alpha \neq 1 \\ \exp(it\delta - |\gamma t|[1 + i\beta\frac{2}{\pi}sign(t)\log|t|]) & \text{si } \alpha = 1 \end{cases} \tag{2}$$

avec $\alpha \in]0,2]$, $\beta \in [-1,1]$, $\gamma \in \mathbb{R}^{+*}$ et $\delta \in \mathbb{R}$.

Ces quatre paramètres sont :
- α est appelé l'exposant caractéristique.
- β est le paramètre de d'asymétrie.
- γ représente le paramètre d'échelle.
- δ indique le paramètre de localisation.

La fonction de densité de probabilité (*fdp*) est obtenue en effectuant une transformée de Fourier de la fonction caractéristique :

$$fdp(x) = \int_{-\infty}^{+\infty} \phi(t) \exp(-itx) dt \tag{3}$$

Plusieurs problèmes rendent difficile la représentation de cette densité de probabilité. Tout, d'abord, l'expression de la fonction caractéristique est complexe. Ensuite, les bornes d'intégration sont infinies. Cependant, Nolan (1997) permet de résoudre ce dernier point en effectuant des changements de variable pour se ramener à des bornes d'intégration finies. Un programme Matlab suivant cette démarche a été développé [1].

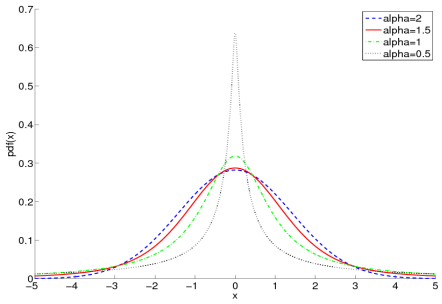


FIG. 1 – *Influence du paramètre* α *avec* $\beta = 0$, $\gamma = 1$ *et* $\delta = 0$.

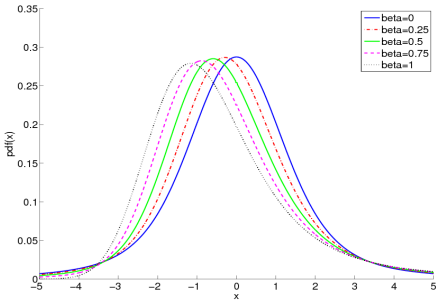


FIG. 2 – *Influence du paramètre* β *avec* $\alpha = 1.5$, $\gamma = 1$ *et* $\delta = 0$.

1. http ://math.bu.edu./people/mveillet/research.html.

Chaque paramètre influe de manière différente sur la représentation de la densité de probabilité d'une loi stable. On remarque que si α est petit, la distribution présente un pic très important (*cf.* figure 1). Lorsque $\beta \to 1$, la distribution a une queue lourde à droite et inversement lorsque $\beta \to -1$ (*cf.* figure 2). Le paramètre γ permet de dilater ou de compresser les distributions (*cf.* figure 3). Enfin, δ permet de positionner le mode de la distribution suivant l'axe des abscisses (*cf.* figure 4).

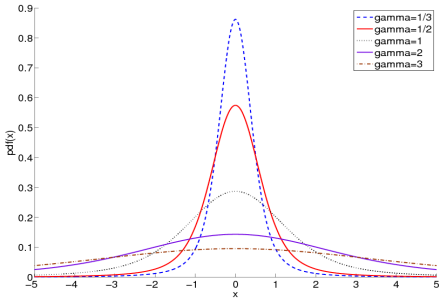


FIG. 3 – *Influence du paramètre γ avec $\alpha = 1.5$, $\beta = 0$ et $\delta = 0$.*

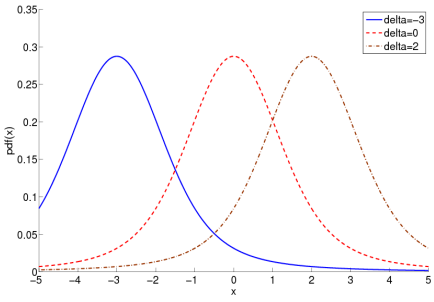


FIG. 4 – *Influence du paramètre δ avec $\alpha = 1.5$, $\beta = 0$ et $\gamma = 1$.*

2.3 Exemples de *fdp*

Malgré la difficulté à représenter les distributions α-stables, il est possible de décrire quelques lois connues. Lorsque $\alpha = 2$ et $\beta = 0$, on retrouve l'expression d'une distribution Gaussienne :

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\delta)^2}{2\sigma^2}\right) \tag{4}$$

avec δ représentant la moyenne et σ^2 la variance. À partir de la définition de la fonction caractéristique, il faut que $\sigma^2 = 2\gamma^2$. Lorsque $\alpha = 1$ et $\beta = 0$, on définit une loi de Cauchy :

$$f(x) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x-\delta)^2} \tag{5}$$

Plus généralement, lorsque $\beta = 0$, on dit que la distribution α-stable est symétrique.
Les données issues des capteurs vont être modélisées par un mélange de distributions α-stables. Ces données sont considérées comme imprécise et/ou incertaine. Des travaux ont été proposés pour classifier des données modélisées à partir d'un mélange de Gaussiennes en s'appuyant sur la théorie des probabilités (voir (Williams, 2009)). La probabilité de chaque sédiment est calculée à partir du mélange de Gaussiennes. Cependant, cette théorie est limitée pour prendre en compte l'incertitude. La théorie des fonctions de croyance permet de prendre en considération cette notion. Par conséquent, la partie suivante a pour but de définir les concepts de cette théorie.

3 La théorie des fonctions de croyance

Les travaux de Dempster (1967) sont à l'origine de la théorie des fonctions de croyance. Shafer (1976) a d'ailleurs repris ces travaux pour formaliser cette théorie. Par conséquent, nous exposons les concepts de base de la théorie des fonctions de croyance dans le cadre discret puis dans le cadre continu.

3.1 Les fonctions de croyance dans le cadre discret

Dans cette section, nous définissons la fonction de masse m, la règle de combinaison permettant de combiner plusieurs fonctions de masses entre elles ainsi que la probabilité pignistique, permettant de prendre une décision.

3.1.1 Définitions

Tout d'abord, la théorie des fonctions de croyance permet de travailler sur l'ensemble $\Theta = \{C_1, \ldots, C_n\}$, appelé cadre de discernement. Θ s'interprète comme toutes les hypothèses possibles d'un problème. Les fonctions de croyance sont définies de 2^Θ dans $[0, 1]$, c'est à dire qu'il est possible d'attribuer une croyance sur des disjonctions de Θ. La quantité m appelée fonction de masse vérifie :

$$\sum_{A \in 2^\Theta} m(A) = 1 \tag{6}$$

À partir de cette fonction de masse, il est possible de définir d'autres fonctions :

$$bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \tag{7}$$

$$pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{8}$$

$$q(A) = \sum_{B \subset \Theta, B \supseteq A} m(B) \tag{9}$$

Ces fonctions représentent la même information que m mais sous des formes différentes. La fonction de crédibilité, notée bel, correspond à la croyance minimum que l'on peut avoir en A.

La fonction de plausibilité, notée pl, correspond à la croyance maximale que l'on peut avoir en A. Enfin, la fonction de communalité, notée q, représente la somme de toutes les masses allouées à un sur ensemble de A et est très utilisée dans l'étape de combinaison.

3.1.2 La règle de combinaison

L'étape de combinaison permet de combiner plusieurs fonctions de masse entre elles. En effet, des experts peuvent avoir des opinions différentes sur un élément $A \subseteq \Theta$ en attribuant une masse m_j à A. Il existe plusieurs règles de combinaison répartissant différemment le conflit entre les sources. La plus employée est la règle de combinaison conjonctive. On obtient comme masse résultante :

$$m(A) = \sum_{C_1 \cap \cdots \cap C_n = A \neq \emptyset} \prod_{j=1}^{M} m_j(B_j) \qquad \forall A \in 2^{\Theta} \tag{10}$$

En pratique, il est difficile de programmer cette formule. On préfère utiliser les fonctions de communalité. Chaque masse m_j est convertit en sa fonction de communalité q_j. Il est possible de calculer la fonction de communalité résultante par :

$$q(A) = \prod_{j=1}^{M} q_j(A) \tag{11}$$

La masse résultante est ensuite obtenue en transformant la fonction de communalité en fonction de masse. Une fois la masse finale de chaque sous ensemble de A, il est nécessaire de prendre une décision.

3.1.3 La transformation pignistique

Plusieurs opérateurs, tels que le maximum de crédibilité et le maximum de plausibilité permettent de prendre une décision. Cependant, l'opérateur le plus utilisé est basé sur la transformée pignistique (Smets, 1990). Le principe est de répartir uniformément la masse des sous-ensembles de Θ sur les singletons qui les composent. La formule est la suivante :

$$betP(C_i) = \sum_{A \in 2^{\Theta}, C_i \in A} \frac{m(A)}{|A|(1 - m(\emptyset))} \tag{12}$$

où $|A|$ représente la cardinalité de A.
On choisit ensuite la décision C_i pour l'observation $\mathbf{x}$ en évaluant $\max\limits_{1 \leq k \leq n} betP(C_k)(x)$.

3.2 Les fonctions de croyance continues

Il est possible de définir les notions définies précédemment dans le cadre continu. Les notions de bases ont été introduites par Shafer (1976), puis par Strat (1984). Récemment, Smets (2005) étend la théorie des fonctions de croyance sur l'ensemble $\mathbb{R} = \mathbb{R} \cup [-\infty, +\infty]$ où il attribue une masse sur des intervalles de $\mathbb{R}$.

3.2.1 Définitions

Considérons $\mathcal{I} = \{[x, y], (x, y], [x, y), (x, y); x, y \in \mathbb{R}\}$ un ensemble d'intervalles fermés, semi-ouverts et ouverts de $\mathbb{R}$. Les éléments focaux sont des intervalles fermés de $\mathbb{R}$. La quantité $m^{\mathcal{I}}(x, y)$ est appelée densité de masse et est reliée à une densité de probabilité. La densité de masse est nulle lorsque x est supérieur à y. Soit $[a, b]$ un intervalle de $\mathbb{R}$. Les fonctions définies dans le cadre discret deviennent :

$$bel([a,b]) = \int_{x=a}^{x=b} \int_{y=x}^{y=b} m^{\mathcal{I}}(x,y)dydx \tag{13}$$

$$pl([a,b]) = \int_{x=-\infty}^{x=b} \int_{y=max(a,x)}^{y=+\infty} m^{\mathcal{I}}(x,y)dydx \tag{14}$$

$$q([a,b]) = \int_{x=-\infty}^{x=a} \int_{y=b}^{y=+\infty} m^{\mathcal{I}}(x,y)dydx \tag{15}$$

3.2.2 La probabilité pignistique

La probabilité pignistique, avec $a < b$, est définie par la formule :

$$BetP([a,b]) = \int_{x=-\infty}^{x=+\infty} \int_{y=x}^{y=+\infty} \frac{|[a,b] \cap [x,y]|}{|[x,y]|} \times m^{\mathcal{I}}(x,y)dxdy \tag{16}$$

Il est possible d'obtenir une densité de masse à partir de la probabilité pignistique. Cependant, il existe plusieurs densités de masse associées à une même probabilité pignistique. Pour simplifier le problème, on choisit la densité de masse consonante, c'est à dire que les éléments focaux sont emboîtés. Les éléments focaux, notés I_u peuvent être rangés dans un index u tel que $I_u \subseteq I_{u'}\ avec\ u' > u$. Cette définition est utilisée pour appliquer le principe de moindre engagement. Le problème avec les fonctions de croyance est qu'elles ne sont pas totalement définies. La seule chose que l'on connaît est son appartenance à une famille de fonctions. Le principe de moindre engagement consiste à choisir la fonction de croyance qui est la moins informative.

3.2.3 Théorème de Bayes généralisé

Supposons que $x \in \mathbb{R}$ un vecteur d'observations. Il est possible de calculer la fonction de masse de chaque $A \in 2^{\Theta}$ connaissant l'observation x à partir du théorème de Bayes généralisé (Smets, 1993; Delmotte et Smets, 2004). Il s'écrit de la manière suivante :

$$m(A/x) = \prod_{C_j \in A} pl_j(x) \prod_{C_j \in A^c} (1 - pl_j(x)) \tag{17}$$

Les fonctions de plausibilité ont été calculées dans le cas symétrique unimodale, qui est vraie pour le cas Gaussian (Smets, 2005) et α-stable symétrique (Fiche et al., 2010a). Il est possible de généraliser le calcul des fonctions de plausibilité dans le cas d'un mélange de Gaussiennes (Caron et al., 2006). Nous reprenons cette démarche pour l'étendre à un mélange de distributions α-stables.

	α			γ			δ			W		
	α_1	α_2	α_3	γ_1	γ_2	γ_3	δ_1	δ_2	δ_3	W_1	W_2	W_3
Premier mélange	1.6	1.4	1.2	1.26	1.5	0.8	-4.5	0	10.1	1/2	1/3	1/6
Second mélange	1.2	1.4	1.6	1.2	3.2	1.5	-13.5	10.1	3	1/2	1/4	1/4
Troisième mélange	1.56	1.24	1.78	0.5	2	4	-8.7	1.3	5.5	1/6	1/6	2/3

TAB. 1 – *Valeurs des paramètres de chaque mélange de distributions α-stables.*

4 Application

Dans cette partie, nous allons classifier des données générées et réelles. Tout d'abord, nous présentons et classifions les données générés. Puis nous effectuons la même démarche que précédemment mais avec des données réelles type images sonar.

4.1 Classification de données générées

4.1.1 Présentation des données

Nous simulons 3 mélanges de distributions α-stables dont les valeurs sont indiquées dans le tableau 1. Chaque mélange est constitué de 3 α-stables générées (voir (Chambers et al., 1976)). Les densités de probabilité peuvent être vues comme des attributs correspondant à différentes classes. Dans la suite, nous allons appliquer la théorie des fonctions de croyance développée précédemment pour classifier nos données.

4.1.2 Résultats

Chaque mélange est constitué de 6000 échantillons. Ces échantillons sont divisés en deux parties : l'une sert à la base d'apprentissage et l'autre à la base de test. Lors de la phase d'apprentissage, il est difficile de choisir un modèle permettant d'estimer les distributions sans perdre d'information. Nous estimons chaque attribut à partir d'un mélange de 3 distributions α-stables. L'outil que nous utilisons pour l'estimation du mélange est l'algorithme Expectation-Maximization. Initialement, il a été développé pour estimer un mélange de Gaussiennes (voir (Dempster et al., 1977)). Nous l'avons étendu dans le cas d'un mélange de distributions α-stables (Fiche et al., 2010b). Une estimation à partir d'un mélange de Gaussiennes est effectuée pour avoir des éléments de comparaisons. Comme chaque mélange est estimé à partir de 3 α-stables, soit 12 paramètres à estimer, nous prenons 3 Gaussiennes pour modéliser chaque mélange, soit 12 paramètres à estimer. À partir de l'estimation de chaque mélange, nous allons classifier les éléments de notre base de test. Tout d'abord, nous calculons chaque fonction de plausibilité dans le cas d'un mélange de Gaussiennes et d'α-stables. Le théorème de Bayes généralisée permet de calculer la fonction de masse attribuée à chaque $A \in 2^{\Theta}$. La transformation pignistique permet de travailler sur les singletons. La décision finale est faite en utilisant le maximum de probabilité pignistique. Nous effectuons cette démarche 5 fois. On obtient alors un taux de bonne classification moyenné. Nous estimons aussi chaque mélange avec 3 et 5 Gaussiennes. Les résultats sont représentés dans le tableau 2. On remarque que le taux de classification sous l'hypothèse d'un mélange de distributions α-stables est sensiblement meilleur que les taux de classification sous l'hypothèse de mélange de Gaussiennes. Cependant, on

	Taux de classification	Intervalles de confiance
Mélange de 3 α-stables	60.51	[58.77 ;62.27]
Mélange de 4 Gaussiennes	53.96	[52.19 ;55.75]
Mélange de 3 Gaussiennes	50.79	[49.76 ;52.58]
Mélange de 5 Gaussiennes	57.83	[56.07 ;59.60]

TAB. 2 – *Taux de classification et intervalles de confiance associés à chaque mélange.*

remarque que plus le nombre de Gaussiennes augmente et plus les taux de classification se rapprochent de ceux du mélange de 3 α-stables. L'inconvénient est qu'on augmente le nombre de paramètres à estimer, par exemple 15 paramètres dans le cas d'un mélange de 5 Gaussiennes. On note qu'il est difficile de choisir le bon modèle lors de la phase d'apprentissage. Le choix d'un mélange de distributions α-stables peut être pertinent pour modéliser des données. Dans la partie suivante, nous allons classifier des données réelles en comparant les taux de classification sous l'hypothèse mélange de Gaussiennes et d'α-stables.

4.2 Classification de données réelles

Nous classifions des images sonar en extrayant des paramètres d'Haraclick. Tout d'abord, nous présentons les données et ensuite nous exposons les résultats obtenus.

4.2.1 Présentation des données

Nous disposons d'une base de données de 42 images sonar (Exemple figure 5) fournie par le GESMA (groupes d'Études Sous-Marines de l'Atlantique). Ces images ont subi un prétraitement afin de corriger la variation de gain et réduire le bruit de chatoiement responsable de l'aspect granulaire des images sonar. Des experts ont attribués à chaque pixel des images un type de sédiments. On en retrouve 5 : sable, rides de sable, vase, cailloutis et roche (Exemple figure 6). Chaque image est découpée en imagettes de taille 32×32 pixels. Une imagette est dite "roche" si tous ses pixels sont classifiés "roche". On extrait des paramètres de texture de chaque imagette grâce aux travaux d'Haralick. Nous en choisissons 8 en particulier : moment des différences inverses, corrélation, contraste, moyenne des sommes, moyenne des variances, entropie des sommes, entropie des différences et mesure de corrélation.

4.2.2 Classification des images sonar

On dispose d'une base de plus de 30 000 imagettes classifiées roche, sable, rides de sable, cailloutis et vase. On effectue un tirage aléatoire de 5000 imagettes. Une moitié est utilisée pour l'apprentissage des mélanges et l'autre utilisée comme base de test. La classification de ces données a déjà été étudiée dans Fiche et Martin (2009) où chaque paramètre de texture était estimé à partir d'un mélange de 5 Gaussiennes. Dans notre étude, nous procédons de la manière suivante : nous fixons le nombre de paramètres à estimer à 12, ce qui revient à estimer chaque paramètre de texture à partir d'un mélange de 4 Gaussiennes et d'un mélange de 3 α-stables. Puis, nous faisons varier le nombre de Gaussiennes du mélange : mélange de 3 Gaussiennes, soit 9 paramètres à estimer, et un mélange de 5 Gaussiennes, soit 15 paramètres

FIG. 5 – *Exemple d'images sonar.*

à estimer. On utilise la même démarche que dans le cas des données générées. On obtient alors des taux de classification moyennés sur 5 tirages.
L'estimation des paramètres à partir d'un mélange de Gaussiennes donne un taux de classification de 65.34 %, avec un intervalle de confiance à 95 % de [63.47 ;67.21], tandis que l'estimation des paramètres à partir d'un mélange de distributions α-stables offre un taux de classificatio de 64.58 %, avec un intervalle de confiance à 95 % de [62.70 ;66.65]. On remarque que l'hypothèse mélange de Gaussiennes permet d'avoir un meilleur taux de classification mais ce n'est pas significatif puisque les intervalles de confiance se chevauchent. Nous effectuons la même démarche avec un mélange de 3 Gaussiennes et avec un mélange de 5 Gaussiennes (*c.f.* Tableau 3). On remarque, en diminuant le nombre de Gaussiennes, qu'on obtient un taux de bonne classification meilleur qu'avec le mélange d'α-stables, mais pas significativement. Ceci peut s'expliquer par le fait que les paramètres de texture extraits des images ont une densité de probabilité de type Gaussienne.

(a) ride (b) sable (c) roche (d) cailloutis (e) vase

FIG. 6 – *Exemples d'imagettes.*

	Taux de classification	Intervalles de confiance
Mélange de 3 α-stables	64.58	[62.70 ;66.45]
Mélange de 4 Gaussiennes	65.19	[63.32 ;67.06]
Mélange de 3 Gaussiennes	64.66	[62.78 ;66.53]
Mélange de 5 Gaussiennes	65.34	[63.47 ;67.21]

TAB. 3 – *Taux de classification et intervalles de confiance associés à chaque mélange.*

5 Conclusion

La classification de données générées montre qu'il peut être intéressant d'utiliser un mélange de distributions α-stables par rapport à un mélange de Gaussiennes lors de la phase d'apprentissage, en considérant le même nombre de paramètres à estimer. Cependant, la limite est qu'en augmentant le nombre de Gaussiennes, il est toujours possible d'avoir une estimation correcte des données. Ensuite, il y a une différence entre la théorie et la pratique. En effet, la classification de données réelles montre que les résultats sont significativement les mêmes suivant l'hypothèse de mélange choisi. Dans ce cas, il est difficile de conclure quant à la pertinence du choix du modèle. Les paramètres extraits des images sonar ont une densité de probabilité plutôt type mélange de Gaussiennes ce qui peut expliquer le fait que les résultats soient significativement les mêmes. Il serait intéressant de travailler sur des données qui se prêtent plus aux distributions α-stables pour avoir des taux de classification meilleurs.

Références

Caron, F., B. Ristic, E. Duflos, et P. Vanheeghe (2006). Least Committed basic belief density induced by a multivariate Gaussian pdf. In *9th International Conference on Information Fusion, Florence, Italie*.

Chambers, J., C. Mallows, et B. Stuck (1976). A method for simulating stable random variables. *Journal of the American Statistical Association 71*(354), 340–344.

Delmotte, F. et P. Smets (2004). Target identification based on the transferable belief model interpretation of Dempster-Shafer model. *Systems, Man and Cybernetics, Part A : Systems and Humans, IEEE Transactions on 34*(4), 457–471.

Dempster, A. (1967). Upper and lower probabilities generated by a random closed interval. *The Annals of Mathematical Statistics 38*, 325–339.

Dempster, A., N. Laird, D. Rubin, et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Fiche, A. et A. Martin (2009). Approche bayésienne et fonctions de croyance continues pour la classification. In *Rencontre francophone sur la Logique Floue et ses Applications, Annecy, France*.

Fiche, A., A. Martin, J. Cexus, et A. Khenchaf (2010a). Continuous belief functions and alpha-stable distributions. In *13th International Conference on Information Fusion, Edinburgh, United-Kingdom*.

Fiche, A., A. Martin, J. Cexus, et A. Khenchaf (2010b). Estimation d'un mélange de distributions alpha-stables à partir de l'algorithme em. In *Rencontre francophone sur la Logique Floue et ses Applications, Lannion, France*.

Haralick, R. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE 67*(5), 786–804.

Haralick, R., K. Shanmugam, et I. Dinstein (1973). Textural features for image classification. *IEEE Transactions on systems, man and cybernetics 3*(6), 610–621.

Kernéis, D. (2007). *Amélioration de la classification automatique des fonds marins par la fusion multicapteurs acoustiques*. Ph. D. thesis, Thèse de doctorat de l'ENSTB.

Laanaya, I. (2007). *Classification en environnement incertain : application à la caractérisation de sédiments marins*. Ph. D. thesis, thèse de doctorat de l'UBO.

Lévy, P. (1924). Théorie des erreurs : La loi de gauss et les lois exponentielles. *Bulletin de la Société Mathématique de France 52*, 49–85.

Nolan, J. (1997). Numerical calculation of stable densities and distribution functions. *Communications in Statistics-Stochastic Models 13*(4), 759–774.

Samorodnitsky, G. et M. Taqqu (1994). *Stable non-Gaussian random processes : stochastic models with infinite variance*. Chapman & Hall.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press Princeton.

Smets, P. (1990). Constructing the pignistic probability function in a context of uncertainty. In *Uncertainty in artificial intelligence*, Volume 5, pp. 29–39.

Smets, P. (1993). Belief functions : The disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning 9*(1), 1–35.

Smets, P. (2005). Belief functions on real numbers. *International journal of approximate reasoning 40*(3), 181–223.

Strat, T. (1984). Continuous belief functions for evidential reasoning. In *Proceedings of the National Conference on Artificial Intelligence, University of Texas at Austin*.

Williams, D. (2009). Bayesian data fusion of multiview Synthetic Aperture Sonar imagery for seabed classification. *Image Processing, IEEE Transactions on 18*(6), 1239–1254.

Zolotarev, V. (1986). One-dimensional stable distributions, Translations of Mathematical Monographs, vol. 65. *American Mathematical Society*.

Summary

This paper shows a classification of data based on the theory of belief functions. The complexity of this problem can be seen as two ways. Firstly, data can be imprecise and/or uncertain. Then, it is difficult to choose the right model to represent data. Gaussian model is often used but is limited when data are complex. This model is a particular case of α-stable distributions. Classification is divided into two steps. Learning step allows to modelize data by a mixture of α-stable distributions and Gaussian distributions. Test step allows to classify data with the theory of belief functions and compare the two models. The classification is realized firstly on generated data and then on real data type sonar images.

Apprentissage itératif pour une connaissance *a priori* des labels

Riwal Lefort*,** Ronan Fablet**
Jean-Marc Boucher**

*Ifremer/STH, Technopole Brest Iroise - 29280 Plouzane, France
riwal.lefort@ifremer.fr,
http://www.ifremer.fr
**Telecom Bretagne/LabSTICC, Technopol Brest Iroise
CS83818, 29238 Brest Cedex, France
riwal.lefort@telecom-bretagne.eu
http://www.telecom-bretagne.eu

Résumé. Ce papier synthétise les travaux présentés dans la revue " RNTI-fouille de données complexes, 2010 ". Nous proposons de traiter le problème de l'apprentissage faiblement supervisé pour lequel l'ensemble d'apprentissage est constitué de données de labels inconnus mais dont les probabilités de classification *a priori* sont connues (Lefort et al. (2011)). Premièrement, nous proposons une méthode pour apprendre des arbres de décision à l'aide des probabilités de classification *a priori*. Deuxièmement, une procédure itérative est proposée pour modifier les labels des données d'apprentissage, le but étant que les *a priori* faibles convergent vers des *a priori* forts. Les méthodes proposées sont évaluées sur des jeux de données issus de la base de données UCI.

1 Introduction

Depuis un grand nombre d'années, la classification automatique d'objets a prodigieusement évoluée dans le domaine de la vision par ordinateur. Ces évolutions ont donné naissance à de nouveaux formalismes des modèles de classification dont la complexité d'apprentissage dépend des données d'apprentissage. Les premiers travaux de Fisher (Fisher (1936)) sur les Iris se basent sur une classification supervisée des pétales de fleurs : à partir d'un ensemble de données labélisées, un modèle de classification est établi. Plus tard, naît la classification non-supervisée dont l'objectif est de classer les objets en groupes homogènes (Lloyd (1982)). L'ensemble d'apprentissage est alors constitué de données sans label. L'obtention de données labélisées est souvent difficile et couteuse, dans ce contexte, les méthodes d'apprentissage semi-supervisé montrent que l'ajout de données sans label à des données labélisées permettent d'améliorer les performances de classification (Chapelle et al. (2006)). Cette méthode est performante, mais elle augmente la complexité de l'ensemble d'apprentissage qui est alors constitué à la fois de données labélisées et de données non labélisées (Blum et Mitchel (1998)). Dans le domaine de la vision par ordinateur, quand on classe des objets dans des images, l'annotation des données d'apprentissage s'effectue par la connaissance de la présence et/ou de l'absence

des classes dans les images (Weber et al. (2000)). Cette annotation entraine la constitution d'un ensemble d'apprentissage pour lequel les objets sont associés à des vecteurs qui donnent les *a priori* pour chaque classe (Ulusoy et Bishop (2005)). Certaines applications produisent des ensembles d'apprentissage pour lesquels les *a priori* des classes sont connus. C'est le cas d'annotations directes par un expert (Rossiter et Mukai (2007)), ou encore, en acoustique halieutique dont les *a priori* des classes sont fournis à l'aide de chalutages qui donnent la probabilité des classes dans des images acoustiques de la colonne d'eau (Lefort et al. (2011)).

Dans ce papier, nous proposons un formalisme d'apprentissage faiblement supervisé qui généralise toutes les formes d'apprentissage précédemment citées. Nous nous plaçons dans le cas d'un ensemble d'apprentissage constitué d'objets associés à un vecteur dont les composantes donnent les probabilités de classification pour chaque classe. Soit $\{x_n, \pi_n\}_{1 \leq n \leq N}$ l'ensemble d'apprentissage où $\{x_n\}_n$ représente les objets dans l'espace des descripteurs et $\{\pi_n\}_n$ les vecteurs des probabilités *a priori* tels que $\pi_n = \{p(y_n = c)\}_c$ où $p(y_n = c)$ est la probabilité que l'objet n soit de la classe c. Dans ce contexte, un modèle génératif et un modèle discriminant ont été proposés (Lefort et al. (2011)). Nous proposons une méthode d'apprentissage faiblement supervisée pour les arbres de décision et nous proposons une procédure itérative qui permet de modifier itérativement les *a priori* faibles vers des *a priori* forts.

Après une présentation succincte des modèles de classification élémentaires (section 2) et de la procédure itérative (section 3), nous proposons des résultats de simulations à partir de jeux de données issues de la base de données UCI (Asuncion et Newman) (section 4).

2 Modèles de classification élémentaires

2.1 Modèle génératif

Le modèle génératif modélise la distribution des classes. Plus exactement, nous faisons l'hypothèse que la distribution des classes suit un mélange de M Gaussiennes dont nous estimons les paramètres $\Theta = \left\{\rho_{c1} \dots \rho_{cM}, \mu_{c1} \dots \mu_{cM}, \sigma^2_{c1} \dots \sigma^2_{cM}\right\}_c$. La distribution des points dans l'espace des descripteurs s'écrit alors comme suit :

$$p\left(x|y = c, \Theta\right) = \sum_{m=1}^{M} \rho_{cm} \mathcal{N}(x|\mu_{cm}, \sigma^2_{cm}) \tag{1}$$

$\mathcal{N}(x|\mu_{cm}, \sigma^2_{cm})$ est la loi normale de moyenne μ_{cm} et de matrice de covariance diagonale σ^2_{cm}, m indexant le mode du mélange. Les paramètres Θ du mélange de Gaussiennes sont estimés à l'aide de deux algorithmes EM (Dempster et al. (1977)) imbriqués. La probabilité de classification *a posteriori* s'obtient en utilisant la loi de Bayes. La méthode d'apprentissage est détaillée dans (Lefort et al. (2011)).

2.2 Modèle discriminant

Le modèle discriminant se base sur l'apprentissage d'un hyperplan séparateur, l'unique objectif étant de séparer les données du mieux possible (Schölkopf et Smola (2002)). Les paramètres à estimer sont $\Theta = \{\omega_c, b_c\}$, les coefficients de l'hyperplan séparateur de la classe

c. La probabilité de classification *a posteriori* prend alors l'expression suivante :

$$p\left(y=c|x,\Theta\right)=\frac{exp\left(\langle\omega_c,\Phi(x)\rangle+b_c\right)}{\sum_{j=1}^{C}exp\left(\langle\omega_j,\Phi(x)\rangle+b_j\right)} \tag{2}$$

La méthode d'apprentissage des paramètres Θ, qui s'appuie sur un critère de Fisher pondéré, est détaillée dans (Lefort et al. (2011)). Cette méthode s'étend aisément au cas non-linéaire (Lefort et al. (2011)) (Schölkopf et Smola (2002)).

2.3 Arbres de décision, forêts aléatoires

Les arbres de décision se basent sur une scission dichotomique de l'espace des descripteurs, de telle sorte que les sous-espaces obtenus soient homogènes en classe (Breiman (1996)) (Quinlan (1993)). De tels arbres s'étendent aisément aux forêts aléatoires (Breiman (2001)). La procédure d'apprentissage consiste à trouver, en chaque noeud de l'arbre, un descripteur d et une valeur de coupure S_d qui maximise le gain d'entropie :

$$\arg\max_{\{d,S_d\}}\left(\sum_{q=1}^{2}-\sum_{c}p_{qc}log(p_{qc})\right)-E^0 \tag{3}$$

où E^0 est l'entropie au noeud considéré, q indice les noeuds *fils* et p_{qc} est la probabilité de la classe c dans le noeud *fils* q. Afin d'étendre l'apprentissage d'un arbre au cas de l'apprentissage faiblement supervisé, nous proposons un critère de fusion pour le calcul des probabilités p_{qc} (Lefort et al. (2010a)). Pour le noeud *fils* q_1 qui regroupe les données telles que $\{x_n^d\}<S_d$, la règle de fusion suivante est proposée :

$$p_{q_1 i}\propto\sum_{\{n\}|\{x_n^d\}<S_d}\pi_{ni} \tag{4}$$

Pour le second noeud *fils* q_2 qui regroupe les exemples tels que $\{x_n^d\}\geq S_d$, une règle équivalente de fusion est proposée :

$$p_{q_2 i}\propto\sum_{\{n\}|\{x_n^d\}\geq S_d}\pi_{ni} \tag{5}$$

L'extension au cas de forêts aléatoires s'effectue aisément, en moyennant les probabilités de classification issues de chaque arbre (Lefort et al. (2010b)).

3 Procédure itérative

Une procédure itérative est proposée pour modifier les *a priori* des données d'apprentissage. En effet, il semble naturel de combiner les données d'apprentissage afin de rendre les données imprécises plus précises. Cela est effectué en classification semi-supervisée telle que les données sans label sont labélisées itérativement à l'aide des données labélisées (Blum et Mitchel (1998)) (Chapelle et al. (2006)). L'idée est donc la suivante : à une itération donnée

j, les données d'apprentissage T^j apprennent un classifieur H^j qui permet de calculer une probabilité de classification *a posteriori* de l'ensemble d'apprentissage. Les *a priori* π^{j+1} de l'ensemble d'apprentissage T^{j+1} sont mis à jour en fusionnant l'*a priori* initial π^0 avec la probabilité de classification *a posteriori* $p(y_n = i|x_n, H^j)$ de la manière suivante :

$$\pi_n^{j+1} \propto \pi_n^0 p(y_n = i|x_n, C_m) \tag{6}$$

Ce critère de fusion a pour but d'exploiter toutes les informations relatives aux données d'apprentissage : les informations *a priori* initiales et les informations de classification issues des classifieurs succesifs.

L'inconvénient de cette procédure intuitive est le phénomène de sur-apprentissage. En effet, les données classées par H^j sont celles qui ont appris H^j. Le classifieur H^j est donc façonner pour être uniquement efficace sur les données d'apprentissage. Pour supprimer ce phénomène de sur-apprentissage, nous proposons à chaque itération de scinder l'ensemble d'apprentissage T^j en deux sous-groupes T_r^j et T_t^j, l'un permettant d'apprendre le classifieur H^j, l'autre est classé par H^j, puis nous appliquons la règle de mise à jour (6) avant de concaténer les deux sous groupes pour former l'ensemble d'apprentissage de l'itération suivante $Tj + 1 = \{T_r^j, T_t^{j+1}\}$. Cette procédure itérative est schématisée dans la figure 1.

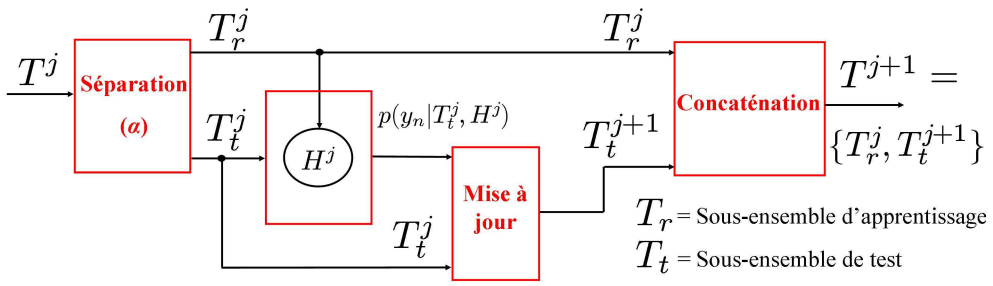


FIG. 1 – *Schéma de la procédure itérative améliorée qui permet d'éviter le phénomène de sur-apprentissage.*

4 Simulation

Afin de maitriser la complexité des données d'apprentissage, des ensembles d'apprentissage faiblement supervisés sont générés à partir de données supervisées provenant de la base de données UCI (Asuncion et Newman) (Lefort et al. (2010a)) (Lefort et al. (2010b)). Quatre types de données sont sélectionnés (Lefort et al. (2010a)) (Lefort et al. (2010b)). Le premier jeu de données (D1) est " Data Segmentation " qui contient 7 classes de texture d'images (ciel, mure, porte, chemin, etc). Chaque image est décrite par un ensemble de 19 descripteurs d'image. Le second jeu de données (D2) est les Iris de Fisher (Fisher (1936)). Le troisième jeu de données (D3, " Synthetic Control Chart Time Series ") est composé de 6 classes de courbes (croissantes, décroissantes, cycliques, etc), les descripteurs de chaque courbe étant la valeur des 60 échantillons de la courbe. Enfin, les formes d'onde de Breiman (Breiman et al. (1984))

sont classées (D4). 3 classes d'ondes sont construites à partir de combinaisons linéaires de courbes bruitées. Chaque onde est décrite par un ensemble de 19 descripteurs.

Dans la figure 2, nous reportons le taux moyen de bonne classification sur l'ensemble des jeux de données pour le modèle génératif (Gen), le modèle discriminant (Discr), les forêts aléatoires (FA) et la procédure itérative associée au modèle discriminant (FA+Iter). Nous faisons varier la complexité du cas de l'apprentissage supervisé au cas de l'apprentissage non supervisé en passant par deux cas faiblement supervisés. Pour l'apprentissage supervisé, la classe des données d'apprentissage générées est connue. Pour l'apprentissage non-supervisé, les classes sont équiprobables, ce qui va engendrer des performances médiocres, étant donné qu'il n'y a aucune connaissance *a priori* sur les classes. Les deux cas intermédiaires proposent des situations pour lesquelles les probabilités de classification *a priori* des données d'apprentissage prennent des valeurs plus ou moins certaines. Pour le premier cas intermédiaire, les probabilités de classification *a priori* sont assez certaines. Par exemple, pour deux classes, les *a priori* peuvent prendre les valeurs [0,9 0,1]. Pour le second cas intermédiaire, des *a priori* moins certains sont attribués aux données d'apprentissage, par exemple [0,6 0,4] dans le cas de deux classes. Cette expérience permet de mesurer la robustesse des méthodes proposées relativement à la complexité des données d'apprentissage.

Premièrement, quel que soit le modèle de classification, plus les *a priori* des données d'apprentissage sont forts, meilleurs sont les résultats de classification. Les résultats indiquent que sans aucune information *a priori* sur les classes, les performances de classification sont médiocres. Ce résultat est logique, du fait que les modèles proposés se basent sur l'utilisation des *a priori* pour apprendre la distribution des classes (modèle génératif), un séparateur des classes (modèle discriminant), ou la distribution hiérarchique des classes (arbre de décision). En revanche, plus la connaissance des classes est certaine, plus la modélisation est bonne, donnant de très bons résultats de classification pour les forêts aléatoires. Deuxièmement, nous constatons que la procédure itérative (FA+Iter) permet d'améliorer nettement les performances de classification par rapport aux forêts aléatoires seules. Cela est dû au filtrage itératif des données incertaines qui permet une meilleure estimation du modèle de classification final.

5 conclusion

Dans ce papier, nous proposons un formalisme original d'apprentissage faiblement supervisé qui généralise les autres formes d'apprentissage. Nous considérons que l'ensemble d'apprentissage est constitué d'objets auxquels sont associés des vecteurs qui donnent les probabilités de classification *a priori* pour chaque classe.

Trois modèles de classifications élémentaires sont étudiés (un modèle génératif, un modèle discriminant et les forêts aléatoires), puis nous proposons une procédure itérative d'apprentissage des forêts aléatoires.

Les performances de classifications montrent comment le filtrage itératif des données d'apprentissage permet d'améliorer les performances de classification.

Références

Asuncion, A. et D. Newman. Uci machine learning repository. *http ://www.ics.uci.edu/∼mlearn/MLRepository.html*.

Blum, A. et T. Mitchel (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual Conference on Computational Learning Theory*, 92–100.

Breiman, L. (1996). Bagging predictors. *Machine Learning 26*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984). *Classification and regression trees*. Chapman & Hall.

Chapelle, O., B. Schölkopf, et A. Zien (2006). *Semi-supervised learning*. MIT Press.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Jour. of the RSS 39, Series B*(1), 1–38.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 179–188.

Lefort, R., R. Fablet, et J.-M. Boucher (2010a). Weakly supervised classification of objects in images using soft random forests. *European Conference on Computer Vision (ECCV)*.

Lefort, R., R. Fablet, et J.-M. Boucher (2010b). Weakly supervised classification with decision trees applied to fisheries acoustics. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Lefort, R., R. Fablet, et J.-M. Boucher (2011). Object recognition using proportion-based prior information : application to fisheries acoutics. *Pattern Recognition Letters 32*(2), 153–158.

Lloyd, S. (1982). Least square quantization in pcm. *IEEE Transactions on Information Theory 28*(2), 129–137.

Quinlan, J. (1993). C4.5 : Programs for machine learning. *Morgan Kaufmann Publishers*.

Rossiter, J. et T. Mukai (2007). Bio-mimetic learning from images using imprecise expert information. *Fuzzy Sets and Systems 158*(3), 295–311.

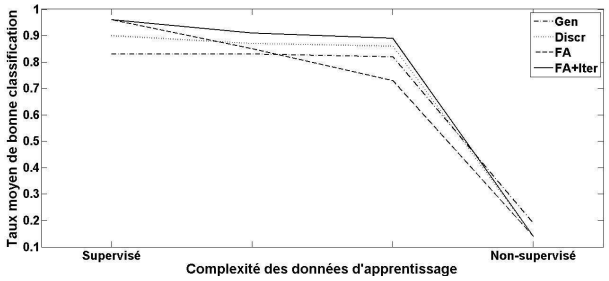
Schölkopf, B. et A. Smola (2002). *Learning with Kernels*. The MIT Press.

Ulusoy, I. et C. Bishop (2005). Generative versus discriminative methods for object recognition. *International Conference on Computer Vision and Pattern Recognition 2*, 258–265.

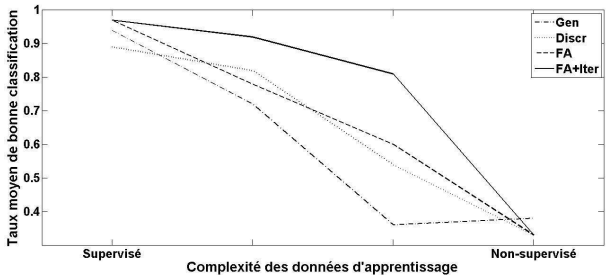
Weber, M., M. Welling, et P. Perona (2000). Towards automatic discovery of object categories. *International Conference on Computer Vision and Pattern Recognition 2*, 101–108.

Summary

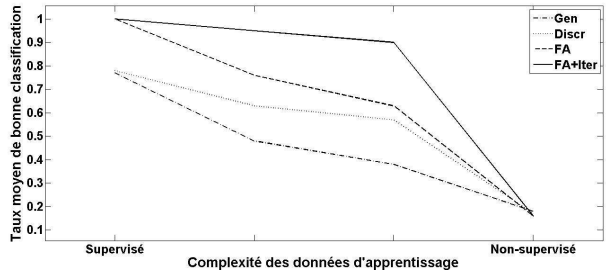
In the field of data mining, depending on the training data complexity, several kinds of classification scheme exist. This paper deals with weakly supervised learning that generalizes the supervised and semi-supervised learning. In weakly supervised learning training data are given as the priors of each class for each sample. We first propose a weakly supervised strategy for learning soft decision trees. Besides, the introduction of class priors for training samples instead of hard class labels makes natural the formulation of an iterative learning procedure. The iterative procedure makes prior refined every iteration. We report experiments for UCI object recognition datasets. These experiments show that recognition performance close to the supervised learning can be expected using the proposed framework. We further discuss the relevance of weakly supervised learning for fisheries acoustics applications.



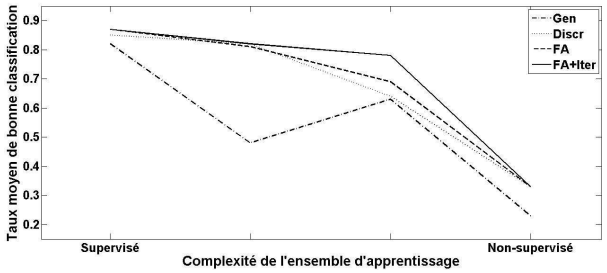
(a)



(b)



(c)



(d)

FIG. 2 – Taux moyen de bonne classification en fonction de la complexité du jeu de données d'apprentissage pour les jeux de données D1 (a), D2 (b), D3 (c), et D4 (d), et pour le modèle génératif (Gen), le modèle discriminant (Discr), les forêts aléatoires (FA), et les forêts aléatoires associées au processus itératif (FA+Iter).

Utilisation de méthodes d'évaluation de sources d'information dans le cadre de la théorie des fonctions de croyance pour une application réelle

Sébastien Régis*, Jérémy Frominville*
Andrei Doncescu** Martine Collard*

* LAMIA, Université des Antilles et de la Guyane Campus de Fouillole 97159 Pointe-à-Pitre cédex
sregis, mcollard@univ-ag.fr, j.e.fromin@gmail.com
http://lamia.univ-ag.fr/
**LAAS-CNRS 7, avenue du Colonel Roche 31077 Toulouse Cedex 4
adoncesc@laas.fr
http://www.laas.f

Résumé. Dans cet article nous étudions deux mesures du conflit de la théorie des fonctions de croyance proposées dans deux approches globalement similaires qui consistent à évaluer la qualité d'une source d'informations. Les tests empiriques montrent que les deux mesures donnent des résultats globalement proches.

1 Introduction

La multiplication et la grande variété des capteurs physiques dans des applications courantes (comme les bioprocédés (Régis et al. (2007)) par exemple) se sont généralisés ces dernières années et le recours à des méthodes de fusion d'informations mutli-sources est de plus en plus systématique. L'utilisation de la théorie des fonctions de croyance dans ces domaines est de plus en plus répandue puiqu'elle permet, entres autres, une agrégation des informations issue de sources différentes, tout en gérant des informations incertaines et incomplètes. La théorie des fonctions de croyance, proposée par Dempster et Shafer connait ainsi un succès croissant dans des applications réelles, bien que la notion de conflit utilisée dans la fusion d'informations conduit parfois à des résultats erronés Zadeh (1984). Diverses formules de fusion ont été proposées pour pallier ce problème ; parmi celles-ci, on peut citer celles proposées par Yager (1987), Dubois et Prade (1988), Smets (1988), Smets et Kennes (1994) ou celle de Lefevre et al. (2002). Une autre approche consiste à utiliser le conflit pour caractériser la qualité des sources d'informations.
Cette approche a été proposée indépendamment par Régis et al. (2004, 2007) et par Chebbah et al. (2010) pour estimer la qualité des sources d'informations mais avec des applications et des objectifs pratiques sensiblement différents. La différence se fait essentiellement au niveau du calcul du conflit. Cependant les deux approches ont le même objectif : caractériser une source d'informations sans *a priori* en se basant sur les informations intrinsèques aux sources et sur l'utilisation du conflit de la théorie des fonctions de croyance. De plus, dans chacune des

méthodes une mesure du conflit est proposée. Régis et al. (2004, 2007) proposent d'évaluer le conflit par une distance entre les fonctions de masses de deux sources tandis que Chebbah et al. (2010) utilisent la distance de Jousselme Jousselme et al. (2001). Cet article propose une comparaison empirique de ces deux mesures du conflit. Il s'agit surtout de comparer les résultats obtenus par chacune des deux méthodes de conflits sur un même exemple.Nous montrons en effet que les résultats sont semblables sur le cas pris en tant qu'exemple.
L'article est structuré comme suit. Après avoir rappelé brièvement les principes de la théorie des fonctions de croyance dans la deuxième section, les deux approches sont présentées dans la section 3. L'exemple utilisé pour le test est présenté dans la section 4 ainsi que les résultats, avant la conclusion en section 5.

2 Théorie des fonctions de croyance

La théorie des fonctions de croyance est une généralisation de la théorie bayésienne qui tient compte des notions d'incertitude et d'imprécision de l'information. Elle a été introduite par Dempster (1968) puis a été formalisée mathématiquement par Shafer (1976).
Considérons l'ensemble de tous les évènements possibles (on parle d'ensemble de toutes les hypothèses) ; cet ensemble est appelé *ensemble de discernement* et est noté Θ. Toutes ces hypothèses sont mutuellement exclusives et sont nommées *singletons*. La théorie des fonctions de croyance porte sur l'ensemble des sous-ensembles A de Θ. Cet ensemble de parties de Θ est noté 2^{Θ}. Une partie A peut être composée d'un singleton ou d'une union de plusieurs singletons. Une fonction de masse m peut être alors définie de 2^{Θ} vers [0,1] avec les propriétés suivantes :

$$\sum_{A \subset \Theta} m(A) = 1 \qquad (1)$$
$$m(\oslash) = 0$$

$m(A)$ est la masse d'évidence associée à A.
Les fonctions de *plausibilité* (Pl) et de *croyance* (Bel) sont définies de 2^{Θ} vers [0,1] comme suit :

$$pl(A) = \sum_{B \cap A \neq \oslash} m(B) \qquad (2)$$
$$bel(A) = \sum_{B \subset A, A, B \in 2^{\Theta}} m(B)$$

Pour obtenir une fusion de l'information de deux sources différentes 1 et 2, une combinaison de leur masses d'évidence appelée règle de Dempster a été définie comme suit :

$$(m_1 \oplus m_2)(A) = m_{1,2}(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B).m_2(C) \quad A, B, C \subset 2^{\Theta}, A, B, C \neq \oslash \qquad (3)$$

où K est défini comme suit :

$$K = \sum_{B \cap C = \oslash} m_1(B).m_2(C) \qquad (4)$$

Le dénominateur $1 - K$ est un facteur de normalisation. Plus précisément K représente la mesure du conflit entre les sources 1 et 2. Plus K est important, plus les sources sont en conflit et moins la fusion a de sens. Si $K = 1$ alors le conflit est total et la fusion n'a pas de sens. On peut généraliser la règle de Dempster à n sources :

$$(\oplus m_i)_{i=1,\ldots,n}(A) = \frac{1}{1-K} \sum_{X_1 \cap \ldots \cap X_n = A} \left(\Pi_{i=1}^{n} m_i(X_i)\right) \quad (5)$$

$$A, X_i \subset 2^{\Theta}; A, X_i \neq \oslash$$

$$K = \sum_{X_1 \cap \ldots \cap X_n = \oslash} \left(\Pi_{i=1}^{n} m_i(X_i)\right)$$

Si les sources sont en conflit fort (K est grand) alors la règle de Dempster peut conduire à des résultats erronés(Zadeh (1984)). La raison de ce comportement de la règle de Dempster provient du fait que la masse d'évidence affecte l'ensemble vide est nulle. Cette contrainte $m(\emptyset) = 0$ implique que l'intersection entre deux hypothèses a une masse nulle. Partant de cette contrainte, deux points de vue sont alors possibles :

- soit l'on travaille dans un *monde fermé*. On considère que les hypothèses décrivent totalement le problème à résoudre. Dans ce cas la solution au problème de classification se trouve forcément parmi les hypothèses données. C'est le point de vue classique de la théorie des fonctions de croyance.
- soit l'on travaille dans un *monde ouvert*, et dans ce cas les hypothèses modélisent partiellement le problème à résoudre. Soit la solution au problème de classification se trouve parmi les hypothèses données, soit il s'agit d'une nouvelle hypothèse omise ou du moins inconnue. Ce point de vue semble *en général* plus réaliste par rapport aux applications pratiques.

Quoiqu'il en soit, plusieurs alternatives ont été proposées pour pallier le manque de cohérence des résultats de la théorie des fonctions de croyance :

- la théorie développée principalement par P. Smets qui suppose que $m(\emptyset) \geq 0$.Cette supposition repose sur l'hypothèse du monde ouvert. Cette approche a été surtout développée et utilisée dans le cadre des probabilités pignistiques(Smets (1988)Smets et Kennes (1994)).
- la combinaison proposée par Yager (1987) qui affecte la masse de tous les conflits à la masse de l'ensemble Ω. Cette approche repose sur l'hypothèse d'un monde fermé.
- La théorie développée par J. Dezert et F. Smarandache (théorie de Dezert-Smarandache) qui suppose que l'ensemble de discernement n'est pas formé uniquement de singletons mutuellement exclusifs mais également d'unions et d'intersections non vides de ces singletons
  (Dezert et Smarandache (2003), Dezert et Smarandache (2004b), Dezert et Smarandache (2004a)). Cette théorie repose sur l'hypothèse du monde fermé.
- la combinaison proposée par Lefevre et al. (2002) qui consiste à pondérer les différentes masses d'évidence. Cette approche peut être utilisée quelle que soit l'hypothèse retenue (monde fermé ou monde ouvert).
- le déconditionnement qui peut être vu comme intermédiaire entre les deux premières alternatives puisqu'il utilise à la fois la notion de monde ouvert et de monde fermé mais

avec des échelles différentes (Bracker (1996)). En effet chaque source est considérée comme un cadre local et non exhaustif (donc comme un monde ouvert) mais l'ensemble Θ est lui considéré comme un cadre global exhaustif (donc comme un monde fermé). Le déconditionnement consiste à exprimer les masses d'évidence données par rapport à un cadre local, dans le cadre global. Il nécessite cependant de donner des informations *a priori* (sous forme de probabilités par exemple) sur les évènements par rapport au cadre global (voir Bracker (1996)).

3 Caractérisation des sources d'informations par le conflit

3.1 Utilisation du conflit pour caractériser les sources d'informations

Dans la plupart des alternatives proposées dans la section précédente, on estime que le conflit est un facteur pouvant générer des erreurs de calcul lors de la fusion. Ce conflit est donc considéré comme un problème à contourner et non comme un indicateur servant à évaluer les sources d'informations. Pourtant cette notion de conflit apporte des renseignements sur les sources elles-même en se basant uniquement sur les données. Schubert (1993, 1995, 1996)) a été l'un des premiers à utiliser le conflit de la théorie de Dempster-Shafer dans ce sens. Son objectif n'était pas de caractériser les sources d'informations mais d'utiliser le conflit pour définir des regroupements de sources d'informations en fonction de ce conflit : les sources qui ont un conflit faible ou nul sont ainsi regroupées dans un même cluster.
Plus récemment, Régis et al. (2004); Régis (2004); Régis et al. (2007) et Chebbah et al. (2010) ont proposé indépendamment d'utiliser le conflit pour estimer la qualité des sources d'informations : il ne s'agit pas ici de faire des regroupements de sources mais d'évaluer directement leur qualité. La différence entre les deux méthodes réside en partie dans la définition de la qualité d'une source. Pour Régis et al. (2007) le critère repose sur la pertinence de la source d'informations, alors que pour Chebbah et al. (2010) il s'agit d'évaluer sa fiabilité.

3.2 Présentation des deux méthodes d'évaluation

3.2.1 Pertinence versus Fiabilité

Les idées sous-jacentes aux travaux de Régis et al. (2007) et de Chebbah et al. (2010) sont similaires : on considère une source d'informations, on calcule son conflit avec les autres sources. Si le conflit est globalement faible ou nul avec les autres sources, la source est de grande qualité, sinon (i.e. si le conflit est important) la source n'est pas de grande qualité. Reste à définir la notion de qualité (on reviendra aussi ultérieurement sur le sens du terme "globalement"). Pour Régis et al. (2007), la qualité est liée à la *pertinence* d'une source, tandis que pour Chebbah et al. (2010) la qualité est liée à la *fiabilité*.
La pertinence reste une notion relativement subjective (il n'y a certainement pas de définition générique de la pertinence) mais elle est pourtant proposée et manipulée dans divers domaines liées à l'informatique : classification (Lazo-Cortès et Ruiz-Schulcloper (1995),Blum et Langley (1997)), détection de fautes(Baluja et Pomerleau (1997)), intelligence des moteurs de recherche (Zadeh (2004)). La définition de cette notion de pertinence varie sensiblement en fonction des disciplines et de l'application. Pour Lazo-Cortès et Ruiz-Schulcloper (1995), la pertinence d'une variable dépend de sa capacité à discriminer différentes classes. Pour Zadeh

(2004), la pertinence globale d'un paramètre est fonction des informations fournies par les autres paramètres disponibles. Blum et Langley (1997) proposent au moins 5 définitions différentes pour la pertinence. Nous prendrons la définition de la pertinence tirée d'un dictionnaire(Larousse (2004)) :
"La pertinence est le caractère de ce qui est approprié, de ce qui se rapporte exactement à ce dont il est question." En nous basant sur cette définition et en tenant compte de notre domaine d'application, la classification, nous proposons donc la définition suivante pour la pertinence des signaux (voir Régis et al. (2007)).

Une source est *pertinente* en terme de classification si elle :

- n'induit pas de résultat aberrant.
- fournit une information significative pour la classification.
- génère des décisions en accord avec la plupart des autres signaux.

La première caractéristique est liée à la présence d'artefacts dans les sources. La deuxième traduit le fait que l'objectif est d'obtenir une classification correcte et montre qu'un signal pertinent a une influence sur la classification.
La dernière repose sur l'hypothèse que la majorité des signaux traduit la vérité au moins dans une certaine mesure. Le fait de considérer que la majorité des signaux traduit la vérité peut apparaître comme une hypothèse "radicale" mais elle est basée sur le fait que dans de nombreuses applications les signaux sont issus de capteurs qui observent et traduisent le même phénomène.
Chebbah et al. (2010) propose d'utiliser le concept de fiabilité pour caractériser la qualité d'une source d'informations (sans toutefois préciser la définition de cette fiabilité). Nous donnons là aussi la définition de la fiabilité tirée du même dictionnaire (Larousse (2004)) :

"La fiabilité est la probabilité de fonctionnement sans défaillance d'un dispositif dans des conditions spécifiées et pendant une période de temps déterminée."
Cette notion de fiabilité est donc proche de celle de pertinence mais cette dernière nous semble *a priori* plus générale et plus appropriée pour définir la qualité d'une source d'informations. En effet, la fiabilité est liée à l'absence de défaillance technique (ou informatique) d'une source alors que la pertinence est surtout liée à l'objectif final (classification, clustering, etc.) et englobe la fiabilité. Ainsi une source peut fonctionner correctement et ne pas détecter un phénomène si ses capteurs n'ont pas été concus dans ce but. Par exemple la vision humaine est une source d'information fiable mais elle ne peut pas capter les sources de chaleurs comme le font les signaux à infrarouge. La vision humaine bien que fiable, n'est donc pas pertinente pour détecter une source de chaleur.
De ce fait, malgré une similitude évidente, la notion de pertinence est de notre point de vue plus adaptée pour définir la qualité d'une source.
Par ailleurs il faut aussi souligner le fait que les deux approches ont été développées dans des contextes applicatifs différents. Ainsi la méthode proposée par Chebbah et al. (2010) a été développée pour caractériser des bases de données évidentielles, et elle fournit une fiabilité globale de la source d'informations : cette fiabilité ne varie pas au cours du temps. La méthode proposée par Régis et al. (2004, 2007) a été développée pour estimer la pertinence de variables biochimiques dont les valeurs sont mesurées au cours d'un bioprocédé fermentaire, et elle fournit une fiabilité locale qui peut évoluer au cours du temps.

3.2.2 Calcul du conflit

Quoiqu'il en soit l'idée de base des deux approches est d'utiliser le conflit pour caractériser une source. Et si une source est considérée comme non pertinente (ou non fiable) alors elle est amoindrie par la méthode de l'affaiblissement (Shafer (1976), Bloch (2005), Appriou (2002)). Le point le plus délicat consiste à calculer le conflit en proposant une alternative à la combinaison de Dempster pour éviter des résultats erronés (voir Zadeh (1984), Régis et al. (2007)). Ainsi Régis et al. (2007) proposent de remplacer la mesure du conflit par une distance métrique. La mesure du conflit est définie comme suit :

$$conf_1(S_1, S_2) = d_{met}(S_1, S_2) = \frac{1}{2}\sum_i |m_1(A_i) - m_2(A_i)| \quad A_i \subset 2^\Theta \tag{6}$$

où $m1$ et $m2$ représente les deux fonctions de masses pour les sources S_i et S_j. Le facteur $\frac{1}{2}$ est un facteur de normalisation.
L'intérêt de cette mesure du conflit est de mesurerla différence de distribution des masses d'évidence entre les deux sources. Son principal défaut est de n'être utilisable que dans la mesure où les fonctions de masses de chacune des sources sont toutes réparties sur les mêmes éléments focaux (dans la pratique c'est souvent le cas mais on peut trouver des cas réels ou simulés où les sources ne travaillent pas sur les mêmes éléments).
Chebbah et al. (2010) utilisent comme mesure du conflit la distance de Jousselme et al. (2001). Celle-ci permet de tenir compte des spécificités des fonctions de croyance puisque cette distance utilise le coefficient de Jaccard $\frac{|A_i \cap A_j|}{|A_i \cup A_j|}$ où A_i et A_j sont deux éléments focaux. Ce coefficient de Jaccard permet de tenir compte des cardinalités des éléments focaux. Une matrice D des coefficient de Jaccard est ainsi définie sur l'ensemble 2^Θ ce qui rend cette distance spécifique aux fonctions de croyance. La distance de Jousselme est donnée par :

$$d_{jous}(m_1, m_2) = \sqrt{\frac{1}{2}.(m_1 - m_2)^t D(m_1 - m_2)} \tag{7}$$

avec

$$\begin{aligned} D(A_i, A_j) &= 1 \quad si \quad A_i = A_j = \emptyset \\ &= \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \quad \forall A_i, A_j \subset 2^\Theta \end{aligned} \tag{8}$$

où m_1 et m_2 sont les deux fonctions de masse des sources S_1 et S_2. La mesure du conflit entre S_1 et S_2 est donnée par :

$$conf_2(S_1, S_2) = d_{jous}(m_1, m_2) \tag{9}$$

L'avantage évident de cette mesure est de pouvoir être utilisée sans condition sur la répartition des masses sur les éléments focaux. Elle pourra donc être utilisée dans tous les cas réels que les sources travaillent ou non sur les mêmes éléments focaux. L'intérêt de la distance de Jouselme est d'être toujours utilisable elle sera donc préférée à la distance métrique pour les tests expérimentaux.

Régis et al. (2007) ont défini une source comme étant en conflit avec une autre source dans le cas où la mesure de ce conflit était supérieure à un seuil donné (ce seuil était *a priori* égal

à 0.5 mais dans la pratique il pouvait être inférieur ou supérieur à 0.5). Selon cette approche, une source est considérée comme non pertinente si elle est en conflit deux à deux avec plus de la moitié des autres sources.
Chebbah et al. (2010) ont proposé de mesurer deux à deux le conflit entre une source et les autres sources puis de considérer la valeur moyenne de ces mesures comme un pourcentage de fiabilité. La valeur de cette moyenne permet alors de dire si une source est fiable ou non.
L'approche proposée par Chebbah et al. (2010) nous paraissant plus simple et plus naturelle, elle a été retenue pour les tests expérimentaux.
Ainsi le conflit moyen d'une source S_i est définie comme suit :

$$ConfMoy(S_i) = \frac{1}{n-1}. \sum_{j,j\neq i} conf(S_i, S_j) \tag{10}$$

où n désigne le nombre de sources.

4 Résultats expérimentaux

Nous avons comparé les deux types de conflit sur un même exemple concret.
Il faut noter que nous nous sommes surtout intéressés à l'impact de l'utilisation de la pertinence sur les résultats en terme de classifications correctes. Nous utilisons les variables biochimiques issues d'une fermentation alcoolique (on parle aussi de bioprocédé). Les paramètres biochimiques mesurés pendant l'expérimentaion représentent donc les sources d'information pour effecter une classification des données. Le regroupement des paramètres en "paquets" permet de caractériser les états physiologiques des micro-organismes par une ou plusieurs classes. La classification permet de caractériser une source et d'anticiper d'éventuels problèmes. Ces paramètres biochimiques se présentent sous la forme de séries temporelles. La classification consiste donc à segmenter les séries temporelles de telle sorte qu'une classe ou un groupe de classes consécutives correspondent à un état physiologique donné. En fait la classification peut être réalisée "manuellement" par un expert en microbiologie (voir figure 1). On cherche donc à se rapprocher le plus possible du travail réalisé manuellement par l'expert en utilisant des méthodes automatiques de classification. Ces méthodes de classification ont fourni des résultats intéressants(Waissman-Vilanova (2000), Goma et al. (2004), Régis (2004)). Pour le bioprocédé, l'expérience dure environ 20 heures et correspond à 1012 points de mesures des paramètres biochimiques. On considère le début du bioprocédé comme étant à l'heure t=0. On rappelle que l'on cherche à détecter trois états physiologiques principaux :

- l'état 1 : la fermentation (production d'éthanol). Elle va de 0h jusqu'à environ 9h ce qui représente un total de 590 points mesurés.
- l'état 2 : la diauxie. Cet état commence à environ 9h et se termine à 9h46 ce qui représente environ 33 points de mesure. C'est le plus petit état physiologique (en temps et en quantité de données) parmi les 3 et le plus difficile à caractériser
- l'état 3 : l'oxydation (production de biomasse). Elle commence à 9h46 et se termine en même temps que la fin de l'expérience à 20h ce qui représente 389 points de mesure.

Il y a 20 paramètres biochimiques et chacun d'eux a donc 1012 éléments. Ces paramètres biochimiques sont donc considérés comme les sources d'informations et les différentes pertinences de ces paramètres biochimiques sont donc évaluées à chaque instant t. Pour définir les

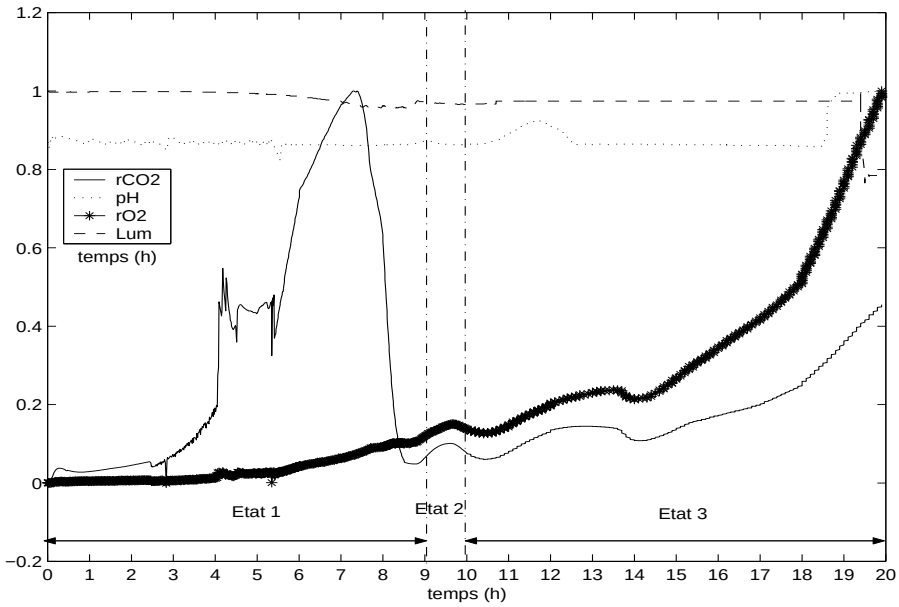


FIG. 1 – *Etats physiologiques fournis par les experts sur un bioprocédé de type batch. L'axe des abscisses représente le temps exprimé en heures, l'axe des ordonnées représente l'amplitude des paramètres biochimiques (les valeurs ont été normalisées). 4 paramètres ont été utilisés : le pH, la vitesse de consommation de l'oxygène (rO2), la vitesse de production de dioxyde de carbone (rCO2) et la luminance (Lum) qui traduit la production de biomasse. On distingue 3 états : la fermentation (état 1), la diauxie (état 2), l'oxydation (état 3).*

masses d'évidence, et puisque les classes étaient connues, nous avons utilisé la méthode de calcul des masses proposée par Denoeux (1995) et basée sur la méthode des k plus proches voisins. 68 échantillons ont été testés (30 pour l'état 1, 7 pour l'état 2 et 31 pour l'état 3). Le calcul des masses d'évidence se fait comme suit : pour chacune des 3 classes C_1, C_2, C_3 et pour l'ensemble Θ la formule est donnée par les équations ci-dessous :

$$m(C_i) = \frac{m_i(C_i) \prod_{j \neq i} m_j(\Theta)}{K} \tag{11}$$

$$m(\Theta) = \frac{\prod_{i=1}^{3} m_i(\Theta)}{K} \tag{12}$$

où K est le facteur de normalisation suivant :

$$K = \sum_{i=1}^{3} \prod_{j \neq i} m_j(\Theta) + (1-3) \prod_{i=1}^{3} m_i(\Theta) \tag{13}$$

Avec :

$$m_i(C_i) = 1 - \prod_{x_{ki} \in C_i} (1 - \alpha_0 e^{-d^{ki,l}}) \tag{14}$$

$$m_i(\Theta) = \prod_{x_{ki} \in C_i} (1 - \alpha_0 e^{-d^{ki,l}}) \tag{15}$$

où $d^{ki,l}$ représente la distance métrique entre chaque mesure à classer x_l et chaque échantillon étiqueté x_{ki} de la classe C_i ($i \in \{1, 2, 3\}$) pour chaque paramètre biochimique ; e représente la fonction exponentielle et α_0 est une valeur fixée comprise entre 0 et 1 (Denoeux (1995)). Pour l'application, nous avons choisi empiriquement $\alpha_0 = 0.95$ et le nombre de plus proches voisins est égal à 7 ($k = 7$).
L'approche que nous proposons peut se résumer comme suit :

- Pour chaque instant donné t :

1. Caractérisation de la pertinence
   - Pour chaque paramètre P
     - calcul du conflit moyen du paramètre P
     - si le conflit moyen du paramètre est inférieur à un seuil donné τ, alors il est pertinent
     - sinon, il est "non pertinent"
2. Fusion de l'information par la combinaison de Dempster de tous les paramètres avec :
   - un poids égal à 1 pour tous les paramètres pertinents
   - un poids égal à 0.5 pour tous les paramètres non pertinents
   - pour l'ensemble Θ, la masse d'évidence m(Θ) reste inchangée lors de la fusion pour les paramètres pertinents, tandis que pour les paramètes non pertinents m(Θ) est mise à jour avec $m(\Theta) = 0.5 \times (1 - m(\Theta))$

Les tests ont été réalisés en utilisant le conflit moyen avec chacune des distances $conf_1$ (basée sur la distance métrique) et $conf_2$ (basée sur la distance de Jousselme). Il est possible

d'utiliser $conf_1$ car dans cet exemple tous les paramètres (i.e. toutes les sources d'informations) travaillent sur les mêmes éléments focaux. Les résultats de pourcentages de classification correcte sont donnés dans le tableau1. Plusieurs valeurs du seuil τ ont été testées. La colonne de gauche du tableau donne les valeurs du seuil τ, les deux colonnes suivantes donnent les pourcentages de classification correcte respectivement pour la distance de Jousselme et pour la distance métrique.

valeur seuil τ	distance de Jousselme	distance métrique
0,4	97%	99%
0,5	74%	88%
0,6	73%	75%
0,7	75%	75%
pas de pertinence/fiabilité	69%	69%

TAB. 1 – *Pourcentage de classification correcte en fonction de la valeur du seuil τ.*

Une première constation est que les résultats sont similaires quelque soit le conflit utilisé : à part le seuil de $0,5$ où il y a une différence de 14 points entre les pourcentages, la différence n'excède pas 2 points pour les autres valeurs du seuil τ. On constate également que quelque soit la méthode, l'utilisation d'une évaluation de la pertinence (ou de la fiabilité) améliore sensiblement les résultats de classification.

5 Conclusion

Nous avons présenté dans cet article deux approches pour évaluer la qualité d'une source d'information dans le cadre de la théorie des fonctions de croyance. Ces deux approches sont similaires et se basent sur l'utilisation du conflit pour estimer la qualité des sources. Des différences existent entre ces deux approches notamment au niveau de la définition des concepts, de l'évaluation numérique du conflit, mais l'expérience menée a fourni des résultats qui sont globalement très proches. On observe que quelque soit l'approche, l'évaluation de la pertinence (ou fiabilité) permet d'améliorer les résultats de la classification. L'intêrèt de la distance de Jousselme qu'utilisent Chebbah et al. (2010) est d'avoir une mesure du conflit générique, alors que la mesure du conflit proposée par Régis et al. (2007) ne peut être utilisé que dans le cas où les sources d'informations ont des fonctions de masses réparties sur les mêmes éléments focaux. Par ailleurs, selon la valeur de seuil choisie, on peut remarquer la supériorité de l'approche basée sur la distance métrique. Des tests plus avancés devront donc être réalisés pour évaluer avec précision l'intérêt relatif de ces mesures de pertinence et de fiabilité.

Références

Appriou, A. (2002). *Décision et Reconnaissance des formes en signal*. Hermes Sciences.

Baluja, S. et D. Pomerleau (1997). Dynamic relevance : vision-based focus attention using artificial neural networks. *Artificial Intelligence 97*, 381–395.

Bloch, I. (2005). Fusion d'informations numériques : panorama méthodologique. In *JNRR'05*.

Blum, A. et P. Langley (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence 97*, 245–271.

Bracker, H. (1996). *Utilisation de la théorie de Dempster-Shafer pour la classification d'images satellitaires à l'aide de données multi-sources et multi-temporelles*. Thèse de Doctorat, Ecole Nationale des Télécommunications de Bretagne.

Chebbah, M., A. Martin, et B. Yaghlane (2010). Modélisation dans les bases de données évidentielles. *EGC-AFDC 1O, Hammamet*, 21–32.

Dempster, A. (1968). A generalisation of bayesian inference. *Journal of the Royal Statistical Society 30*, 205–247.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE trans. on systems, man, and cybernetics 25*(5), 804–813.

Dezert, J. et F. Smarandache (2003). On the generation of hyper-powersets for the dezert-smarandache thoery. In *Fusion 2003*, pp. 1118–1125.

Dezert, J. et F. Smarandache (2004a). *Advances and Application in Dezert-Smarandache Theory*, Chapter Combining Uncertain and Paradoxical Evidences in Dezert-Smarandache Thoery.

Dezert, J. et F. Smarandache (2004b). The generalized pignistic transformation. In *Fusion 2004*, Stockholm Suède.

Dubois, D. et H. Prade (1988). Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence 4*, 244–264.

Goma, G., J.-L. Uribelarrea, V. Guillouet, et C. Jouve (2004). Tackling complexity in industrial microbiology for bioprocess. In *4rth International Congress on Bioprocess in Food Industries*, Clermont-Ferrand.

Jousselme, A.-L., D. Grenier, et E. Bossé (2001). A new distance between two bodies of evidence. *Information Fusion 2*, 91–101.

Larousse (2004). le Petit Larousse Illustré. Edt. Larousse.

Lazo-Cortès, M. et J. Ruiz-Schulcloper (1995). Determining the feature relevance for non-classically described objects and a new algorithm to compute typical fuzzy testors. *Pattern Recognition Letters 16*, 1259–1265.

Lefevre, E., O. Colot, et P. Vannoorenberghe (2002). Belief function combination and conflict management. *Information Fusion 3*, 149–162.

Régis, S. (2004). *Segmentation, classification, et fusion d'informations de séries temporelles multi-sources : application à des signaux dans un bioprocédé*. Thèse de Doctorat, Université des Antilles et de la Guyane.

Régis, S., J. Desachy, et A. Doncescu (2004). Evaluation of biochemical sources pertinence in classification of cell's physiological states by evidence theory. In *FUZZ'IEEE*, Budapest, Hongrie.

Régis, S., A. Doncescu, et J. Desachy (2007). Théorie des fonctions de croyance pour la fusion et l'évaluation de la pertinence des sources d'informations : application à un bioprocédé fermentaire. *Traitement du signal 24*(2), 115–132.

Schubert, J. (1993). On non specific evidence. *International Journal of Intelligent Systems 8*, 711–725.

Schubert, J. (1995). Finding a posterior domain probability distribution by specifying non-specific evidence. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 3*, 163–185.

Schubert, J. (1996). Specifying nonspecific evidence. *International Journal of Intelligent systems 11*, 525–563.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. New Jersey : Princeton University Press.

Smets, P. (1988). *Non standard Logics for Automated Reasoning*, Chapter Belief Functions, pp. 29–39. Academic Press.

Smets, P. et R. Kennes (1994). The transferable belief model. *Artificial Intelligence 66*, 191–234.

Waissman-Vilanova, J. (2000). *Construction d'un modèle comportemental pour la supervision de procédés : application à une station de traitement des eaux*. Thèse de Doctorat, LAAS - CNRS, Institut National Polytechnique de Toulouse.

Yager, R. (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences 41*, 93–138.

Zadeh, L. (1984). A mathematical theory of evidence (book review). *AI magazine 5*(3), 81–83.

Zadeh, L. (2004). A note on web intelligence, world knowledge and fuzzy logic. *Data and Knowledge Engineering 50*, 291–304.

Summary

In this article we study two measures of conflict (in the context of the belief functions theory) proposed in two similar approaches that aim to assess the quality of a information source. Empirical tests show that both measures give globally close results.

Stratégie de fusion d'informations exploitant le réseau des sources

Thomas Bärecke*, Marie-Jeanne Lesot*
Herman Akdag*, Bernadette Bouchon-Meunier*

* LIP6 - Université Pierre et Marie Curie-Paris6, UMR7606
4 place Jussieu 75252 Paris cedex 05
prénom.nom@lip6.fr

Résumé. Dans le processus de la cotation d'une information, la phase de fusion vise à agréger les confiances accordées aux diverses déclarations qui s'y rapportent, afin de quantifier la confiance globale en cette information. Nous considérons la prise en compte, dans cette étape, de relations d'affinité et d'hostilité entre les sources. Nous proposons de décomposer le graphe valué représentant le réseau de sources en sous-graphes amicaux liés par des relations d'hostilité. Nous proposons ensuite une stratégie de fusion qui exploite cette partition pour effectuer une agrégation nuancée des confiances individuelles.

1 Introduction

La cotation d'information vise à évaluer la qualité d'une information, et en particulier la confiance qu'on peut lui accorder. La formalisation initiale du standard STANAG2022 (OTAN, 1997) repose sur deux critères : la fiabilité de la source et la plausibilité de l'information, comprise comme sa confirmation par d'autres sources. Elle a été enrichie par de nombreux facteurs (Demolombe, 2004; Cholvy, 2004; Besombes et Revault d'Allonnes, 2008; Bärecke et al., 2010) tels que la sincérité de la source, sa compétence, la vraisemblance de l'information par rapport à des connaissances a priori, sa véracité par rapport à des informations déjà récoltées ou encore l'incertitude exprimée par la source à travers des marqueurs linguistiques comme le conditionnel ou des adverbes. Différents formalismes ont été exploités pour représenter la confiance calculée, tels que la logique multi-valuée (Revault d'Allonnes et al., 2007), la théorie de l'évidence (Cholvy, 2010) ou la théorie des possibilités (Bärecke et al., 2010).

De façon générale, la cotation se déroule en deux étapes : la première évalue la qualité d'éléments informationnels individuels, la seconde, appelée fusion, agrège les confiances individuelles des éléments qui se rapportent à l'information à traiter. Dans cet article, nous considérons la seconde étape qui réalise une agrégation basée sur les confirmations et infirmations de l'information. Plus précisément, nous proposons une définition de cette notion de corroboration qui tient compte d'une connaissance a priori sur les relations entre sources : nous considérons qu'il est réaliste de supposer qu'un analyste est un expert qui possède des informations sur les sources, et est en mesure d'indiquer, au moins partiellement, des relations d'affinité ou d'hostilité entre les sources. Ces connaissances peuvent reposer sur des connaissances générales,

supposant par exemple des affinités entre un chef de l'état et le porte-parole du gouvernement en général, sur des expertises géopolitiques renseignant sur les relations entre pays et donc leurs déclarations officielles ou sur des expertises plus spécifiques.

En effet, ces connaissances peuvent être exploitées pour nuancer l'agrégation des confiances individuelles : le principe général est de considérer que des sources indépendantes, voire hostiles, qui fournissent une même information lui donnent plus de poids qu'un ensemble de sources en relation d'affinité qui produisent naturellement une information plutôt redondante.

Aussi, nous considérons la tâche de fusion des confiances accordées à des éléments informationnels isolés dans le cas où l'on dispose d'un graphe dont les nœuds correspondent aux sources et les arêtes sont étiquetées comme relation d'affinité ou d'hostilité. Nous proposons une stratégie de fusion basée sur une partition cohérente et séparable du graphe, c'est-à-dire sur une décomposition en sous-réseaux de sources amicales, liés entre eux par des relations d'hostilité. La section 2 décrit la méthode de partition proposée, et la section 3 la stratégie de fusion qui comporte deux étapes : une fusion partielle à l'intérieur de chaque sous-groupe amical, puis la fusion des résultats fournis par les différents groupes.

2 Partitionnement du graphe de relations entre sources

La première étape de l'approche proposée consiste à décomposer le réseau indiquant des relations d'affinité et d'hostilité entre paires de sources, afin d'identifier des sous-groupes obéissant à deux contraintes, de cohérence et de séparabilité. La première impose l'absence de liens hostiles à l'intérieur d'un sous-graphe amical : si elles sont liées, les sources d'un même sous-groupe doivent être en relation d'affinité. La séparabilité impose l'absence de liens d'affinité entre des sous-graphes distincts : ceux-ci doivent être indépendants ou hostiles.

Un réseau réel, comme illustré par la figure 1, n'est généralement pas cohérent et séparable. Il s'agit donc de trouver la partition qui minimise les violations des deux contraintes, c'est-à-dire le nombre de modifications du réseau qui le transforment en un graphe cohérent et séparable, en minimisant les suppressions d'arêtes : dans la figure 1, la suppression du seul lien d'affinité entre les sources S_3 et S_5 suffit à obtenir deux sous-groupes vérifiant les contraintes.

Après avoir décrit la fonction de coût formalisant ce principe ainsi que des méthodes d'optimisation, nous discutons des problèmes de non transitivité de la relation d'affinité.

2.1 Formalisation

Fonction de coût On note le graphe $G = (V, E)$, avec V l'ensemble des sources et E l'ensemble des arêtes ; v et w étant deux sommets, on note vHw (resp. vAw) s'ils sont liés par une relation d'hostilité (resp. d'affinité). On note enfin $\mathcal{C} = \{C_1, ..., C_n\}$ une partition des sommets, telle que $\bigcup_{i=1}^{n} C_i = V$ et $\forall i \neq j \; C_i \cap C_j = \emptyset$. La fonction de coût à minimiser sur l'ensemble des partitions $\mathcal{C}$ possibles qui formalise les principes précédents s'écrit alors

$$f(\mathcal{C}) = \alpha \sum_{i=1}^{n} \left| \{(v, w) \in C_i^2 | vHw\} \right| + \beta \sum_{(i,j)|j>i} |\{(v, w) \in C_i \times C_j | vAw\}|$$

Le premier terme exprime la contrainte de cohérence : il impose de minimiser les relations hostiles entre nœuds d'un même sous-groupe. Le second terme représente la contrainte de

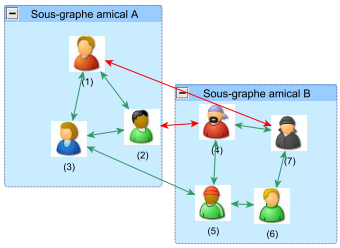


FIG. 1 – *Réseau de sources S_1 à S_7 non cohérent et sa partition optimale (lien d'affinité en vert, d'hostilité en rouge).*

séparabilité : il impose de minimiser les relations amicales entre nœuds affectés à des sous-groupes différents. Les paramètres α et β permettent de pondérer l'importance relative des deux critères et d'indiquer si l'on préfère a priori supprimer des arêtes d'affinité ou d'hostilité. A titre d'exemple, pour le réseau et la partition représentés sur la figure 1, on a $f(\mathcal{C}) = \beta$, car un seul lien d'affinité, entre les sources S_3 et S_5, n'est pas respecté.

L'originalité de cette fonction de coût par rapport au problème classique de recherche de coupe minimale de graphes réside dans la prise en compte des deux relations simultanément, l'affinité A et l'hostilité H, au lieu d'un seul critère.

Détermination de la partition optimale La partition optimale qui minimise la fonction de coût f ci-dessus peut alors être obtenue par l'algorithme A* (Hart et al., 1968) adapté au problème de coupe minimale précédent. Dans le pire cas, cet algorithme a une complexité exponentielle en nombre de nœuds, il reste néanmoins très performant sur de petits graphes.

D'autres approches plus appropriées pour les réseaux de taille élevée incluent les stratégies de type "diviser pour régner" ou les méta-heuristiques (Talbi, 2009) comme les algorithmes évolutionnaires, le recuit simulé ou la recherche taboue, qui permettent de trouver une solution approchée en un temps de calcul très réduit.

2.2 Non transitivité de la relation d'affinité

Le manque de cohérence ou de séparabilité des réseaux provient de la non transitivité de la relation d'affinité : sur la figure 1 par exemple, les sources hostiles S_2 et S_4 sont liées par la chaîne d'affinité S_3-S_5, de même S_1 et S_7 sont liées par le chemin d'affinité S_3-S_5-S_6.

Le cas le plus simple de non transitivité est illustré par le réseau de gauche de la figure 2 : deux sources ayant un ami commun sont en relation d'hostilité. La suppression d'une arête quelconque suffit à restituer un réseau cohérent et séparable. En notant $\mathcal{C}_1 = \{\{S_1, S_2, S_3\}\}, \mathcal{C}_2 = \{\{S_1, S_2\}, \{S_3\}\}$ et $\mathcal{C}_3 = \{\{S_1\}, \{S_2, S_3\}\}$, les partitions optimales sont $\mathcal{C}_1$, $\mathcal{C}_2$ et $\mathcal{C}_3$, ex aequo si $\alpha = \beta$; $\mathcal{C}_1$ si $\alpha < \beta$; $\mathcal{C}_2$ et $\mathcal{C}_3$, si $\alpha > \beta$. Aussi, dans la majorité des cas il n'existe pas de solution optimale unique. Le graphe de droite de la figure 2 illustre la non transitivité avec une chaîne d'affinité longue entre les deux sources hostiles S_1 et S_7. Là aussi, il suffit de supprimer une arête quelconque pour rendre le réseau cohérent : si toutes les sources sont affectées à un unique groupe, un lien d'hostilité est violé, conduisant à un coût égal à α ; toutes les décompositions optimales en deux sous-groupes suppriment un lien

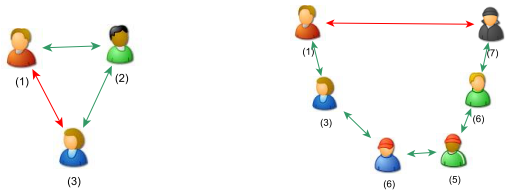


FIG. 2 – *Exemples de réseaux non transitifs.*

d'amitié, et ont un coût de β. Suivant les valeurs de α et β, un grand nombre de solutions est donc indistinguable.

La résolution de tels conflits est possible par exemple si l'on dispose d'une pondération numérique des arêtes, indiquant un degré d'affinité et d'hostilité. Toutefois, cette approche repose sur une hypothèse de connaissances disponibles peu réaliste.

Dans le cas de longue chaîne d'affinité, une heuristique peut consister à préférer couper un lien dans la chaîne plutôt que le lien d'hostilité, en supposant un affaiblissement de l'amitié en raison de la longueur du chemin, et à considérer ensuite une faible hostilité entre les deux sous-groupes ainsi créés. Ainsi, il semble préférable de considérer que S_1 et S_6 ne sont pas dans le même sous-groupe, plutôt que d'imposer que S_1 et S_7, en relation d'hostilité immédiate, soient dans le même groupe. Le choix de la position de coupure reste néanmoins délicat.

Dans le cas particulier où seul un sous-ensemble des sources rapporte des informations, ce choix peut être basé sur la longueur des liens d'amitié : si par exemple seules les sources S_1, S_5 et S_7 se sont exprimées, cette longueur est 2 pour S_5 et S_7 et 3 pour S_5 et S_1. Les sources S_5 et S_7 formeront donc un sous-graphe tandis que la source S_1 reste isolée et en faible relation d'hostilité. Il est également envisageable de prendre en compte les contradictions observées pour choisir la coupure appropriée. Ainsi, dans l'exemple précédent, il est vraisemblable que les informations fournies par la source S_1 contredisent celles de S_5 et S_7.

3 Fusion par partition

A partir de la partition du réseau de sources, on peut réaliser une agrégation nuancée des confiances attribuées à chaque élément d'information : le principe est d'accorder moins d'importance aux informations confirmées par des sources amies qu'à celles confirmées par des sources indépendantes ou hostiles. En effet, on peut s'attendre à une redondance a priori des informations fournies par des sources en relation d'affinité, moins informative que des sources indépendantes ou hostiles. La fusion comporte alors deux étapes détaillées successivement ci-dessous : une fusion intra-groupe, qui agrège les informations fournies par des sources considérées comme globalement amicales, puis une fusion des résultats ainsi obtenus, entre groupes.

Il faut noter que l'opérateur d'agrégation lui-même dépend du formalisme adopté pour représenter les confiances individuelles (logique multi-valuée, théories de l'évidence ou des possibilités par exemple). Nous examinons ci-dessous la sémantique qui doit lui être associée, selon la classification des opérateurs distinguant les opérateurs conjonctifs, disjonctifs, de compromis et à attitude variable (Detyniecki, 2000).

3.1 Fusion intra-groupe

Pour des sources au sein d'un même groupe, l'unanimité est attendue, et ne doit pas renforcer la confiance. Aussi, on est amené à considérer des opérateurs de compromis, qui fournissent une valeur agrégée intermédiaire entre les valeurs individuelles : si toutes les sources sont en accord, sur une confiance élevée ou une confiance faible, le résultat se situe dans la même plage de valeurs. S'il existe des contradictions dans le groupe, c'est-à-dire si une source a une valeur nettement distincte des autres, celle-ci peut être compensée par les autres valeurs.

Une autre possibilité, plus sévère, consiste à rejeter et pénaliser de telles contradictions internes, en exigeant que les sources amies soient cohérentes. Ainsi, un opérateur conjonctif peut également être envisagé : il suffit alors par exemple qu'une source du groupe ne soit pas confiante pour que l'on diminue drastiquement la confiance globale. On peut également pondérer ce comportement, en tenant compte d'un degré global d'affinité au sein du groupe, en utilisant par exemple des mesures de connectivité dans le sous-graphe.

Dans le cas où la confiance calculée est représentée dans le cadre de la théorie de l'évidence, des règles de combinaison telles que la règle prudente ou ses variantes, proposées par Denoeux (2006), peuvent être envisagées.

3.2 Fusion inter-groupes

L'étape suivante fusionne les résultats fournis par les différents groupes, considérés comme globalement indépendants ou en relation d'hostilité. Aussi, une redondance peut ici être interprétée comme un renforcement : si des groupes de sources indépendants voire hostiles accordent une confiance élevée à une même information, cette unanimité peut être jugée comme significative. Les groupes se renforcent alors mutuellement, et conduisent à une confiance agrégée plus élevée encore. De même, s'il y a unanimité sur une confiance faible, on peut conclure à une valeur plus faible encore. Enfin, dans les cas où un désaccord est observé, on peut adopter simplement un comportement de compromis. L'opérateur d'agrégation doit donc être conjonctif pour les valeurs faibles, disjonctif pour les valeurs élevées et de compromis dans les cas intermédiaires : de tels comportements, dits à attitude variable et à renforcement total (Detyniecki, 2000), sont par exemple présentés par la somme symétrique (Silvert, 1979).

Par ailleurs, il semble approprié de tenir compte de la cardinalité des groupes considérés : un groupe plus grand doit avoir plus d'importance qu'un groupe réduit, sans toutefois que sa contribution au résultat final domine celle du second groupe. Aussi, nous proposons d'associer à chaque groupe un poids, défini en fonction de sa cardinalité. Afin d'atténuer l'influence des groupes les plus importants sans ignorer les groupes plus petits, nous proposons de définir ces poids comme la racine de leurs cardinalités.

3.3 Exemple illustratif

Pour illustrer la méthode proposée, nous considérons le réseau de sources représenté sur la figure 1, et quatre assertions exprimées par les sources S_1, S_3, S_4 et S_7. Nous considérons de plus que la cotation se fait dans le formalisme de la théorie des possibilités (Dubois et Prade, 1988), tel que décrit par Bärecke et al. (2010) : l'évaluation d'une assertion relative à un événement e lui associe un couple de deux valeurs, donnant la possibilité exprimée par la

source pour e et la possibilité de son contraire $\neg e$. La certitude d'une source est alors calculée comme le complément à 1 de la valeur minimale de ce couple.

Considérons à titre d'exemple que l'évaluation des assertions individuelles ait fourni les résultats suivants :

$$\begin{array}{llll} S_1 & (1 \quad 0.4) & S_4 & (1 \quad 0.7) \\ S_3 & (1 \quad 0.2) & S_7 & (1 \quad 0.2) \end{array}$$

Ainsi, les quatre sources sont en accord globalement, et accordent une possibilité maximale à e. Elles varient par la certitude qu'elles lui accordent : les sources S_1, S_3 et S_7 sont plutôt sûres d'elles (leurs degrés de certitude respectifs sont 0.6, 0.8 et 0.8), la source S_3 accorde une possibilité élevée à $\neg e$ également et son degré de certitude est 0.3 seulement.

Pour la fusion de ces valeurs, étant donné que la partition du réseau est déjà disponible, deux étapes sont à effectuer : pour la fusion intra-groupe, nous choisissons d'utiliser un opérateur de compromis, la moyenne. Deux groupes sont à considérer : d'une part les sources S_1 et S_3 qui conduisent à la distribution $(1 \quad 0.3)$, d'autre part les sources S_4 et S_7, qui conduisent à $(1 \quad 0.45)$.

Ainsi, des groupes en relation d'hostilité sont en accord sur leur évaluation de l'événement, un comportement de renforcement est donc attendu. Pour l'étape de fusion inter-groupes, nous utilisons la somme symétrique basée sur le produit, définie par $Ag(x, y) = xy/(xy + (1 - x)(1 - y))$ (Silvert, 1979). On obtient alors la distribution finale $(1 \quad 0.26)$.

Ainsi la certitude dans l'événement e augmente par rapport aux certitudes des sources considérées individuellement, on a bien un renforcement.

4 Conclusion et perspectives

Nous avons proposé une stratégie de fusion prenant en compte des connaissances sur les relations entre les sources émettant les éléments d'information, suivant qu'elles sont en relation d'affinité ou d'hostilité. La stratégie consiste à décomposer le réseau des sources en groupes amicaux en relation d'hostilité, pour atténuer la redondance des sources amies et renforcer les confirmations entre sources indépendantes ou hostiles.

Outre des expérimentations permettant de tester l'approche proposée et de la valider sur un problème réel, les travaux en cours visent à enrichir encore la fusion, en tenant compte de la temporalité des éléments d'information, pour les pondérer en fonction de leur ancienneté. En effet, dans de nombreux domaines, les informations se succèdent très rapidement et l'obsolescence est élevée. D'autre part, les dynamiques de confirmation et d'infirmation influencent les schémas de confiance : une alternance d'éléments informationnels contradictoires n'a pas le même impact qu'une période de confirmation suivie d'une infirmation. La formalisation de ces profils dynamiques constitue également une perspective à la fusion enrichie pour la cotation d'informations.

Remerciements

Ces travaux ont été réalisés dans le cadre du projet CAHORS, financé par l'ANR - CSOSG'08.

Références

Bärecke, T., T. Delavallade, M.-J. Lesot, F. Pichon, H. Akdag, B. Bouchon-Meunier, P. Capet, et L. Cholvy (2010). Un modèle de cotation pour la veille informationnelle en source ouverte. In *Actes du 6ème colloque Veille Stratégique Scientifique & Technologique*.

Besombes, J. et A. Revault d'Allonnes (2008). An extension of STANAG2022 for information scoring. In *Proc. of the Int. Conf. on Information Fusion*, pp. 1635–1641.

Cholvy, L. (2004). Information evaluation in fusion : a case study. In *Proc. of IPMU'04*, pp. 993–1000.

Cholvy, L. (2010). Evaluation of information reported : a model in the theory of evidence. In *Proc. of IPMU'10*, pp. 258–267.

Demolombe, R. (2004). Reasoning about trust : a formal logical framework. In *Int. Conf. on iTrust*.

Denoeux, T. (2006). The cautious rule of combination for belief functions and some extensions. In *9th Int. Conf. on Information Fusion*, pp. 1–8.

Detyniecki, M. (2000). *Mathematical Aggregation Operators and their Application to Video Querying*. Ph. D. thesis, University of Pierre and Marie Curie.

Dubois, D. et H. Prade (1988). *Possibility Theory*. Springer.

Hart, P., N. Nilsson, et B. Raphael (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. on Systems Science and Cybernetics 4*(2), 100–107.

OTAN (1997). Annex to STANAG2022. Information handling system.

Revault d'Allonnes, A., H. Akdag, et O. Poirel (2007). Trust-moderated information-likelihood. a multi-valued logics approach. In *Proc. of the 3rd Conf. on Computability in Europe, CiE 2007*, pp. 1–6.

Silvert, W. (1979). Symmetric summation : a class of operations on fuzzy sets. *IEEE Trans. on Systems, Man and Cybernetics 9*, 659–667.

Talbi, E.-G. (2009). *Metaheuristics : from design to implementation*. Wiley.

Summary

The fusion step of the cotation process aims at aggregating the confidence values associated to individual assertions related to the same piece of information, in order to assess the global confidence in this information. In this paper, we propose a fusion procedure that exploits relationships between the sources, more precisely whether they are friendly or hostile. From the representation of the source network as an edge-valued graph, we propose a decomposition method to identify affinity subgraphs linked by hostility relationships, and a fusion method that exploits this partition to perform the aggregation of the individual confidences.

Représentativité et graphe de représentants : une approche inspirée de la théorie du choix social pour la fouille de données relationnelles

Frédéric Blanchard*, Cyril de Runz*
Herman Akdag**, Michel Herbin*

*CReSTIC,
IUT de Reims, Rue des crayères, BP 1025,
51687 REIMS cedex 2
frederic.blanchard@univ-reims.fr,
http://crestic.univ-reims.fr/
**LIP6
4 place Jussieu
75252 Paris cedex 05

Résumé. Après avoir défini la *représentativité* dans un ensemble de données relationnelles, nous utilisons cette notion pour construire un *graphe des représentants*. Ce graphe permet de faire émerger des structures arborescentes et des regroupements possibles des données en partitions. Le calcul de la représentativité et la construction du graphe ne requièrent qu'une relation entre les paires d'objets. Notre concept est particulièrement adaptée aux données complexes pour lesquelles on ne dispose pas d'hypothèse a priori. Nous proposons une application directe de notre outil en classification automatique de données complexes.

1 Introduction

Dans la démocratie athénienne, le peuple est réuni sur l'agora et la participation au débat est directe. Le développement des réseaux sociaux et des applications sur internet conduit à réunir sur une agora virtuelle un grand nombre de participants à une activité. Un débat nécessite alors le regroupement à la volée des opinions exprimées, les plus semblables ou similaires sont diffusées par l'intermédiaire d'un représentant. Ces représentants sont fugaces, ils changent et sont désignés à chaque mouvement d'opinion. Ils participent directement au débat mais ne représentent qu'un regroupement d'opinion à un instant donné.

Le travail présenté dans ce papier est une contribution à la détermination de ces représentants dans un flot de données complexes ou d'opinions exprimées. Pour exposer ce travail, nous avons repris le schéma classique de la classification des données.

Nous n'avons pas trouvé, dans la littérature, de définition formelle de la représentativité. Si les statisticiens emploient ce terme lorsqu'ils évoquent la représentativité d'un échantillon, le

sens est toutefois très éloigné de celui que nous évoquions plus haut. En théorie des catégories, la notion de *typicalité* se rapproche de celle que nous proposons. Le degré de typicalité (voir Lesot (2006) ainsi que Lesot et al. (2007) et Rifqi (1996)) est défini sur le principe suivant : « un objet est d'autant plus typique qu'il ressemble beaucoup aux membres de sa classe et qu'il est très différent des membres des autres classes ». La différence majeure avec la notion que nous mettons en place est que dans cette approche, le partitionnement en classe doit précéder le calcul du degré de typicalité. Le degré de typicalité est donc complètement lié à la classification. Dans notre concept, aucune hypothèse n'est faite sur l'ensemble des données, ni sur l'éventuel espace contenant ces données. Le degré de représentativité dont nous proposons la définition n'est calculé qu'à partir des relations entre les données. Elle est complètement indépendante du problème de classification. Elle ne requiert, pour être mise en place, qu'une relation entre les objets pris deux à deux. Notre définition est donc particulièrement adaptée à l'analyse des données complexes, pour lesquelles on ne dispose généralement pas de métrique simple.

Notre définition repose dans son interprétation sur des idées empruntées à la théorie du choix social ce qui permet de la rendre plus « expressive » (nous avions dans un précédent travail (voir Blanchard et al. (2010)), proposé une définition de la représentativité, pour des données numériques, reposant la théorie des ensembles flous, mais dont l'interprétabilité n'était pas aisée). Formellement, elle utilise des outils mathématiques simples et robustes. On peut décomposer son calcul en trois étapes :

1. Expression des préférences individuelles des objets : les objets à étudier sont présentés à travers les relations valuées entre les objets, deux à deux.
2. Transformation des préférences individuelles : les relations sont transformées en rangs, et les rangs en scores de rangs (scores de rangs de Borda, voir Chamberlin et Courant (1983)).
3. Calcul du degré de représentativité par agrégation des préférences individuelles : les scores de rangs obtenus par chaque objet sont agrégés.

La deuxième partie de notre contribution consiste à exploiter ce degré de représentativité pour faire émerger une structuration des objets étudiés. Nous définissons sur ces objets un *graphe des représentants* de la manière suivante :

1. Les degrés de représentativité de tous les objets sont calculés.
2. On détermine les voisinages des objets au sens des k-plus proches voisins (en utilisant la relation pour définir les proximités).
3. On associe à chaque objet, celui, dans son voisinage, dont le degré de représentativité est le plus élevé. Ce deuxième objet est en quelque sorte le « représentant local » du premier.

On obtient ainsi un graphe dont chaque composante connexe est un arbre et possède un unique meilleur représentant (le sommet qui n'a pas de successeur).

Nous proposons enfin, comme conséquence immédiate, une application à la classification automatique de données. En effet, le partitionnement de l'ensemble des objets induit par celui en composantes connexes du graphe des représentants, constitue une classification de l'ensemble initial.

Dans la suite de ce document, nous décrivons toutes les étapes qui permettent de construire le degré de représentativité, puis le graphe des représentants. Nous proposons ensuite une application à la classification automatique. Enfin, nous terminons par les traditionnelles conclusions et perspectives.

2 Degré de représentativité

La première étape de notre méthode consiste à définir la *représentativité* de chaque donnée dans son ensemble initial, et de calculer son *degré de représentativité*. D'un point de vue mathématique, ce calcul repose sur la transformation en rangs des dissimilarités entre les objets (données) considérés, puis sur une agrégation de ces rangs pour produire un indice quantifiant « combien » chaque objet est représentatif de son ensemble.

2.1 Les données relationnelles

Comme nous l'avons déjà expliqué, nous nous intéressons dans ce travail à l'analyse des données relationnelles. On suppose ainsi que l'on dispose d'un ensemble $\mathcal{O} = \{O_1, O_2, \ldots, O_n\}$ de n objets à étudier. Contrairement au cas classique où ces objets sont décrits par des vecteurs numériques (caractérisant les « mesures » observées des différentes variables sur ces objets), les données relationnelles sont constituées de valeurs numériques quantifiant la *relation* entre les objets pris deux à deux. On représente ainsi ces données sous la forme d'une matrice $n \times n$:

$$R_{\mathcal{O}} = \begin{pmatrix} R_{1,1} & \cdots & R_{1,n} \\ \vdots & \ddots & \vdots \\ R_{n,1} & \cdots & R_{n,n} \end{pmatrix}$$

où $R_{i,j}$ quantifie la relation entre l'objet O_i et l'objet O_j.

Nous supposerons, dans la suite de ce document, que la relation quantifie un éloignement, une différence, entre les objets. Les mesures de *dissimilarité* et de *distance* sont les exemples les plus typiques de ce genre de relation.

2.2 Scores de rangs

La première étape consiste à transformer cette matrice de dissimilarité en matrice de rangs. La transformation en rang est une technique de prétraitement des données qui confère au processus qui l'utilise une robustesse et qui permet d'échapper aux hypothèses sur la distribution des données (voir Friedman (1937)) et de limiter l'impact d'éventuelles données aberrantes. Pratiquement, cette opération consiste à transformer chaque colonne R^i de la matrice $R_{\mathcal{O}}$ en remplaçant chaque valeur de cette colonne par son rang dans l'ensemble trié des valeurs. On obtient ainsi une matrice $n \times n$:

$$RG_{\mathcal{O}} = \begin{pmatrix} RG_{1,1} & \cdots & RG_{1,n} \\ \vdots & \ddots & \vdots \\ RG_{n,1} & \cdots & RG_{n,n} \end{pmatrix}$$

où :

$$RG_{i,j} = 1 + \sum_{k=1}^{n} \mathbf{1}_{]-\infty;R_{i,j}[}(R_{k,j})$$

En empruntant le langage de la théorie du choix social, l'opération précédente peut sembler plus intuitive. Considérons un objet quelconque O_j de $\mathcal{O}$. On peut voir la colonne RG^j comme étant le classement effectué par O_j sur l'ensemble des objets de $\mathcal{O}$, en fonction de ses préférences, ou de sa ressemblance avec les autres objets. Ainsi $RG_{i,j} = k$ signifie que O_i est le $k^{\text{ième}}$ objet préféré par O_j. Naturellement, O_j est l'objet préféré de O_j.

À l'issue de cette étape, chaque objet classe tous les autres en fonction de ses préférences, et par conséquent, chaque objet se voit classé par les n objets de $\mathcal{O}$. Ainsi, chaque objet peut-être caractérisé par l'ensemble des rangs qu'il obtient dans les n classements. Pour un objet O_i ces rangs sont contenus dans la $i^{\text{ème}}$ ligne de $RG_{\mathcal{O}}$.

L'étape suivante consiste ensuite à agréger les scores correspondant aux rangs obtenus afin d'obtenir un *score global* pour chaque objet, que nous appellerons *degré de représentativité*.

2.3 Degré de représentativité

Pour agréger les « préférences » des objets nous avons choisi d'utiliser la méthode de Borda. Cette méthode consiste à transformer chaque rang en un score puis à les agréger en les sommant. Un objet classé 1^{er} par un autre reçoit n points. Il reçoit $n-1$ points lorsqu'il est classé $2^{\text{ème}}$, $n-k+1$ lorsqu'il est classé $k^{\text{ième}}$, et 1 seul point lorsqu'il est dernier. Chaque objet O_i reçoit ainsi n scores de rangs. En sommant ces scores on obtient un score global qui, divisé par n, définit ce que nous appelons l'indice de représentativité de l'objet O_i dans $\mathcal{O}$ (remarque : diviser par n n'est pas toujours utile ; on préfèrera, par exemple pour mieux discriminer les objets, conserver la somme sans effectuer cette normalisation).

On a donc :

$$DR(O_i) = \frac{1}{n} \sum_{j=1}^{n} (n - RG_{i,j} + 1)$$

et donc :

$$DR(O_i) = n - \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbf{1}_{]-\infty;R_{i,j}[}(R_{k,j})$$

Cette notion, comme son nom l'indique, quantifie combien un objet représentatif de son ensemble. En observant la façon dont elle a été construite, on peut facilement vérifier que *plus un objet est « préféré » par les autres, plus son indice de représentativité est élevé (et inversement)*.

2.4 Exemple

Pour illustrer le processus de calcul de l'indice de représentativité, nous allons l'utiliser sur un exemple simple.

On considère les points de $\mathbb{R}^2$ suivants :

A = (1.00 , 2.00)
B = (1.00 , 3.00)
C = (2.00 , 2.00)
D = (3.00 , 4.00)
E = (1.00 , 0.00)
F = (3.00 , 1.00)

dont les relations deux à deux sont induites par la distance euclidienne et représentées dans la matrice de dissimilarité suivante :

R	A	B	C	D	E	F
A	0.00	1.00	1.00	2.83	2.00	2.24
B	1.00	0.00	1.41	2.24	3.00	2.83
C	1.00	1.41	0.00	2.24	2.24	1.41
D	2.83	2.24	2.24	0.00	4.47	3.00
E	2.00	3.00	2.24	4.47	0.00	2.24
F	2.24	2.83	1.41	3.00	2.24	0.00

La transformation des relations en rangs, puis en scores de rangs donnent :

RG	A	B	C	D	E	F
A	1	2	2	4	2	3
B	2	1	3	2	5	5
C	2	3	1	2	3	2
D	6	4	5	1	6	6
E	4	6	5	6	1	3
F	5	5	3	5	3	1

Sco.	A	B	C	D	E	F
A	6	5	5	3	5	4
B	5	6	4	5	2	2
C	5	4	6	5	4	5
D	1	3	2	6	1	1
E	3	1	2	1	6	4
F	2	2	4	2	4	6

Enfin, les scores sont agrégés (en lignes) pour former les degrés de représentativité (sans normalisation) :

	DR
A	28
B	24
C	29
D	14
E	17
F	20

En représentant graphiquement les points dans le plan et en affectant un niveau de gris d'autant plus sombre que le point est représentatif, on obtient la figure 1.

3 Graphe de représentativité

Dans cette partie, nous allons exploiter le degré de représentativité pour construire, sur l'ensemble des objets étudiés, un graphe que nous appelons *graphe des représentants*.

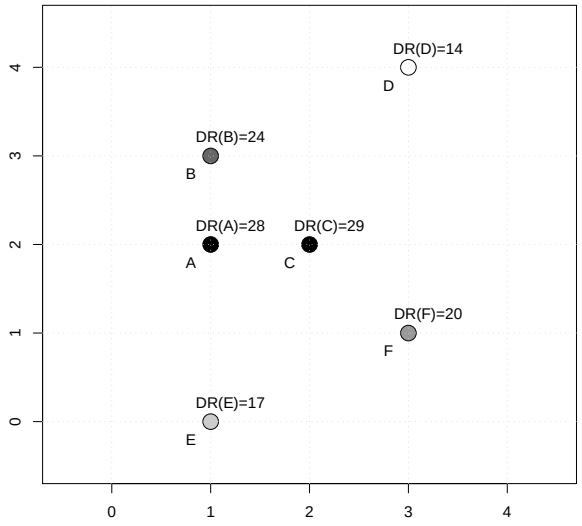


FIG. 1 – *Calcul de la représentativité sur un ensemble de six points de* $\mathbb{R}^2$.

3.1 Construction du graphe

Le principe est d'associer, à chaque objet, son voisin le plus représentatif. Le voisinage d'un objet est déterminé par les k objets qu'il préfère. Autrement dit, le voisinage considéré est un voisinage au sens des k plus proches voisins (k-ppv), dans lequel la notion de proximité est définie grâce à la relation initiale entre les objets.

On obtient ainsi un graphe orienté $G_k = (X, U)$ (avec $k \in [\![1, n]\!]$) dont :
– l'ensemble des sommets X est l'ensemble des objets étudiés ($\mathcal{O}$) ;
– l'ensemble des arcs U est défini de la manière suivante :

$$\forall (i,j) \in [\![1,n]\!]^2, \qquad (O_i, O_j) \in U \Leftrightarrow O_j = \underset{O_p \in V_k(O_i)}{\operatorname{argmax}} \ (DR(O_p))$$

où $V_k(O_i)$ est l'ensemble des k-ppv de O_i : $V_k(O_i) = \{O_p \in \mathcal{O} / RG_{p,i} \leq k\}$
Cette définition permet à un objet d'être son propre représentant, lorsqu'il est le plus représentatif parmi son propre entourage. D'autre part, le graphe obtenu est une forêt. En effet, il n'est pas nécessairement connexe, mais ne peut contenir ni cycle ni circuit.

La construction de ce graphe entraine une autre propriété intéressante : chaque composante connexe du graphe contient un et un seul *puits* (un puits est un sommet ne possédant aucun successeur). Ces sommets jouent un rôle particulier dans le graphe. Il ont en effet la particularité d'être leur propre meilleur représentant et d'être, par transitivité, les meilleurs représentants de

leurs composantes connexes respectives.

3.2 Composantes connexes

Le nombre de composantes connexes est directement lié au paramètre k utilisé pour déterminer les voisinages des objets. Plus k est élevé et plus le nombre de composantes connexes diminue. Lorsque $k = n$ (avec $n = Card(\mathcal{O})$), le graphe et connexe. Lorsque $k = 1$, chaque objet est son propre meilleur représentant, et le graphe possède n composantes connexes.

Cette relation est illustrée dans l'exemple de la partie suivante (figure 4).

3.3 Exemple

Nous appliquons maintenant notre méthode à un exemple simple pour en illustrer les principaux points.

Considérons un ensemble $\mathcal{O} = \{O_1, O_2, \ldots, O_{20}\}$ constitué de 20 points de $\mathbb{R}^2$. Cet ensemble est constitué de deux classes $\mathcal{C}_1 = \{O_1, O_2, \ldots, O_{10}\}$ et $\mathcal{C}_2 = \{O_{11}, O_{12}, \ldots, O_{20}\}$. La figure 2 représente l'ensemble des points dans le plan. Chaque objet est représenté par un disque dont le niveau de gris est d'autant plus foncé que le degré de représentativité est élevé.

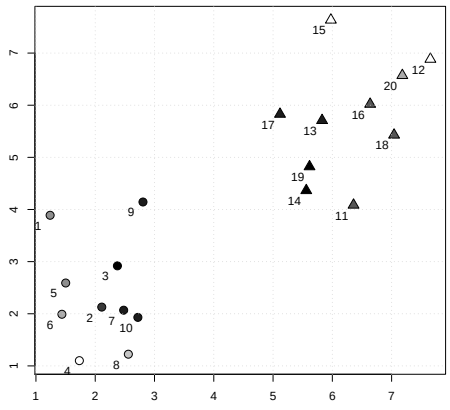


FIG. 2 – *Points en 2D : les données et leur degré de représentativité (le niveau de gris est d'autant plus sombre que le degré de représentativité est élevé)*

Enfin, la figure 3 illustre le graphe des représentants sur l'ensemble de points.

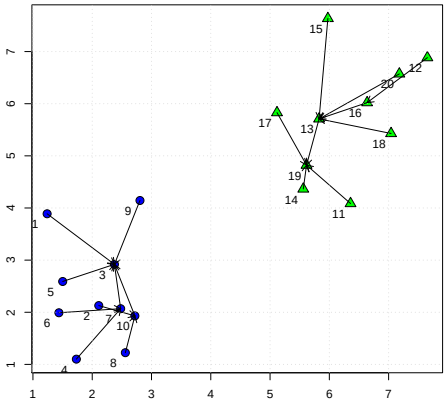


FIG. 3 – *Points en 2D : les données et le graphe des représentants*

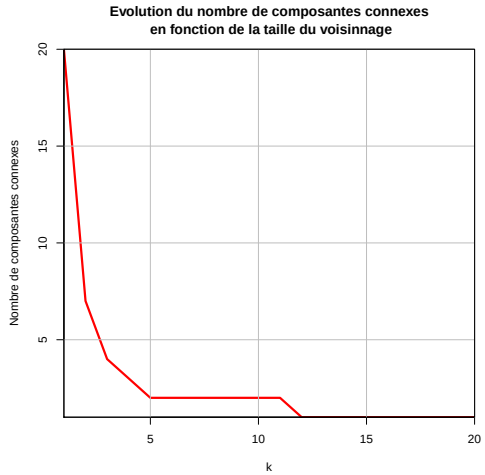


FIG. 4 – *Nombre de composantes connexes Vs. k*

4 Application à la classification automatique

4.1 Partitionnement

Le calcul de la représentativité et du graphe des représentants associé présente deux intérêts immédiats en classification automatique. Tout d'abord la partition du graphe de représentants permet de façon évidente de partitionner l'ensemble des objets étudiés en classes. En faisant varier le paramètre k (la « taille » du voisinage), on obtient un partitionnement plus ou moins fin. Une conséquence triviale de ce qui a été dit avant est que *plus k est petit et plus le partitionnement est fin (i.e. plus le nombre de classe est élevé)*.

La classification effectuée ainsi possède un certain nombre d'avantages hérités des méthodes impliquées dans le calcul de la représentativité.

Le premier avantage est lié à la nature des données traitées. Les données relationnelles sont un cas plus général que celui où l'on dispose de descriptions vectorielles des objets à étudier, et qui est généralement le pré-requis des algorithmes classiques.

Ensuite, la transformation préalable des dissimilarités en rangs est un outil souvent utilisé par les statisticiens pour obtenir des méthodes non paramétriques. Dans notre contexte, elle permet d'effectuer une classification sans faire d'hypothèse sur la distribution des données. Par ailleurs aucune hypothèse n'est -même implicitement- faite sur la forme des classes. La transformation par rangs offre aussi une robustesse vise à vis des valeurs aberrantes.

Enfin, la simplicité des outils théoriques impliqués ainsi que les champs sémantiques des domaines auxquels ils sont empruntés, offrent des possibilités de compréhension, d'interprétation et d'explication que ne permettent pas la plupart des algorithmes.

Remarque : Le choix de la meilleure valeur du paramètre k pour la classification n'est pas aisée. Empiriquement, nous avons déterminé que le choix optimal du paramètre se situait après le « coude », lorsque l'on observe la courbe du nombre de composantes connexes en fonction de k. Cette plage correspond aux valeurs de k pour lesquelles on atteint le premier plateau de stabilisation du nombre de composantes connexes.

4.2 « Meilleurs » représentants

Contrairement aux méthodes de classification comme celles de la famille des k-means, notre méthode de classification ne calcule pas des centres de classes « virtuels » (obtenus par moyenne par exemple), mais extraits, parmi les objets initiaux, des représentants « réels ». Là encore, les calculs sont effectués directement sur les données, à partir de leurs relations deux à deux, et pas dans un hypothétique espace sous-jacent. Cet aspect présente un intérêt non négligeable, notamment dans les situations où l'on souhaite analyser plus précisément les classes à travers leurs représentants.

4.3 Exemple sur des données synthétiques

Nous illustrons maintenant ce processus sur deux jeux de données synthétiques simulées. Les résultats sont illustrés sur la figure 5.

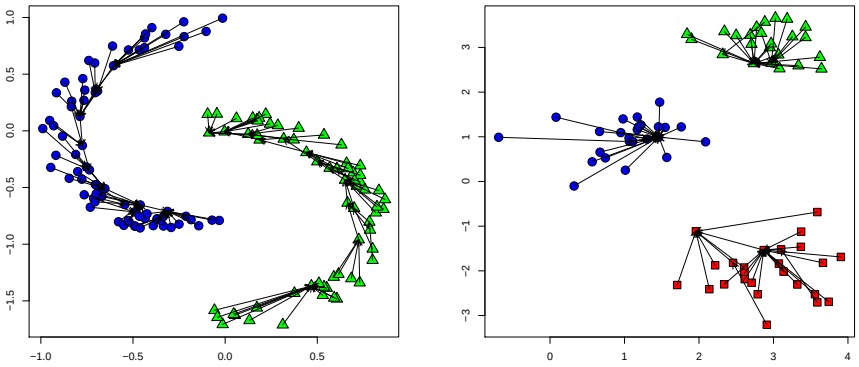


FIG. 5 – *Application à la classification automatique*

5 Conclusion et perspectives

Nous avons proposé, dans ce papier, une définition de la représentativité sur un ensemble de données relationnelles. Cette notion permet de quantifier « combien » un objet est représentatif de son ensemble. Le calcul du degré de représentativité s'effectue à partir des relations valuées entre les objets, deux à deux. Cet indice est robuste et permet facilement d'être interprété grace à l'utilisation de la sémantique de la théorie du choix social à laquelle nous empruntons des outils pour effectuer le calcul. Par ailleurs, son calcul ne nécessite aucun a priori sur la distribution des données ou sur l'espace qui les contient.
Nous avons utilisé le degré de représentativité pour construire un graphe des représentants. Ce graphe permet de faire émerger une structuration des objets, en associant à chacun d'entre eux, « celui qui le représente le mieux » parmi « ceux qu'il préfère ».
L'utilisation de ces deux concepts en classification est immédiat et assez intuitif. Nous avions par ailleurs déjà utilisé le concept de représentativité, dans une version préliminaire, pour l'extraction d'éléments représentatifs en archéologie, (de Runz et al., 2008).

L'utilisation de ce travail pour l'analyse des réseaux sociaux nous semble particulièrement opportune et nous envisageons donc d'orienter les applications dans ce domaine. Sur le plan théorique, nous pensons pouvoir exploiter l'aspect hiérarchique du graphe des représentants et l'utiliser pour développer une méthode hybride de classification que nous comparerons aux approches classiques.

Références

Blanchard, F., P. Vautrot, H. Akdag, et M. Herbin (2010). Data representativeness based on fuzzy set theory. *Journal of Uncertain Systems 4*(3), 216–228.

Chamberlin, J. R. et P. N. Courant (1983). Representative deliberations and representative decisions : Proportional representation and the borda rule. *The American Political Science Review 77*(3), pp. 718–733.

de Runz, C., F. Blanchard, E. Desjardin, et M. Herbin (2008). Fouilles archéologiques : à la recherche d'éléments représentatifs. In *Atelier Fouilles de Données Complexes - Conférence Extraction et Gestion des Connaissances - AFDC@EGC*, Sophia Antipolis, France, pp. 95–103.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Statist. Ass. 32*, 675–701.

Hataway, R. J. et J. C. Bezdek (2003). Visual cluster validity for prototype generator clustering models. *Pattern Recognition Letters 24*, 1563–1569.

Hataway, R. J., J. C. Bezdek, et J. W. Davenport (1996). On relational data versions of c-means algorithms. *Pattern Recognition Letters 17*, 607–612.

Lesot, M.-J. (2006). Typicality-based clustering. *Int. Journal of Information Technology and Intelligent Computing 1*(2), 279–292.

Lesot, M.-J., M. Rifqi, et B. Bouchon-Meunier (2007). *Fuzzy prototypes: From a cognitive view to a machine learning principle*, Chapter Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, pp. 431–452. Springer.

Rifqi, M. (1996). Constructing prototypes from large databases. In *Proc. IPMU'96*, pp. 301–306.

Summary

As complex objects are often modeled as relational data, we propose a (complex) data mining approach using relations between objects. After defining data representativeness in a relational dataset, we use it to build a graph of representatives (delegates). This graph makes possible the emergence of tree structures and of data groups forming partitions. Finally, our approach is illustrated with a clustering application.

Représentation, classification par voisignages successifs de descriptions morphologiques complexes

Henri Ralambondrainy*,David Grosser*,Noël Conruyt*

*LIM-IREMIA, Université de la Réunion
Parc Technologique Universitaire, Bâtiment 2
2, rue Joseph Wetzell - 97490 Sainte-Clotilde
{ralambon,grosser, conruyt}@univ-reunion.fr

Résumé. Afin de classifier des descriptions morphologiques issues de bases de connaissances en biologie, nous proposons une méthode de fouille de données incrémentale, interactive et semi-dirigée. Cette méthode est fondée sur la construction itérative du voisinage de la description partielle de l'objet à classer. Nous proposons différents indices de similarité adaptés à la nature complexe des données considérées (multi-valuées, incomplètes et structurées), pour sélectionner les descriptions les plus proches. Les connaissances du domaine sont utilisées aux différentes étapes du processus de classification, notamment pour le choix de variables discriminantes. A partir de la base de connaissances sur les coraux des Mascareignes, une application montre l'intérêt de cette approche.

1 Introduction

Outre l'étude des liens de parenté entre espèces (phylogénie), la description de *spécimens* biologiques à des fins d'identification et de classification est une part essentielle du travail des systématiciens. L'automatisation de ce processus par des outils informatiques dans le but de construire des systèmes classificatoires pose d'intéressants problèmes de représentation et de traitement de la connaissance. Ceci est particulièrement vrai en biologie marine et pour certains taxons comme les coraux, où la nature polymorphe des individus (les colonies coralliennes) rend difficile leur description par des représentations classiques de type attribut-valeur. Les relations de dépendance entre caractères, induites par une forte variabilité morphologique engendrent notamment des difficultés de description. Pour pallier à ces contraintes et permettre la construction de bases de connaissances sur la biodiversité des milieux marins, une méthodologie d'acquisition des connaissances descriptives adaptée aux besoins particuliers des biologistes, s'appuyant sur *les logiques descriptives en sciences de la vie* a été proposée par Conruyt (1994). Cette méthode permet de définir une connaissance d'ordre ontologique sur la nature des taxons et de décrire de manière structurée les spécimens utilisés pour la conception de la Taxonomie des milieux considérés. Elle offre un cadre signifiant pour représenter des objets biologiques complexes, permettant notamment d'exprimer des relations de dépendance entre composants d'une description. Ces objets complexes sont caractérisés par des attributs de type nominal, ordinal ou continu, ensemble ou intervalle, dont les valeurs peuvent être manquantes,

inconnues, variées ou imprécises selon les choix de définition des experts et les difficultés d'observation rencontrées. Les *logiques descriptives en sciences de la vie* sont à l'origine du modèle de représentation des connaissances par objets **Codesc** proposé par Grosser et al. (2003) et implanté au sein du système IKBS (*Iterative Knowledge Base System*) développé en Java par Grosser (2002). Ce système de gestion de bases de connaissances (SGBC) a notamment été utilisé pour la conception de la *base de connaissances sur les coraux des Mascareignes*. Celle-ci rassemble plus de 150 taxons et 800 descriptions de spécimens, soit une représentation d'environ 80% des coraux de cet archipel du Sud-Ouest de l'Océan Indien.

Pour identifier un spécimen et lui associer un nom, les biologistes utilisent traditionnellement des clefs d'identification (ou de détermination [1]). Cette méthode historique offre l'avantage d'être véhiculée sur support écrit et d'être très explicite, la diagnose pouvant être associée à l'identification. Avec l'informatisation des données systématiques et biologiques et l'essor de la taxonomie numérique initiée par Sokal et Sneath (1963), de nombreuses méthodes d'identification et de classification ont été proposées sur des données tabulaires. Parallèlement, en analyse discriminante et en apprentissage, d'autres méthodes de classification par arbre ont également été développées, telles que les arbres de classification ou de décision (Quinlan (1993)). Ces différentes méthodes peuvent être utilisées dans le cadre de la classification de données biologiques complexes, mais ne sont pas toujours satisfaisantes car elles ne prennent en compte ni les relations entre attributs, ni les données manquantes ou imprécises. Plus important encore, ces méthodes n'intègrent pas de connaissances *a priori* sur les domaines considérés, pré-requis indispensable pour construire des systèmes classificatoires, fiables et robustes (Conruyt (1994)). Or, les experts émettent des hypothèses sur l'objet étudié et procèdent généralement en deux phases pour déterminer la classe d'un spécimen. D'abord une phase *synthétique*, par observation globale des caractères les plus visibles permet de réduire le champ d'investigation en sélectionnant des descriptions présentant des similitudes. Puis une phase *analytique*, par l'observation fine de caractères discriminants permet d'affiner la recherche jusqu'à obtention du résultat.

La méthode de classification par voisinages successifs (CVS) développée dans cet article propose une méthode de fouille fondée sur ce type de raisonnement. Elle s'appuie sur la recherche du voisinage d'une description partiellement renseignée, possédant éventuellement des erreurs. Cette méthode est rendue possible grâce aux différents indices de similarité qui tiennent compte de la structure et du contenu des descriptions (partie 3). Elle utilise une phase de sélection de variables discriminantes afin de compléter l'information et d'affiner progressivement le processus d'identification. La méthode est interactive et itérative. La partie 4 propose une implémentation de cette méthode au sein du système IKBS et une étude comparative de l'approche proposée avec les méthodes de discrimination : arbre de décision et k-plus proches voisins classiques (partie 5) à partir d'une base de connaissance relative aux coraux.

2 Représentation des données morphologiques

Le modèle de représentation des données morphologiques propose deux entités de premier ordre pour la conception d'une base de connaissances : le **modèle descriptif** et les **descriptions**. Un modèle descriptif permet de structurer les caractères observables d'un groupe

1. succession d'alternatives dichotomiques portant sur les caractères d'un spécimen qui permet d'en déterminer le rang taxinomique.

taxinomique donné. Il est composé d'objets (ou composants) ainsi que d'attributs munis d'un domaine de valeurs et de relations. Les relations sont par défaut des relations de composition (*part-of*) entre composants ou entre un composant et des propriétés (caractères). Les attributs sont ainsi attachés aux composants qu'ils caractérisent. Par exemple sur la figure 1, le modèle descriptif de la famille des *Fungiidae* comporte cinq attributs (*label*, *descripteur*, *taxon*, etc.). Ces attributs sont attachés au composant *identification*. Dans cette représentation, les composants *identification* et *contexte* définissent les méta-données associées aux descriptions. Ils sont situés dans le même plan que ceux relatifs aux caractères morphologiques (branche description) afin de disposer d'une vision globale des données et méta-données des descriptions.

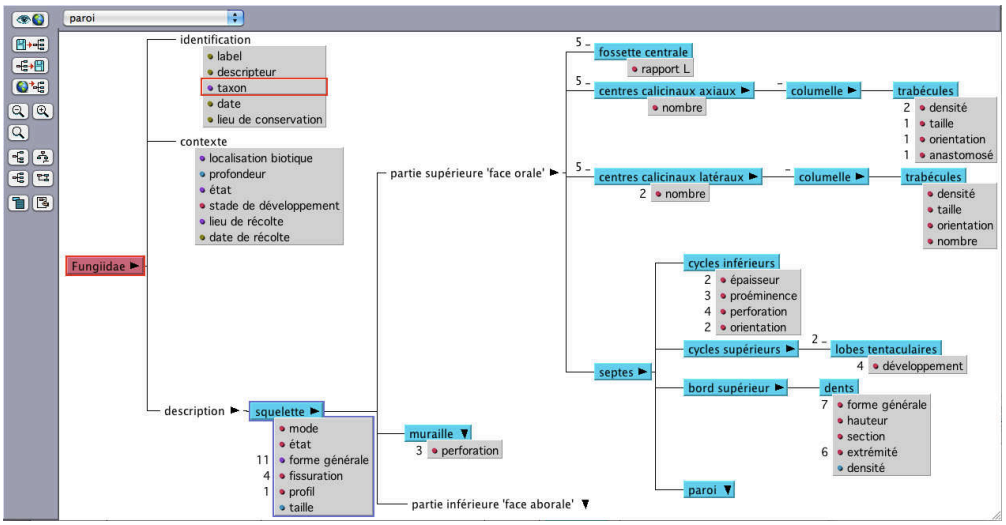


FIG. 1 – *Partie du modèle descriptif de la famille des Fungiidae, extrait d'IKBS.*

Formellement, la structure d'un modèle descriptif est une arborescence $\mathcal{M} = (\mathcal{A}, \mathcal{U})$, où l'ensemble des sommets $\mathcal{A}$ est un ensemble des objets et attributs. Les noeuds non-terminaux sont des objets, appelés également attributs structurés et sont notés $A_j :< A_1, \ldots, A_p >$ où A_j est la racine du sous-arbre dont les fils sont les $A_1, \ldots, A_p$,. Une arête $(A_j, A_l) \in \mathcal{U}$ exprime que A_l est un composant de A_j. Les attributs sont les feuilles de l'arborescence.

L'objectif principal de cette modélisation est de permettre aux experts de bâtir une représentation d'un groupe taxinomique donné, dans l'exemple, une famille corallienne. La structuration est une donnée essentielle du modèle car elle permet de regrouper les caractères qui définissent le même objet observable (*le squelette*, *les différentes sortes de columelles*, *la muraille*, etc.). Certains objets peuvent être absents des descriptions. Par exemple, les *columelles* des différents centres calicinaux (axiaux et latéraux) peuvent être absentes et certains individus du groupe *Fungiidae*[2] peuvent en être dépourvus, ce qui est visualisé par la présence d'un signe - à gauche de l'objet. Ces composants sont qualifiés "absents possibles" ou contingents.

2. La famille des Fungiidae se distingue par le fait que les colonies sont constituées d'un organisme unique (polype) de grande taille, à l'inverse des autres familles.

Les caractères sont définis comme des attributs complexes décrits par plusieurs facettes telles que le type, le co-domaine, la question associée, la pondération, un commentaire explicatif, des liens hypertextuels, des illustrations.

Des règles de dépendance (ou d'implication) entre caractères peuvent être exprimées. Elles sont du type SI $c_1 = v_1$ ALORS $c_2 = v_2$ où les c_i sont des caractères et v_i des états. Ces règles permettent essentiellement de garantir la cohérence des descriptions.

Une description est une sous-arborescence dérivée de celle de A, obtenue par valuation de la présence/absence des objets et par valuation des caractères. Dans l'exemple de la figure 2, les caractères ont été valués par des valeurs complexes. Les valeurs peuvent être de type nominal, ordinal, taxonomique, numérique, discrète ou intervalle, conjonctive ou disjonctive, etc. Les descriptions sont réunies au sein d'une base de cas. La structure arborescente d'une description est appelée *squelette*. Un *squelette* décrit la structure morphologique d'une observation, il renseigne le statut de chaque composant, dont l'état peut être présent (+), absent (-) ou inconnu ($*$) (figure 2). Notons $L = \{+, -, *\}$, une fonction dite "label" $\lambda : \mathcal{A} \rightarrow L$ définit le squelette H_λ par l'arborescence étiquetée par L : $H_\lambda = (\mathcal{A}_\lambda, \mathcal{U})$ avec $\mathcal{A}_\lambda = \{(A_j, \lambda(A_j))_{j \in J}\}$, avec $A_j \in A$.

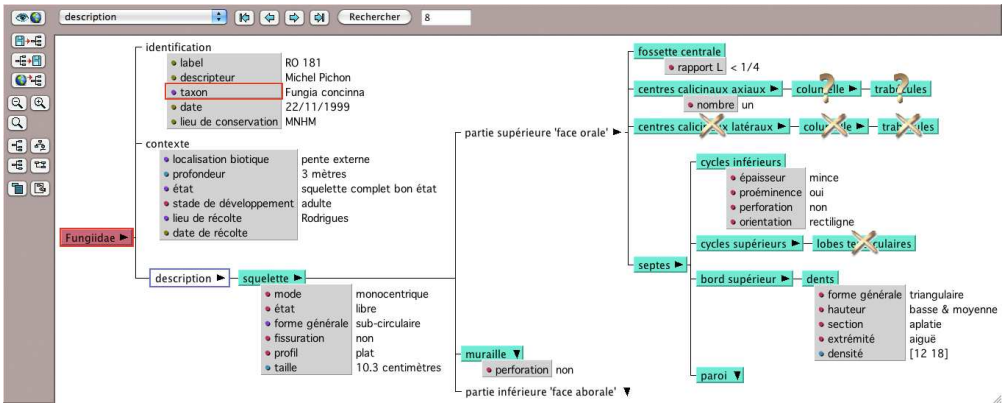


FIG. 2 – *Partie d'une description. Espèce Fungia concina, famille des Fungiidae.*

Sur l'ensemble des squelettes les règles de cohérence suivantes sont imposées. Les fils d'un noeud absent sont absents, les fils d'un noeud inconnu sont inconnus ou absents, les fils d'un noeud présent peuvent être présents, absents ou inconnus. On note $\mathcal{H}$ l'ensemble des squelettes vérifiant les règles de cohérence précédentes.

3 Mesures de similarité

3.1 Mesure de similarité structurelle

Comment mesurer la similarité entre des arbres étiquetés enracinés ? Afin d'accorder une importance différente aux noeuds, selon leur degré de profondeur, nous définissons une fonction de poids m sur les noeuds. Le poids d'un noeud est le nombre de sommets du sous-arbre dont il est la racine. Puis, on se donne une mesure de similarité σ à valeurs dans $[0, 1]$ sur les

TAB. 1 – *La mesure de similarité σ pour un noeud donné*

$\lambda_1 \diagdown \lambda_2$	$+$	$-$	$*$
$+$	1	α_j^1	β_j^1
$-$	α_j^2	1	γ_j
$*$	β_j^2	γ_j	1

fonctions labels. Si $\lambda_1, \lambda_2 \in L^{\mathcal{A}}$, on a $\sigma(\lambda_1(A_j), \lambda_2(A_j)) = 1 \iff \lambda_1(A_j) = \lambda_2(A_j)$ et $\sigma(\lambda_1(A_j), \lambda_2(A_j)) = \sigma(\lambda_2(A_j), \lambda_1(A_j))$.

La similitude structurelle de deux squelettes est alors évaluée comme la moyenne des valeurs de similitude des noeuds. La mesure de similarité structurelle pondérée, notée $\zeta_{SPondere}$ a pour expression :

$$\zeta_{SPondere}(H_1, H_2) = \frac{\Sigma_{j \in J} m(A_j) \sigma(\lambda_1(A_j), \lambda_2(A_j))}{\Sigma_{j \in J} m(A_j)}.$$

Le tableau 1 donne l'expression générale de la mesure de similarité σ pour un noeud donné. Cette mesure de similarité généralise des mesures classiques comme celle de Sokal ou la distance d'édition d'arbre pour un choix particulier des poids des noeuds et des valeurs de σ. Nous donnons ci-dessous des indications pour le choix des valeurs de la mesure de similarité σ (tableau 1) entre les noeuds.

- $\sigma(\lambda_1(A_j) = +, \lambda_2(A_j) = -) = \alpha_j^1$. Si le noeud A_j est une feuille alors $\alpha_j^1 = 0$, sinon sa valeur est le rapport entre le nombre de fils absents du noeud A_j de H_1 sur le nombre de ses fils. α_j^1 mesure l'importance du sous-arbre absent du noeud présent dans H_1, plus ce noeud a des fils absents plus la similarité sera grande avec celui de H_2.
- $\sigma(\lambda_1(A_j) = +, \lambda_2(A_j) = *) = \beta_j^1$. La valeur inconnue $*$ correspond à une présence $(+)$ ou une absence $(-)$ du noeud donné. On calcule dans l'échantillon les probabilités de présence simultanée $Pr(\lambda_1(A_j) = +, \lambda_2(A_j) = +)$ et de présence-absence $Pr(\lambda_1(A_j) = +, \lambda_2(A_j) = -)$. Si la co-présence est la plus probable alors $\beta_j^1 = 1$ sinon $\beta_j^1 = \alpha_j^1$. La valeur par défaut de ce paramètre est 0.
- $\sigma(\lambda_1(A_j) = -, \lambda_2(A_j) = *) = \gamma_j$. Si la configuration $(\lambda_1(A_j) = -, \lambda_2(A_j) = -)$ est la plus probable dans l'échantillon alors $\gamma_j = 1$ sinon sa valeur est 0. Par défaut, sa valeur est 0.
- Les valeurs α_j^2, β_j^2 du tableau 1 sont calculées de manière symétrique.

Les mesures de similarité structurelle pondérée, en sommant sur tous les noeuds, introduisent une redondance pour ceux absents (resp. inconnus), car leurs descendants sont absents (resp. inconnus) selon les règles de cohérence relatives aux squelettes. L'idée est d'affiner cette valeur en considérant la similitude des descendants uniquement pour les noeuds simultanément présents dans les deux squelettes. Soient les squelettes H_1 et H_2. Le degré de similitude est défini de manière récursive pour les noeuds simultanément présents, à l'aide d'une mesure de similitude σ_r définit comme suit : Si A_j est une feuille alors $\sigma_r(\lambda_1(A_j) = +, \lambda_2(A_j) = +) = 1$, Sinon c'est un noeud non terminal $A_j :< A_k >_{k \in K}$ alors :

$$\sigma_r(\lambda_1(A_j) = +, \lambda_2(A_j) = +) = \frac{1 + \Sigma_k m(A_k) \sigma_r(\lambda_1(A_k), \lambda_2(A_k))}{m(A_j)} \quad (1)$$

TAB. 2 – *La mesure de similarité récursive σ_r pour un objet $A_j =< A_k >_{k\in K}$*

$\lambda_1 \diagdown \lambda_2$	$+$	$-$	$*$
$+$	$\frac{1+\Sigma_k m(A_k)\sigma_r(\lambda_1(A_k),\lambda_2(A_k))}{m(A_j)}$	α_j^1	β_j^1
$-$	α_j^2	1	γ_j
$*$	β_j^2	γ_j	1

qui vérifie bien les propriétés d'une mesure de similarité. Le tableau 2 donne l'expression générale de σ_r. La mesure de similarité entre deux squelettes est calculée par une "descente récursive" à partir de la racine A :

$$\zeta_{SRecursive}(H_1, H_2) = \sigma_r(\lambda_1(A), \lambda_2(A)) \tag{2}$$

Dans ce parcours, la détermination des valeurs de similarité entre les noeuds non simultanément présents n'est faite que pour les noeuds les plus élevés de l'arbre.

3.2 Mesure de similarité sur les valeurs

Dans cette partie, nous nous intéressons aux similarités portant sur les valeurs relatives aux attributs (les feuilles de l'arborescence). L'ensemble des attributs est noté : $\{(A_q, D_q)|q \in Q\}$, la valeur d'une observation pour un attribut A_q est dans son domaine D_q si et seulement si l'attribut est présent. On suppose donnée une mesure de similarité s_q sur chaque domaine d'un attribut : $s_q : D_q \times D_q \rightarrow [0, 1]$. Dans le logiciel IKBS, diverses distances adaptées à différents types d'attribut sont disponibles : la distance euclidienne, la distance du khi-deux pour des attributs qualitatifs,... (Grosser et al. (2000)). L'indice de "similarité valeur" entre deux observations portera sur les valeurs des attributs présents en commun. Soit une observation $o \in O$, dont la fonction label est λ_o, on note l'ensemble de ces indices des attributs $Q_+(o) = \{q \in Q|\lambda_o(A_q) = +\}$. L'ensemble des valeurs de o est : $v(o) = (v_q|v_q \in D_q, q \in Q_+(o))$. La mesure de similarité valeur $\zeta_V : O \times O \rightarrow [0, 1]$ est définie pour deux observations o et o' dont les valeurs sont $v(o) = (v_q), v(o') = (v'_q)$ comme :

$$\zeta_V(o, o') = \frac{\Sigma_{q\in Q_+(o)\cap Q_+(o')} s_q(v_q, v'_q)}{|Q_+(o) \cap Q_+(o')|}$$

pour $|Q_+(o) \cap Q_+(o')| \neq \emptyset$ et $\zeta_V(o, o') = 0$ sinon.

3.3 Mesures de similarité globale

Une observation $o \in O$ est décrite par son squelette H_o et ses valeurs $v(o) = (v_q)$. Pour définir un indice de similarité unique qui tienne compte à la fois de la structure et du contenu, on peut considérer la moyenne pondérée des deux indices : structurelle (indice $\zeta_{SPondere}$ ou $\zeta_{SRecursive}$) et valeur ζ_V. Cependant on remarque que les attributs présents dans deux squelettes, sont comptabilisés deux fois, une fois dans la mesure de similarité structurelle pondérée ou récursive, avec une valeur égale à 1 et une autre fois avec la valeur de la mesure de similarité valeur. Pour éviter cette redondance, nous proposons la similarité globale ζ_{GR} en modifiant la

mesure de similarité structurelle récursive proposée de la section 3.1 pour qu'elle prenne en compte la similitude au niveau des attributs présents en commun. Si A_j est une feuille alors on pose : $\sigma_r(\lambda_o(A_j) = +, \lambda_{o'}(A_j) = +) = s_j(v_j, v'_j)$, sinon c'est un noeud non terminal $A_j :< A_k >_{k \in K}$, et l'expression (1) est appliquée. La mesure de similarité globale $\zeta_{GR}(o, o')$ est calculée récursivement à partir de la racine par la formule (2).

4 Classification par voisinages successifs

Afin de déterminer l'appartenance à une classe d'une description partielle décrite par un utilisateur, nous proposons une méthode itérative et interactive de classification par voisinages successifs (CVS). La méthode est hybride, elle met en oeuvre conjointement une stratégie de type k-plus-proches voisins ainsi qu'une étape de sélection de variables discriminantes de type arbre de décision. Dans la suite, on note $(\mathcal{E}, d)$ l'espace métrique des descriptions où $\mathcal{E} = \mathcal{H} \times \mathcal{V}$ avec $\mathcal{H}$ l'ensemble des squelettes, $\mathcal{V}$ l'ensemble des valeurs des observations et $d = 1 - \zeta_{GR}$ l'indice de dissimilarité dérivé de l'indice de similarité global ζ_{GR} précédement défini.

On note par e la description partielle d'un spécimen dont la classe est à déterminer. Les règles d'inférence suivantes sont d'abord appliquées : les descendants d'un noeud non renseigné sont valués à "$*$", les descendants d'un noeud absent sont valorisés à "$-$", les ancêtres d'un noeud présent ou d'une feuille valuée sont considérés comme "présent". Après cela, nous obtenons une description $e \in \mathcal{E}$.

4.1 Principes de l'algorithme CVS

L'algorithme CVS permet de calculer la classe d'appartenance d'un spécimen donné e. Il est itératif et comporte les étapes suivantes :

1. Initialiser la valeur du rayon Δ à la distance maximum de e à l'ensemble des observations.
2. Déterminer l'ensemble des voisins comme l'ensemble des objets à l'intérieur de la sphère de rayon Δ et de centre e,
3. Calculer les scores de classification des classes *a priori*,
4. Calculer la nouvelle valeur du rayon Δ à partir de l'ensemble des voisins,
5. Répéter les étapes 2,3,4 jusqu'à ce que les critères d'arrêt soient satisfaits,
6. A partir des scores de classification, proposer une classe d'appartenance de la description à l'utilisateur.
   Si satisfaction alors fin.
   - sinon à partir de l'ensemble des voisins, calculer le pouvoir discriminant des attributs non encore valués,
   - L'utilisateur est invité à compléter la description avec une ou plusieurs valeurs des attributs suggérés.
   - Réitérer l'algorithme de discrimination par voisinage successif avec cette description complétée.

L'ensemble des voisins de e à l'itération m (étape 2) est défini comme l'ensemble des objets dans la sphère de rayon Δ_m et de centre e : $voisin_{(m)} = \{o \in O | d(e, o) < \Delta_m\}$. la valeur

du rayon est déterminée comme suit : soit la mesure de dissimilarité maximum entre e et un ensemble A : $D_{max}(e, A) = max_{a \in A} d(e, a)$ on choisit pour valeur de Δ_m la distance maximum :

$$\Delta_m = D_{max}(e, N_{(m-1)}). \tag{3}$$

La suite de nombre positifs Δ_m est décroissante, car par définition de l'ensemble des voisins $\Delta_{m+1} = D(e_m, voisin_{(m)}) = max_{o \in voisin_{(m)}} d(e_m, o) < \Delta_m$.

Soient $\{C_l\}_{l \in K}$ l'ensemble des classes a priori d'appartenance possibles. La probabilité de la classe C_l dans le contexte des voisins $N_{(m)}$ est noté $Pr(C_l|N_{(m)}) = \frac{|C_l \cap N_{(m)}|}{|N_{(m)}|}$ ou la fréquence relative de C_l dans l'ensemble de voisins $N_{(m)}$. La classe qui sera proposée comme celle d'appartenance du spécimen e sera choisie à partir des classes telles que les probabilités $Pr(C_l|N_m)$ seront significativement différentes des probabilités à priori des classes $Pr(C_l|O) = \frac{|C_l|}{|0|}$. Les tests statistiques usuels de comparaison de fréquences peuvent être appliqués à ce stade. Le score de classification de la classe C_l (étape 3) à l'itération m est :

$$R_l = \frac{Pr(C_l|N_m)}{Pr(C_l|O)}. \tag{4}$$

La méthode de sélection des variables discriminantes (étape 6) est interactive, car cette liste est calculée à chaque étape du processus d'identification, en fonction de plusieurs critères et d'une réponse demandée à l'utilisateur. Elle est semi-dirigée dans le sens où l'utilisateur peut choisir une autre variable parmi la liste proposée ou bien apporter une réponse inconnue. Dans ce cas, la variable en seconde position sera automatiquement choisie.

La méthode utilise une combinaison de plusieurs critères de choix, de façon à minimiser le nombre de questions et prendre également en compte certaines connaissances de fond du domaine relatives à la qualité des variables, exprimées dans le modèle descriptif :

1. Construction de la liste des attributs éligibles. Cette liste est contextuelle, constituée des attributs pouvant être sélectionnés. Elle dépend aussi de la structure du modèle et des attributs déjà renseignés. La présence des attributs éligibles doit notamment être vérifiée pour chaque composant contingent. Une question booléenne portant sur sa présence/absence est ajoutée à la liste.
2. Choix de critère classique de calcul du gain d'information utilisé en apprentissage, tel que l'*entropie de Shannon* ou le *Gini Index*. Ce type de critère permet de minimiser la longueur de la diagnose en privilégiant les attributs les plus discriminants.
3. Pondération des attributs. Chaque attribut peut être pondéré par l'expert sur une échelle réelle de 0 à 1. Le poids 0 est affecté à un attribut non discriminant. Les pondérations sont très utiles pour introduire une connaissance d'ordre stratégique sur les caractères. Certains attributs sont en effet particulièrement difficile à observer (nécessite du matériel spécifique), à décrire ou sujet à interprétation. L'expert peut ainsi choisir de minimiser leur occurrence.

Plusieurs critères d'arrêt sont considérés : le score de classification d'une classe C_l (cf. formule 4) supérieure à un seuil fixé par l'utilisateur. le nombre minimal de voisins, car le rayon Δ_m est une suite décroissante (cf. formule 3). le nombre maximal d'itérations. la description est complète ou bien il n'y a plus de variables discriminantes disponibles. Notons que ces différents critères sont exclusifs et ne peuvent être vérifiés simultanément.

5 Applications

Dans cette partie, nous procédons à différents tests de validation en grandeur réelle sur des données structurées issues de la base de connaissances sur les coraux des Mascareignes. L'objectif poursuivi est double. D'une part (section 5.1), illustrer l'exécution de la méthode CVS dans l'environnement IKBS sur un exemple réel. D'autre part (section 5.2), comparer les performances relatives de la méthode CVS par rapport à un classement par arbre d'identification et un classement par la méthode des k-plus-proches voisins.

5.1 Classement d'une description

L'exemple choisi est l'illustration du processus de classement d'une description de référence e_r appartenant à l'espèce *Fungia concinna*, de la famille des *Fungiidae*, donné pour exemple dans la partie 2. La famille des *Fungiidae* est structurée en quatre genres (*Cycloseris*, *Fungia*, *Herpolita* et *Podabacia*) et onze espèces (cf. figure 3). La base *Fungiidae* compte 63 descriptions (cas) comportant 94 caractères (attributs) et 15 taxons (classes).

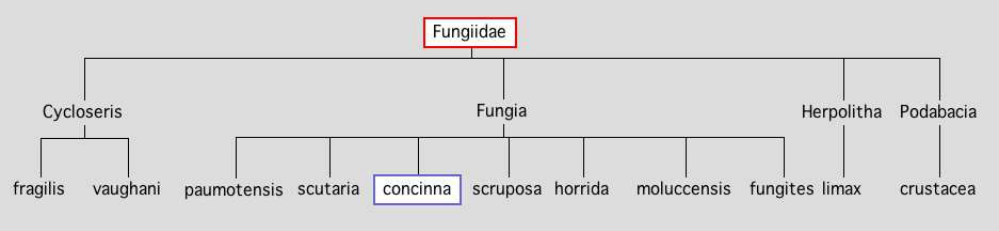


FIG. 3 – *Taxonomie de la famille des Fungiidae, comprenant quatre genres et onze espèces.*

L'algorithme cherche à associer à e_m (le cas test, initialement vide) un élément de la hiérarchie de taxons (fig. 3), en exploitant les informations du modèle descriptif et de la base de cas de référence. Afin de simuler l'interactivité avec l'utilisateur, la source de données de e_m est e_r. A chaque itération, e_m est alimenté par un attribut de e_r. L'étiquette de e_r est utilisée à la fin du processus pour vérifier que le classement est correct ou non. La mesure du gain d'information utilisée dans cette application est la mesure de l'entropie de Shannon (gain d'information). Dans l'exemple, l'algorithme CVS a conduit au bon classement de e_m en 21 itérations.

La table 3 illustre un sous-ensemble choisi d'itérations qui ont conduit au bon classement de e_t. Pour chaque itération, les informations suivantes sont renseignées : le numéro de l'itération, l'attribut sélectionné et le gain d'information associé, la valeur correspond à l'attribut pour le cas de référence affectée à e_m, ainsi que des informations relatives au voisinage de e_m. Pour chaque voisinage, sont donnés les numéros des cas de la base de référence les plus proches, le taxon (classe) associé, la valeur de l'attribut sélectionné, ainsi que la distance du cas e_m. Notons que pour des raisons pratiques, seuls les trois cas les plus proches du voisinage sont affichés.

Différents types d'attributs sont sélectionnés lors de cette application : type hiérarchique,nominal, continu, intervalle ou booléen portant sur la présence d'un composant (présence des lobes-

TAB. 3 – *Illustration du classement d'une description structurée par la méthode CVS.*

Num iter	Attribut	Valeur	Voisinage			
			cas	classe	valeur	Dist
1	forme[squelette]	circulaire	62	fungites	circulaire	0,2216
	gain = 0,628		46	fungites	circulaire	0,2239
			51	scruposa	circulaire	0,2259
7	taille[squelette]	10.3 cm	62	fungites	9,5 cm	0,1905
	gain = 0,423		46	fungites	9 cm	0,1930
			60	fungites	15,5 cm	0,1974
8	densité[épines]	[10 16]	62	fungites	[6 10]	0,1912
	gain = 0,423		46	fungites	6	0,1946
			60	fungites	[6 8]	0,1978
9	profil[squelette]	plat	60	fungites	plat	0,1924
	gain = 0,394		62	fungites	convexe	0,1935
			46	fungites	plan convexe	0,1968
19	dév[côtes]	sub égales	46	fungites	sub égales	0,1901
	gain = 0,211		**39**	**concinna**	sub égales	0,1915
			60	fungites	inégales	0,1930
20	lobes-tentaculaires	absent	46	fungites	absent	0,1901
	gain = 0,173		**39**	**concinna**	absent	0,1915
			60	fungites	absent	0,1930
21	forme[épines]	cylindre	39	**concinna**	cylindre	0,1912
	gain = 0,153		46	fungites	conique	0,1935
			62	fungites	cylindre	0,1941

tentaculaires). Pour les besoins de notre application, le critère d'arrêt qui s'applique ici est la correspondance entre la classe du cas de référence et la classe du premier voisin.

L'information la plus intéressante à observer dans ce tableau est la liste des cas les plus proches. Les variations de position des voisins permet d'évaluer intuitivement dans quelle mesure l'adjonction d'une information supplémentaire modifie les distances des éléments du voisinage au cas e_m et par conséquent modifie l'ordre des voisins. Ainsi par exemple, lors du passage de l'itération 8 à 9, le cas 60 passe en première position du fait que la valeur du profil du squelette (valeur *plat*) correspond à la valeur de référence. A l'itération 19, apparait dans la liste des trois candidats le cas 39 en position 2, dont la valeur de classe correspond à la classe recherchée. Enfin, à l'itération 21, le cas 39 passe en première position devant le cas 46, du fait d'une correspondance avec l'attribut forme[épines], ce qui aboutit à un bon classement.

5.2 Performances de la méthode CVS

Dans cette seconde expérience, nous souhaitons tester les performances relatives de l'algorithme CVS par rapport à deux méthodes de référence disponibles dans IKBS : les arbres d'identification et les k-plus-proches voisins. La méthode des arbres d'identification (Grosser (2002)) est une extension de la méthode de classification supervisée C4.5 (Quinlan (1993))

adaptée aux données complexes de notre modèle de représentation. La méthode est monothétique (les caractères sont proposés un à un) et interactive. Le modèle inductif utilisé pour le classement des attributs discriminants est identique à celui utilisé par la méthode CVS lors de la phase de sélection. Le second algorithme est de type k-plus-proches voisins. Dans l'expérience, le facteur k est simplement positionné à 1. La mesure de similarité utilisée est la mesure récursive ζ_{GR} exposée précédemment. La classe du cas le plus proche est affectée au cas test et la totalité de l'information du cas de référence est utilisée. La méthode est polythétique, un ensemble de caractères devant être renseignés *a priori*. La mesure de similarité utilisée est la même que celle de l'algorithme CVS. La méthode de validation utilisée, de type "Leave-on-out", consiste à classer chaque cas de la base en utilisant les autres cas comme base de référence. La méthode est appliquée pour les trois algorithmes dans des conditions identiques. Les paramètres de classement sont identiques à ceux utilisés pour la première expérience.

TAB. 4 – *Tests de validation "Leave-on-out" de différentes bases*

Bases	nb class	nb cas	nb attr	Arbre ident		K-voisins		CVS	
				score	taux	score	taux	score	taux
Faviinae	36	92	146	65	70,65%	85	92,39%	84	91,30%
Montastreinae	15	24	118	17	70,83%	22	91,66%	19	79,16%
Fungiidae	15	63	94	47	74,60%	58	92,06%	55	87,30%
Mussidae	15	56	28	49	87,50%	54	96,42%	51	91,07%
Poritidae	28	28	87	22	78,57%	24	85,71%	19	67,85%
Siderastreidae	14	60	99	49	81,67%	56	93,33%	57	95,00%

La table 4 illustre les résultats des trois méthodes de classification sur 6 bases extraites de la base coraux. Pour chaque base, les informations suivantes sont affichées : nombre de taxons pour le modèle (nb class), nombre de cas (nb cas), nombre d'attributs (nb attr). Ensuite pour chaque méthode est donné le nombre (score) et le taux de cas bien classés.

Globalement, la méthode la moins robuste est la méthode par arbre d'identification. Les erreurs de classement sont fréquentes, de 12.5% sur la base *Mussidae* à $29,35\%$ pour la base *Faviinae*. La méthode la plus robuste est la méthode de type k-voisins, qui exploite la totalité de l'information disponible lors du calcul des similarités 2 à 2. Les taux d'échec pour celle-ci ce situent entre 14.29% (base *Poritidae*) et 3.58% (base *Mussidae*). La méthode CVS offre un score intermédiaire relativement proche de la méthode K-voisins et parfois très supérieur à la méthode par arbre. Observer par exemple les résultats de la base *Faviinae* qui montrent un écart de plus de 20% de bonnes identifications avec la méthode CVS.

La méthode K-voisins donne de bons résultats, mais est difficilement exploitable par des utilisateurs non-experts dans des cas réels d'utilisation. Il est en effet très difficile de choisir *a priori* un sous-ensemble de caractères discriminants, sans une connaissance approfondie du domaine. Il est donc très utile de disposer d'un processus interactif qui guide l'observation et suggère les caractères à décrire. L'approche monothétique (un caractère à la fois) proposée par la méthode des arbres d'identification offre cette facilité. C'est la raison pour laquelle cette méthode est très utilisée, par les utilisateurs d'IKBS, experts et biologistes et par toutes les personnes qui ont besoin d'identifier des organismes, en particulier pour l'évaluation et la

conservation de la biodiversité. Nous pensons que la méthode CVS offre une solution alternative intéressante, monothétique et interactive, plus performante que les méthodes par arbres.

6 Conclusion

Dans cet article nous avons proposé un modèle de représentation et une méthode de classification de données complexes implantée et évaluée à l'aide de l'outil IKBS pour la conception d'une base de connaissances sur les coraux des Mascareignes. Une expérimentation sur différentes bases montre que la méthode présente une assez bonne résistance aux bruits comparativement aux méthodes de classement par arbres d'identification, tout en offrant des caractéristiques intéressantes d'interactivité et d'explication. L'approche est générique et applicable à tout domaine où les données se présentent sous forme d'arborescences composées d'attributs hétérogènes.

Références

Conruyt, N. (1994). *Amélioration de la robustesse des systèmes d'aide à la description, à la classification et à la détermination des objets biologiques*. Thèse de doctorat, Université Paris IX-Dauphine.

Grosser, D. (2002). *Construction de bases de connaissances descriptives et classificatoires avec la plate-forme à objets IKBS. Application à la systématique des coraux des Mascareignes*. Ph. D. thesis, Université de la Réunion.

Grosser, D., N. Conruyt, et Y. Geynet (2003). Représentation de connaissances descriptives et classificatoires : le modèle codesc. In *Actes des 9èmes journées francophones "Langages et modèles à objets", Revue Sciences et Technologies de l'information (RSTI), série l'Objet, Hermès (Ed.)*.

Grosser, D., J. Diatta, et N. Conruyt (2000). Improving dissimilarity functions with domain knowledge, applications with ikbs system. In *PKDD*, pp. 409–415.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning.

Sokal, R. et P. Sneath (1963). *Principles of numerical taxonomy*. San Francisco et Londres: W.H. Freeman et Cie.

Summary

We propose a new classification method for complex biological data which is iterative, interactive and semi-directed. It combines inductive techniques for the choice of discriminating variables and search for nearest neighbors based on various similarity measures which take into account structures and values of the objects for the neighborhood computation.

Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets: Application to cancer expression data

Sidahmed Benabderrahmane*, Marie-Dominique Devignes*, Malika Smail-Tabbone*,
Olivier Poch **, Amedeo Napoli*, Wolfgang Raffelsberger**,
Dominique Guenot ***, N.Hoan Nguyen **,
and Eric Guerin ***.

*LORIA (CNRS, INRIA, Nancy-Université), Équipe Orpailleur, Campus scientifique,
54506 Vandoeuvre-lès-Nancy Cedex, France.
benabdsi@loria.fr, http://www.loria.fr/~benabdsi.
**LBGI, CNRS UMR7104, IGBMC, 1 rue Laurent Fries, 67404 Illkirch, France.
***INSERM U682, 3 avenue Molière, Strasbourg, France.

Abstract. Clustering algorithms rely on a similarity or distance measure that directs the grouping of similar objects into the same cluster and the separation of distant objects between distinct clusters. Our recently described semantic similarity measure (*IntelliGO*), that applies to functional comparison of genes, is tested here for the first time in clustering experiments. The dataset is composed of genes contained in a benchmarking collection of reference sets. Heatmap visualization of hierarchical clustering illustrates the advantages of using the *IntelliGO* measure over three other similarity measures. Because genes often belong to more than one cluster in functional clustering, fuzzy C-means clustering is also applied to the dataset. The choice of the optimal number of clusters and clustering performance are evaluated by the F-score method using the reference sets. Overlap analysis is proposed as a method for exploiting the matching between clusters and reference sets. Finally, our method is applied to a list of genes found dysregulated in cancer samples. In this case, the reference sets are provided by expression profiles. Overlap analysis between these profiles and functional clusters obtained with fuzzy C-means clustering leads to characterize subsets of genes displaying consistent function and expression profiles.

1 Introduction

1.1 Transcriptomic data analysis

In recent years, DNA microarrays technologies have become an important tool in genomics, allowing the measure of the *expression level* of several thousands of genes in different biological situations. Using these technologies and clustering approaches, *expression profiles*

can be produced by grouping together genes displaying similar expression levels in a set of situations.
Usually a functional analysis is then applied to genes from the same expression profiles in order to associate the profiles with one or more common biological functions, derived from functional annotations. The main purpose of this processing, known as functional profiling, is to identify and characterize genes that can serve as diagnostic signatures or prognostic markers for different stages of cancer.
Among the most commonly used functional annotations of genes are the Gene Ontology terms. The Gene Ontology (GO) is one of the most important tool in bioinformatics, consisting of about 30,000 terms. It is organized as a controlled vocabulary, represented as a rooted Directed Acyclic Graph (rDAG) in which GO terms are the nodes connected by different hierarchical relations (mostly *is_a* and *part_of* relations). This rDAG is covering three orthogonal aspects or taxonomies, namely the *biological process* (BP), *molecular function* (MF), and *cellular component* (CC) aspects of gene annotation (Consortium, 2010). The process of annotating a gene with a given GO annotation is summarized by an evidence code (EC), which reflects the quality of this association (Rogers and Ben-Hur, 2009).
GO annotations are widely used in several complex data mining problems relating to bioinformatics domains. However, it is still a challenge for biologists and computer scientists to analyze and use such a huge amount of data, growing in an exponential way. Authors as (Khatri and Draghici, 2005; Speer et al., 2004; Huang et al., 2009), used gene functional analysis in order to interpret DNA microarrays experiments, using the assumption commonly admitted that genes having similar expression profile should share similar biological function(s). Functional similarity between genes or gene products relies on measuring the similarity between their GO annotation terms. Many GO similarity measures have been described so far (Pesquita et al., 2009), some of which have been used for functional clustering (Speer et al., 2005; Adryan and Schuh, 2004; Brameier and Wiuf, 2007).

1.2 Functional Similarity Measures

The notion of similarity measure is usually applied to objects sharing common attributes or characteristics (Blanchard et al., 2008). In the biological domain, these objects are generally genes or gene products annotated with GO terms. As the GO terms are organized in a rDAG, it is then possible to exploit the relationships between terms and define semantic similarity measures. In (Pesquita et al., 2009; Benabderrahmane et al., 2010) is presented the state of the art of a variety of semantic similarity measures. At the level of the individual GO terms, two categories of measures are reviewed, namely the *edge-based* measures which rely on edge counting in the GO graph, and the *node-based* measures which exploit the information content (*IC*) of both terms of the comparison and of their closest common ancestor. At the level of genes or gene products, various strategies, either *pair-wise* or *group-wise*, are used to combine the similarities between annotation terms. We describe here the four semantic similarity measures involved in this study.
One of the most known measures in the *pair-wise* and *node-based* category is the *Lord* similarity measure (Lord et al., 2003), which is based on the generic semantic similarity measure described by Resnik (Resnik, 1995). Referring to the information theory, the (*IC*) of a concept represents the probability of occurring of this term or any of its descendants in an annotation

corpus, i.e. $IC(t_i)$=$-Log(p(t_i))$. Given two terms t_i and t_j, Resnik introduces the use of the *IC* of their least common ancestor (LCA). The Resnick measure is then defined as:

$$SIM_{Resnik}(t_i, t_j) = IC(\textit{LCA}(t_i, t_j)) \tag{1}$$

If g_1 and g_2 are two genes or gene products represented respectively by two sets of n and m GO terms noted $t_{1,i}$ and $t_{2,j}$, the functional similarity defined by Lord's measure is calculated as the average of all pair-wise Resnik's similarities between their annotation GO terms:

$$SIM_{Lord}(g_1, g_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} SIM_{Resnik}(t_{1,i}, t_{2,j})}{n * m}. \tag{2}$$

Nagar and Al-Mubaid proposed a *pair-wise* and *edge-based* approach in which the shortest path length (*SPL*) between each pair of annotation terms is calculated (Nagar and Al-Mubaid, 2008a). Comparing g_1 and g_2 involves calculating SPL($t_{1,i}$,$t_{2,j}$), $\forall$ i$\in$[1,n], $\forall$ j$\in$[1,m]. Then, the average of all *SPL* values is calculated, it represents the average *SPL* between the two genes:

$$SPL(g_1, g_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} SPL(t_{1,i}, t_{2,j})}{n * m}. \tag{3}$$

Finally, to convert this average *SPL* value into a similarity value, a transfer function is applied to it:

$$SIM_{Al-Mubaid}(g_1, g_2) = e^{-0.2.SPL(g_1,g_2)}. \tag{4}$$

An example of *group-wise* measure is given by the Sim_{GIC} (Graph Information Content) measure which involves both *node-based* and *edge-based* similarities (Pesquita et al., 2008). Here, the sets of annotation terms representing the genes or the gene products are extended with all ancestor terms up to the root. They will be noted g_1^+ and g_2^+. Then, the similarity between g_1 and g_2 is calculated using the *Jaccard* union-intersection formula, by replacing the cardinal of the union and intersection sets with the sum of the *IC*s of the terms they contain. This measure is also known as the weighted Jaccard measure:

$$SimGIC(g_1, g_2) = \frac{\sum_{t \in g_1^+ \cap g_2^+} IC(t)}{\sum_{t \in g_1^+ \cup g_2^+} IC(t)}. \tag{5}$$

In (Benabderrahmane et al., 2010), we have proposed a new vector-based measure, called *IntelliGO*, defined in a novel *annotation Vector-Space Model* (VSM), analogous to the classical VSM described for document retrieval (Salton and McGill, 1983; Polettini, 2004). In the *IntelliGO VSM*, each gene is represented as a vector $\vec{g}$ in a k-dimensional space, where the base vectors $\vec{e_i}$ correspond to k annotation terms. To take into account the semantic relationships between terms, a specific dot product is defined as:

$$\vec{e_i} * \vec{e_j} = \frac{2 * Depth(LCA)}{MinSPL(t_i, t_j) + 2 * Depth(LCA)}. \tag{6}$$

This definition is an extension of the *generalized Cosine Similarity* measure introduced by (Ganesan et al., 2003) and applied here to the GO rDAG. Moreover, we include in the *IntelliGO* VSM an original weighting scheme α_i, assigned to each $\vec{e_i}$ so that the gene representation becomes: $\vec{g} = \sum_i \alpha_i.\vec{e_i}$. The coefficients ($\alpha_i$) combine on one hand, a weight $w(g, t_i)$ depending

on the evidence code tracking the annotation of a gene g by a GO term t_i, and, on the other hand, the *Inverse Annotation Frequency* ($IAF(t_i)$), which is an *IC* estimation of a term t_i.

In total, the similarity between $\vec{g_1}$ and $\vec{g_2}$ is given by the following generalized cosine formula:

$$SIM_{IntelliGO}(g,h) = \frac{\vec{g_1} * \vec{g_2}}{\sqrt{\vec{g_1} * \vec{g_1}}\sqrt{\vec{g_2} * \vec{g_2}}}. \quad (7)$$

1.3 Functional Clustering

Clustering algorithms rely on a similarity or distance measure that directs the grouping of similar objects into the same cluster and the separation of distant objects between distinct clusters (Macqueen, 1967). Clustering algorithms have been used in several domains, with the purpose of data reduction, hypothesis testing and prediction (Theodoridis and Koutroumbas, 2006). There are a multitude of clustering algorithms, but all of them are based on the same basic steps: feature selection, choice of the similarity or distance measure, grouping criterion and techniques, validation and interpretation of the results (Theodoridis and Koutroumbas, 2006; Rousseeuw, 1987).

In biology, clustering is often required for grouping genes or gene products with similar functions. The so-called functional clustering relies on a variety of metrics applied to expression levels, GO annotations, etc (Eisen et al., 1998; Huang et al., 2009; Adryan and Schuh, 2004; Wang et al., 2007). Two major categories of clustering algorithms are used in bioinformatics. *Hierarchical clustering* algorithms are popular because the resulting dendrograms are easily interpreted visually (Eisen et al., 1998). *Al Mubaid et al.* used hierarchical clustering for validating their functional similarity measure, by calculating the silhouette index of clusters generated with genes belonging to yeast pathways (Nagar and Al-Mubaid, 2008b). One limitation of this category of algorithms is that they do not allow overlap between clusters.

The second category concerns *partitional clustering* algorithms like the k-means and fuzzy C-means (FCM) algorithms. Gash and Eisen used the FCM algorithm to identify overlaps that may exist between clusters relating to yeast gene expression data (Gasch and Eisen, 2002). In (Speer et al., 2005), the authors presented a functional clustering approach using the k-means method and the functional similarity measure presented in (Jiang and Conrath, 1997).

The two categories of clustering algorithms are used in this paper with the *IntelliGO* semantic similarity measure. Previous results had shown that this measure displays a robust discriminating power between predefined sets of genes (Benabderrahmane et al., 2010). However clustering results obtained with this measure were neither reported nor compared with other measures. In a first step we explore a dataset of genes representing a collection of reference sets (KEGG pathways, Kanehisa et al. (2010)). Using hierarchical clustering and heatmap visualization, we compare the results obtained with the *IntelliGO* measure and those obtained with three other similarity measures. Then we optimize the *IntelliGO*-based FCM clustering using the reference sets and the F-score method (van Rijsbergen, 1979). We also propose an approach called *overlap analysis* that aims to exploit the matching between clusters and reference sets. In a second step, we explore a list of genes selected from a transcriptomic cancer study. We confront *IntelliGO*-based clustering results with the *fuzzy Differential Expression Profiles* (fuzzy DEP) defined in (Benabderrahmane et al., 2009). Overlap analysis of the *IntelliGO*-based FCM clus-

ters leads to the identification of consistent subsets of genes which are further characterized with respect to GO-term enrichment.

2 Experimental Design

2.1 Presentation of the Datasets

We decided to evaluate the clustering results using a collection of reference sets composed of 13 human KEGG pathways (Kanehisa et al., 2010). In Table(1) are presented the pathways with the number of genes they contain. The similarity values were calculated by considering only the BP aspect of GO, assuming that genes belonging to the same pathway are often referring to similar biological process. Let us consider $List1$ the list containing the 280 genes present in the 13 human pathways.

Human Pathways accession (Hsa:)	00040	00920	00140	00290	00563	00670	00232	03022	03020	04130	03450	03430	04950
Genes Nb	26	13	17	11	23	16	7	38	29	38	14	23	25

TAB. 1 – *Reference dataset composed of a list of 13 human KEGG pathways. The number of genes present in each pathway is displayed (Gene Nb).*

Another list of genes relating to colorectal cancer was used as an applicative example after the evaluation the *IntelliGO* clustering based method. This list of 128 genes that were found dysregulated in cancer samples, is named $List2$, and corresponds to the 222 genes studied in (Benabderrahmane et al., 2009) from which 94 genes were excluded because they lack GO annotation.

2.2 Calculation of similarity matrices

Pair-wise similarity matrices were calculated for $List1$ and $List2$ using the *IntelliGO* measure, and for $List1$ only the three other similarity measures described in section (1.2). These matrices serve as input for the clustering process. The *C++* programming language was used to implement the *Lord, Al-Mubaid* and *IntelliGO* [1] measure, due to its good memory management and calculation speed. The Sim_{GIC} measure is available in the *csbi.go* package within R Bioconductor [2] (Ova).

2.3 Clustering programs

Hierarchical clustering, heatmap visualization and FCM clustering were performed using R Bioconductor. The F-score method and the strategy of overlap analysis were implemented using *C++* programming language.

1. http://bioinfo.loria.fr/Members/benabdsi/intelligo_project/
2. www.bioconductor.org

3 Results

3.1 Comparison of heat maps obtained with four different functional similarity measure results

Four pairwise similarity matrices were generated from $List1$, using three semantic similarity measures, namely: SIM_{Lord}, $SIM_{Al-Mubaid}$, SIM_{GIC}, in addition to our measure $SIM_{IntelliGO}$. The heatmaps generated after hierarchical clustering are presented in Figure 1. Color scale ranges from dark red for very similar genes to dark blue for very dissimilar genes. The heatmap visualization obtained with the SIM_{Lord} measure (Panel A) reveals a very fuzzy

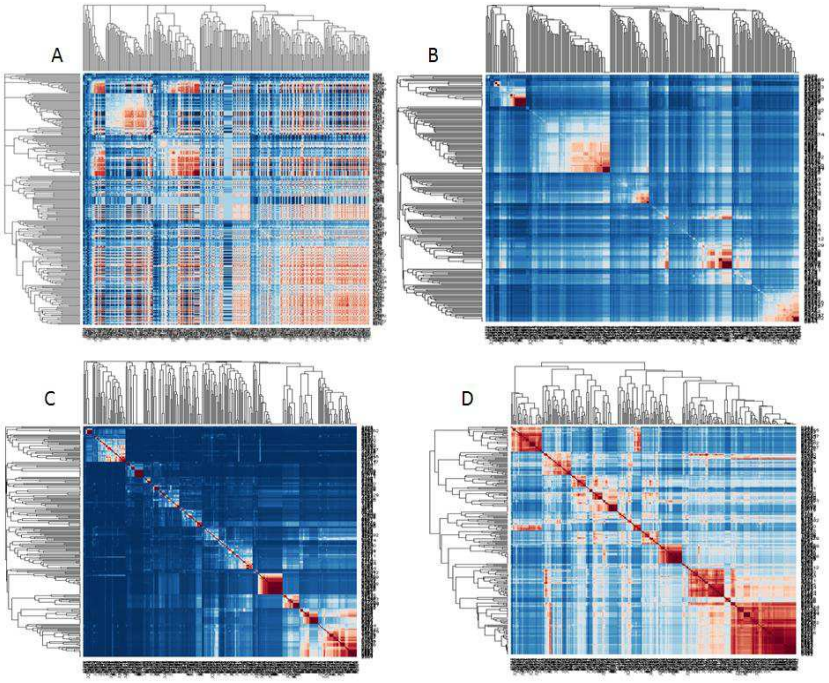


FIG. 1 – *Heatmaps generated after hierarchical clustering using the similarity matrices obtained with (A): SIM_{Lord}, (B): $SIM_{Al-Mubaid}$, (C): SIM_{GIC}, and (D): $SIM_{IntelliGO}$ semantic similarity measures. Genes belong to human pathways.*

color distribution associated with a quite imperfect grouping of similar genes in clusters around the diagonal. This confirms our previous observation that the SIM_{Lord} measure does not efficiently discriminate between genes belonging to two different pathways (Benabderrahmane et al., 2010). The situation is globally reversed with the $SIM_{Al-Mubaid}$ and SIM_{GIC} measures (Panels B and C respectively). These two heatmaps present a limited number of small well delineated clusters around the diagonal, with some other clusters displaying weak (light blue color) intra-set similarity and very few if any cross-similarity between clusters. This also confirms our previous findings concerning the variations in the discriminative power of these measures depending on the pathway (Benabderrahmane et al., 2010). Finally the heatmap ob-

tained with the $SIM_{IntelliGO}$ measure (Panel D) appears well balanced in terms of color scale usage. More than 10 clusters of various sizes can be clearly identified around the diagonal, as well as cross-similarities between clusters.
Further observation of the heatmaps reveals that for two of them, the cells in the diagonal of the heatmap are seldom of dark red color which means that the self-similarity is rarely maximal with these two measures (SIM_{Lord} and $SIM_{Al-Mubaid}$). On the contrary it can be checked that with the SIM_{GIC} and $SIM_{IntelliGO}$ measures, the self-similarity is always maximal (equal to 1) as expected.
It thus appears that heatmaps constitute an interesting visual mean of estimating the performance of a similarity measure for clustering genes from a collection of reference sets. It should be noted here that very similar results were obtained when using another collection of reference sets, namely yeast instead of human pathways.
In this study we will continue working with the $SIM_{IntelliGO}$ measure which produced the most informative heatmap. However as mentioned above, hierarchical clustering is not the most appropriate method for functional clustering of genes because a given gene often belongs to more than one cluster. In fact, in the collection of genes studied here, several genes are involved in multiple biological processes, and therefore belong to multiple pathways.

3.2 Fuzzy clustering approach using a collection of reference sets

The same list of genes ($List1$) was studied for fuzzy clustering using the $SIM_{IntelliGO}$ measure for producing the similarity matrix and the fuzzy C-means (FCM) algorithm. A matching analysis leading to the calculation of an average F-score has to be conducted in order to discover the optimal (k) number of fuzzy clusters (Cleuziou, 2010).

Knowing that our $List1$ corresponds to a collection of 13 pathways, we varied the number of generated clusters k from 11 to 17. For each k value, we calculate the precision and the recall of the reference sets in the best matching clusters leading to individual F-scores which are then averaged to give an average F-score reflecting the quality of the fuzzy clustering. The results are presented in Table (2).

k value	11	12	13	14	15	16	17
Average F-Score using IntelliGO	0.59	0.61	0.61	0.62	0.56	0.55	0.54

TAB. 2 – *Variation of the F-Score when varying (k) number of FCM clusters with IntelliGO measure.*

It can be seen that all F-Score values are greater than 0.5, with a maximum value of 0.62 for $k = 14$. This means that the genes of the 13 human pathways considered in $List1$ are grouped at best with our measure into 14 functional clusters. This result can be easily explained by the fact that pathways of the KEGG database do present some overlaps due to genes being involved in multiple biological processes. Similar results have been observed when dealing with genes from 13 yeast pathways (not presented here).

3.3 Overlap analysis between cluster and reference sets

A possible exploitation of our fuzzy clustering experiment relies on a careful investigation of cluster content by domain experts. For this purpose, we defined and applied a generic overlap analysis method between cluster (C_i) and reference set (R_j). The strategy is illustrated in Figure 2.

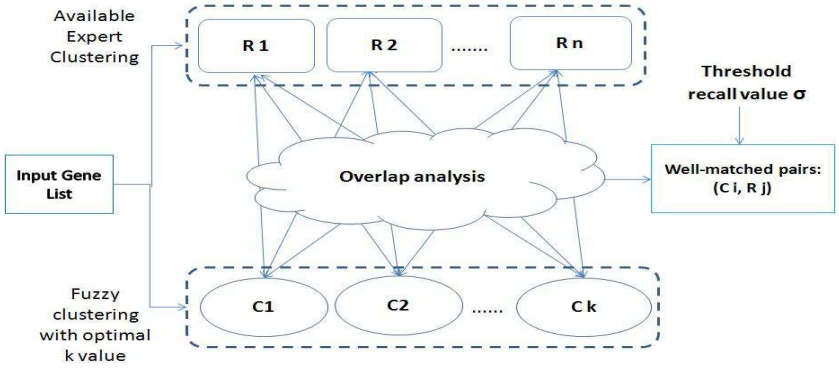


FIG. 2 – *Strategy for overlap analysis between clusters (C_i) and reference sets (R_j).*

Each of the k clusters produced with the optimal k value (see above) is compared with each reference set. A recall value is calculated as the ratio between the number of genes from the reference set present in the cluster and the total number of genes in the reference set. A well-matched pair is thus defined as the association of a cluster with a reference set displaying a recall value above a certain threshold. This threshold recall value is set to get at least one well-matched pair for each cluster and each reference set. In consequence more than one well-matched pairs can be produced for some clusters.
Well-matched pairs (C, R) constitute interesting datasets for further analyses. The intersection $C \cap R$ is expected to display a highly homogeneous content composed of genes known as members of a reference set and found most similar by clustering. Alternatively, the two set-theoretic differences $C \backslash R$ and $R \backslash C$ can be studied in order to discover missing information. The former difference $(C \backslash R)$ contains genes that are similar to genes from the reference set but not counted among its members. This difference content can be presented to an expert in order to check whether some genes from $C \backslash R$ could be missing members of R. The latter difference $(R \backslash C)$ contains genes that are members of the reference set but do not get clustered with most other members on the basis of similar functional annotation. The annotation of genes from $R \backslash C$ can be scruted by an expert in order to check whether some terms could be missing. A GO-term enrichment study of cluster C (Eden et al., 2009) can then be conducted in order to propose the most relevant GO terms for completing gene annotation in $R \backslash C$.

3.4 Application to a dataset relating to colorectal cancer

In this section, we present an application of the *IntelliGO*-based clustering and overlap analysis approach using $List2$ which is composed of 128 genes relating to colorectal cancers.

The idea here is to confront functional clusters generated with *IntelliG0* measure and *fuzzy Differential Expression Profiles* (fuzzy DEP) obtained from the same list of genes. We believe that overlap analysis may lead to discover hidden relationships between gene expression and biological function. Fuzzy DEPS are considered here as a collection of reference sets for overlap analysis. More precisely, eight fuzzy DEPs containing genes with GO annotation are retained from our previous study (Benabderrahmane et al., 2009).
The pair-wise similarity matrix was generated for the 128 genes of $List2$. Then, as a first step, the heatmap showing the resulting of hierarchical clustering was produced Figure (3). Despite of a high level of cross similarities in this dataset, several clusters can be distinguished around the diagonal of the heatmap. Fuzzy clustering was applied in a second step. The number of

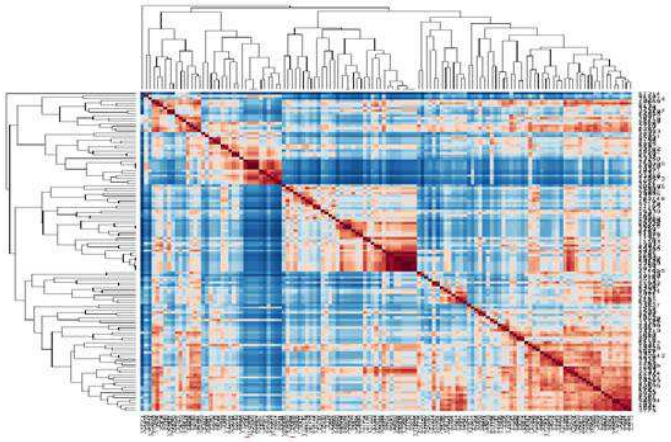


FIG. 3 – *Heatmap generated from IntelliGO pair-wise similarities of colorectal cancer genes ($List2$).*

clusters, k, was optimized with the F-score method using the 8 fuzzy DEPs as reference sets. Table (3) shows the values obtained for k varying from 2 to 14. The optimal value is 0.4 for $k = 3$.

K generated clusters using IntelliGO	2	3	4	5	6	7	8	9	10	11	12	13	14
Average F-Score	0.39	0.4	0.37	0.32	0.26	0.27	0.29	0.29	0.29	0.28	0.26	0.25	0.16

TAB. 3 – *Variation of the F-Score for k value varying from 2 to 14. FCM clustering was performed on $List2$, a list of 128 genes found dysregulated in colorectal cancer samples. The 8 fuzzy DEPs previously extracted from these 128 genes are taken as reference sets for F-score calculation.*

The three functional clusters produced for $k = 3$ were then studied by overlap analysis as described above, in order to extract lists of genes displaying both functional similarity and similar expression profile. The list of well-matched pairs maximizing the recall value between

clusters and reference sets and *Fuzzy DEP* and the number of genes contained in their intersection are summarized in Table (4).

	Fuzzy Differential Expression Profiles							
	P1 (34)	P2 (51)	P3 (32)	P13 (6)	P14 (5)	P15 (4)	P20 (31)	P22 (1)
Cluster1 (45)			14	3				1
Cluster2 (64)	15	28				2	17	
Cluster3 (19)					3			

TAB. 4 – *Overlap analysis between the three functional clusters obtained with FCM and the 8 fuzzy DEPs. Between brackets is indicated the number of genes present in each set. For each well-matched pair, the number of genes of the intersection $C \cap R$ is reported in the corresponding cell. The threshold recall value is set here to 0.4 in order to get well-matched pairs for each reference set.*

Various methods exist for characterizing the biological relevance of signature genes obtained from high-throughput experimental results. One of them is the simple GO term enrichment analysis which allows to discover among all GO terms associated with all genes in a given cluster, statistically significant GO terms displaying low P_Value $< 10^{-4}$ or 10^{-5}. The P_value is calculated for a given gene list versus a background list (here all human genes) displaying GO annotation in the NCBI repository file (GEN), using the *hyper geometric test* (Eden et al., 2009). Only the BP annotations are considered here. Table (5) presents the results obtained with each well-matched pair. It can be seen that quite specific BP terms are assigned to each subset of genes delineated by the intersection $C \cap R$ of a functional cluster and an expression profile. In the case of Cluster_2 ∩ P2 and Cluster_2 ∩ P20, the same general GO term (*cell differentiation*) is found at the 1st position, but distinct GO terms appear at the 2nd and 3rd positions which correspond to biological processes which were mixed together in Cluster 2 but are now associated to two distinct expression profiles (P2 and P20). This example illustrates how our overlap analysis appears capable of extracting consistent subsets of genes with respect to biological function and transcriptional behavior.

Cluster_1 ∩ P3		Cluster_2 ∩ P1		Cluster_2 ∩ P2		Cluster_2 ∩ P20		Cluster_3 ∩ P14	
GO term	P_Value	GO term	P_Value	GO term	P_Value	GO term	P_Value	GO term	P_Value
regulation of transcription, DNA-dependent	9.95E-04	chromosome organization	9.55E-05	cell differentiation	7.35E-05	cell differentiation	5.97E-05	Water transport	2.08E-05
NADH oxidation	2.98E-04	strand break repair via homologous recombination	1.1012E-04	vascular endothelial growth factor receptor signaling pathway	1.06E-04	multicellular organismal development	9.58E-05		
		response to estrogen stimulus	9.6584E-04	angiogenesis	5.83E-04	insulin secretion	1.42E-04		

TAB. 5 – *GO term enrichment in the well-matched pairs. Only the top GO terms (with P_Value lower than 10^{-3}) characterizing the genes present in the intersection $C \cap R$ are displayed here.*

4 Conclusion and Perspectives

In this paper, we have tested our recently described semantic similarity measure *IntelliGO* in various clustering approaches. A collection of reference gene sets composed of selected KEGG human pathways has been used. Heatmap vizualisation of hierarchical clustering has provided visual evidence that the IntelliGO measure is more advantageous than other measures for clustering genes with respect to semantic similarity. Fuzzy C-means clustering was successfully optimized with F-score values reaching a maximum value of 0.62. A method for overlap analysis between clusters and reference sets has been described and implemented. It has been applied to a set of genes that are dysregulated in cancer using expression profiles as reference sets. It then allows to retrieve at the intersection of functional clusters and expression profiles, relevant subsets of genes that can be meaningfully characterized.

An important motivation of this work was to compare the performance of our *IntelliGO* similarity measure with other measures for clustering purposes. We have illustrated how the visualization with heatmaps of hierarchical clustering results may help to globally appreciate such performance. We intend to make our collections of reference sets of genes available on-line in a comparison tool complementary to the Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) tool (Pesquita et al., 2008). Users would download the datasets, produce their own similarity matrices using the measure to be tested and submit these matrices on-line for hierarchical clustering and heatmap generation.

The fuzzy C-means clustering belongs to overlapping clustering methods that attract more and more attention, because of their application to many domains. Recently, some overlapping variants of the K-means algorithms have been proposed (Cleuziou, 2008, 2010), namely *Okm, Okmed*, and *Wokmed*. These algorithms could now be tested with our *IntelliGO* measure and benchmarking collection of genes.

Optimizing fuzzy clustering remains challenging, especially in the absence of reference sets. In our application with cancer genes, we used expression profiles as reference sets. The influence of this choice on clustering results should be tested, an alternative solution being the clustering optimization without any reference sets (Ammor et al., 2008).

The overlap analysis method proposed here leads to a pairing of clusters and reference sets, which may be used for mismatch analysis. Indeed the genes present in a cluster but not in the corresponding reference set may be proposed as missing members of this reference set. Reciprocally, some genes from a reference set that are absent from the corresponding cluster may be enriched with features required for its grouping with other members of this cluster. Thus, the proposed overlap analysis may reveal a mean for discovering missing information.

Applied to a list of genes from a transcriptomic cancer study, our method also leads to identify subsets of genes displaying consistent expression and functional profiles. Promising results have been obtained using a simple GO term enrichment procedure. More sophisticated tools such as DAVID (Huang et al., 2009) and GSEA (Subramanian et al., 2005) tools, could be used to improve the biological interpretation of these subsets of genes.

References

The csbl.go package. `http://csbi.ltdk.helsinki.fi/anduril/`.

The NCBI gene2go file. `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz`.

Adryan, B. and R. Schuh (2004). Gene-Ontology-based clustering of gene expression data. *Bioinformatics 20*(16), 2851–2852.

Ammor, O., A. Lachkar, K. Slaoui, and N. Rais (2008). Optimal fuzzy clustering in overlapping clusters. *Int. Arab J. Inf. Technol. 5*(4), 402–408.

Benabderrahmane, S., M.-D. Devignes, M. Smaïl-Tabbone, A. Napoli, O. Poch, N.-H. Nguyen, and W. Raffelsberger (2009). Analyse de données transcriptomiques: Modélisation floue de profils dexpression différentielle et analyse fonctionnelle. In *INFORSID*, pp. 413–428.

Benabderrahmane, S., M. Smail-Tabbone, O. Poch, A. Napoli, and M.-D. Devignes (2010). Intelligo: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics 11*(1), 588.

Blanchard, E., M. Harzallah, and P. Kuntz (2008). A generic framework for comparing semantic similarities on a subsumption hierarchy. In *18th European Conference on Artificial Intelligence (ECAI)*, pp. 20–24.

Brameier, M. and C. Wiuf (2007). Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps. *J. of Biomedical Informatics 40*(2), 160–173.

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *ICPR*, pp. 1–4. IEEE.

Cleuziou, G. (2010). Two variants of the okm for overlapping clustering. In F. Guillet, G. Ritschard, D. Zighed, and H. Briand (Eds.), *Advances in Knowledge Discovery and Management*, Volume 292 of *Studies in Computational Intelligence*, pp. 149–166. Springer Berlin / Heidelberg. 10.1007/978-3-642-00580-0_9.

Consortium, T. G. O. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucl. Acids Res. 38*(suppl-1), D331–335.

Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini (2009). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics 10*(1), 48.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America 95*(25), 14863–14868.

Ganesan, P., H. Garcia-Molina, and J. Widom (2003). Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst. 21*(1), 64–93.

Gasch, A. and M. Eisen (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology 3*(11), research0059.1–research0059.22.

Huang, D. W. a. . W., B. T. Sherman, and R. A. Lempicki (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols 4*(1), 44–57.

Jiang, J. J. and D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pp. 9008+.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids*

research 38(Database issue), D355–360.

Khatri, P. and S. Draghici (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics 21*(18), 3587–3595.

Lord, P. W., R. D. Stevens, A. Brass, and C. A. Goble (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics 19*(10), 1275–1283.

Macqueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Nagar, A. and H. Al-Mubaid (2008a). A new path length measure based on go for gene similarity with evaluation using sgd pathways. In *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS 08)*, Washington, DC, USA, pp. 590–595. IEEE Computer Society.

Nagar, A. and H. Al-Mubaid (2008b). Using path length measure for gene clustering based on similarity of annotation terms. In *Computers and Communications, 2008. ISCC 2008. IEEE Symposium on*, pp. 637–642.

Pesquita, C., D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto (2008). Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics 9*(Suppl 5), S4.

Pesquita, C., D. Faria, A. O. Falcao, P. Lord, and F. M. Couto (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol 5*(7), e1000443.

Polettini, N. (2004). The vector space model in information retrieval- term weighting problem.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pp. 448–453.

Rogers, M. F. and A. Ben-Hur (2009). The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics 25*(9), 1173–1177.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math. 20*(1), 53–65.

Salton, G. and M. J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Speer, N., H. Frohlich, C. Spieth, and A. Zell (2005). Functional grouping of genes using spectral clustering and gene ontology. In *In Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 298–303. IEEE Press.

Speer, N., C. Spieth, and A. Zell (2004). A memetic co-clustering algorithm for gene expression profiles and biological annotation.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*(43), 15545–15550.

Theodoridis, S. and K. Koutroumbas (2006). *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

Wang, J. Z., Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics 23*(10), 1274–1281.

Résumé

Les algorithmes de classification (*Clustering*) reposent sur des mesures de similarité ou de distance qui dirigent le regroupement des objets similaires dans un même groupe et la séparation des objets différents entre des groupes distincts. Notre nouvelle mesure de similarité sémantique (IntelliGO), récemment décrite, qui s'applique à la comparaison fonctionnelle des gènes, est testée ici dans un processus de clustering. L'ensemble de test est composé des gènes contenus dans une collection de classes de référence (*Pathways KEGG*). La visualisation du clustering hiérarchique avec des cartes de densité (*heatmaps*) illustre les avantages de l'utilisation de la mesure IntelliGO, par rapport à trois autres mesures de similarité. Comme les gènes peuvent souvent appartenir à plus d'un cluster fonctionnel, la méthode C-means floue est également appliquée à l'ensemble des gènes de la collection. Le choix du nombre optimal de clusters et la performance du clustering sont évalués par la méthode F-score en utilisant les classes de référence. Une analyse de recouvrement entre clusters et classes de référence est proposée pour faciliter des analyses ultérieures. Enfin, notre méthode est appliquée à une liste de gènes dérégulés, concernant le cancer colorectal. Dans ce cas, les classes de référence sont les profils d'expression de ces gènes. L'analyse de recouvrement entre ces profils et les clusters fonctionnels obtenus avec la méthode C-means floue conduit à caractériser des sous-ensembles de gènes partageant à la fois des fonctions biologiques communes et un comportement transcriptionel identique.

Un modèle génératif pour la comparaison de métriques en classification de profils d'expression de gènes

Alpha Diallo[*,**], Ahlame Douzal-Chouakria[*] et Françoise Giroud[**]

[*]LIG AMA, Université Joseph Fourier
BP 53 - 38041 Grenoble cedex 9
(alpha.diallo, ahlame.douzal)@imag.fr
[**]TIMC-IMAG RFMQ (CNRS-UMR 5525), Université Joseph Fourier
F-38706 La Tronche Cedex, France
francoise.giroud@imag.fr

Résumé. La prolifération cellulaire traduit le processus cyclique de division des cellules responsable de la croissance des tissus. Un tissu tumoral peut alors être caractérisé par la présence de cellules cancéreuses qui présentent une prolifération anormale. Ce papier s'intéresse à la classification des gènes différentiellement activés au cours du processus de division cellulaire. Les gènes étudiés ici sont décrits par leurs profiles d'expression (c-à-d des données temporelles d'expression de gènes) tout au long de 3 cycles cellulaires successifs. Le travail présenté concerne l'évaluation de l'efficacité de 4 métriques majeures pour la classification et le classement des profiles d'expression de gènes. Il est basé sur la mise en œuvre d'un modèle périodique aléatoire pour la simulation de gènes d'expression cyclique au long du cycle de division cellulaire. Le modèle traduit les événements cycliques de régulation moléculaire du cycle cellulaire, avec des variations en décalage dans le temps des profiles d'expression de différents gènes. Le modèle rend compte également de phénomènes de désynchronisation cellulaire provoquant à la fois l'atténuation en amplitude des valeurs d'expression et des modifications de périodicité des cycles cellulaires successifs.

1 Introduction

Toutes les cellules de notre corps contiennent les mêmes gènes, mais tous n'interviennent pas dans chaque cellule : les gènes sont activés ou exprimés au besoin. De tels gènes spécifiques définissent le modèle moléculaire lié à une fonction spécifique d'une cellule et apparaissent dans la plupart des cas comme organisés dans des réseaux de régulation moléculaire. Pour comprendre comment les cellules réalisent une telle spécialisation, il est nécessaire d'identifier quels gènes s'expriment dans chaque type de cellules (par exemple, des tissus cancéreux versus des tissus sains). La technologie des puces à ADN nous permet d'étudier simultanément les niveaux d'expression de plusieurs milliers de gènes, au cours de processus biologiques importants, pour déterminer ceux qui sont exprimés dans un type de cellule spécifique (Eisen et Brown, 1999). Les techniques de classification et de classement sont utilisées et se

sont montrées particulièrement efficaces pour comprendre la fonction des gènes, des voies de régulation et des processus cellulaires (e.g., Liu et al. (2008), Park et al. (2008), Scrucca (2007)). Nous distinguons au moins deux principales approches de classification et de classement de profils ou de séries temporelles. D'une part, les approches paramétriques consistant à projeter les séries temporelles dans des espaces de fonctions correspondant, par exemple, aux polynômes d'un modèle ARIMA, aux transformées de Fourier, ou plus généralement aux paramètres d'un modèle approximant les séries temporelles. Des mesures conventionnelles peuvent ensuite être utilisées dans le nouvel espace de projection (e.g.,Bar-Joseph et al. (2003), Caiado et al. (2006), Garcia-Escudero et Gordaliza (2005)). D'autre part, on distingue les approches non-paramétriques dont l'objectif est la proposition de nouvelles mesures de proximités définies dans l'espace de description initial et intégrant la dimension temporelle des données (e.g.,Anagnostopoulos et al. (2006), Heckman et Zamar (2000), Keller et Wittfeld (2004)). Dans le cadre des approches non-paramétriques, nous proposons d'étudier l'efficacité de quatre métriques majeures pour la classification et le classement des profils temporels d'expression de gènes. Cette étude est basée sur la mise en œuvre d'un modèle périodique aléatoire pour la simulation de gènes d'expression cyclique. Ce modèle tient compte des caractérisques principalement observées sur les profils de gènes du cycle cellulaire : l'amplitude initiale du profil, la période du profil, l'atténuation des amplitudes dans la longueur du temps et les effets de tendance. La suite de l'article est organisé en quatre sections. La section suivante définit ce que sont les données d'expression de gènes et présente le problème biologique abordé. La section 3 présente les quatre principales métriques à évaluer et discute de leurs caractéristiques. La section 4 indique comment les mesures seront comparées par la classification et le classement des profils de gènes. Enfin, l'ensemble des méthodes d'évaluation basées sur le modèle utilisé et la discution des résultats obtenus sont présentés dans la section 5.

2 Identification des gènes exprimés au cours du cycle cellulaire

Le problème biologique d'intêret est l'analyse de la progression de l'expression des gènes durant le processus de la division cellulaire. La division cellulaire est le processus principal assurant la prolifération des cellules , et se décompose en quatre phases principales (G_1, S, G_2 et M) et trois transitions de phase (G_1/S, G_2/M et M/G_1). Le processus de division commence à la phase G_1 pendant laquelle la cellule se prépare à la synthèse de l'ADN. Vient la phase S où l'ADN est dupliqué (c-à-d chaque chromosome est dupliqué), suivie par la phase G_2 pendant laquelle la cellule se prépare à sa division. Enfin vient la mitose M où la cellule se divise en deux cellules filles. Pendant ces quatre phases, certains gènes sont actifs (fortement exprimés) à des périodes spécifiques, d'autres pas. Un des objectifs consiste à identifier les gènes fortement exprimés et caractérisant chaque phase du cycle cellulaire. Ceci fournit des informations importantes, par exemple, pour comprendre comment le traitement hormonal peut induire la prolifération cellulaire par l'activation de gènes spécifiques. Afin d'accroître la compréhension de l'expression des gènes au cours du processus de la division cellulaire, des molécules d'ADN représentant les différents gènes sont placées sur des spots discrets régulièrement répartis en une matrice ligne/colonne (appelée puce à ADN). En déposant sur ces puces à ADN des extraits cellulaires on peut mesurer le niveau d'expression de chaque gène

au sein des populations cellulaires étudiées. En échantillonnant au cours du temps une population cellulaire initialement synchronisée, chaque gène étudié peut être décrit par son profil d'expression observé au cours du temps sur un ou plusieurs cycles de la division cellulaire.

3 Proximité entre profils d'expression de gènes

La classification et le classement des données d'expression implique le plus souvent la distance euclidienne ou le coefficient de corrélation de Pearson. Cette section introduit quatre métriques majeures pour l'analyse des gènes et leurs spécifications, en tenant compte des proximités en valeurs et en forme des profils de gènes. Soit $g_1 = (u_1, ..., u_p)$ et $g_2 = (v_1, ..., v_p)$ les niveaux d'expression de deux gènes observés aux instants $(t_1, ..., t_p)$.

3.1 La distance euclidienne

La distance euclidienne δ_E entre g_1 et g_2 est définie par :

$$\delta_E(g_1, g_2) \quad = \quad \left(\sum_{i=1}^{p} (u_i - v_i)^2 \right)^{\frac{1}{2}}. \tag{1}$$

Il ressort de cette définition, que la proximité entre deux gènes dépend de la proximité des valeurs de leurs niveaux d'expression, sans tenir compte de la forme de leurs profils. En d'autres termes, la distance euclidienne ignore la dépendance temporelle des données.

3.2 Le coefficient de corrélation de Pearson

De nombreux travaux utilisent le coefficient de corrélation de Pearson comme mesure de proximité en forme entre 2 séries temporelles. Sans perte de généralité, supposons que les valeurs de g_1 et g_2 évoluent dans $[0, N]$. Les gènes g_1 et g_2 sont dits de formes similaires si à chaque période d'observation $[t_i, t_{i+1}]$, ils croient ou décroient simultanément (monotonicité), avec un taux d'accroissement égal. En revanche, g_1 et g_2 sont de formes opposées si dans chaque période d'observation $[t_i, t_{i+1}]$ où g_1 croit , g_2 décroit et vice-versa avec le même taux d'accroissement en valeur absolue. Afin d'illustrer la limite du coefficient de corrélation à mesurer la proximité en forme des gènes, considérons son expression basée sur les différences entre les valeurs observées :

$$\textsc{Cor}(g_1, g_2) = \frac{\sum_{i,i'} (u_i - u_{i'})(v_i - v_{i'})}{\sqrt{\sum_{i,i'} (u_i - u_{i'})^2} \sqrt{\sum_{i,i'} (v_i - v_{i'})^2}}. \tag{2}$$

En impliquant les différences entre tous les couples d'observations (c-à-d, observées à tous les couples d'instants (i, i')), le coefficient de corrélation de Pearson fait l'hypothèse d'indépendance entre les données observées. Conséquence, ce coefficient peut surestimer la proximité en forme. Par exemple, la section 4 illustre le cas de données dotées d'un effet de tendance où deux gènes de formes opposées peuvent avoir un coefficient de corrélation de valeur positive forte.

3.3 Le coefficient de corrélation temporelle

Pour surmonter les limites du coefficient de corrélation de Pearson (Eq. (2)), le coefficient de corrélation temporelle est utilisé. Il réduit le coefficient de corrélation de Pearson aux différences de premier ordre :

$$\text{CORT}(g_1, g_2) = \frac{\sum_i (u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_i (u_{(i+1)} - u_i)^2}\sqrt{\sum_i (v_{(i+1)} - v_i)^2}}. \quad (3)$$

avec $\text{CORT}(g_1, g_2) \in [-1, 1]$. La valeur $\text{CORT}(g_1, g_2) = 1$ indique que g_1 et g_2 présentent une forme similaire. La valeur $\text{CORT}(g_1, g_2) = -1$ signifie que g_1 et g_2 sont de formes opposées. Enfin, $\text{CORT}(g_1, g_2) = 0$ exprime que les taux d'accroissement de g_1 et g_2 sont stochastiquement linéairement indépendants, identifiant ainsi des gènes de formes différentes (non similaires ni opposées).

3.4 Mesures de proximité alliant forme et valeurs

Pour une mesure de proximité couvrant simultanément les écarts en forme et en valeurs des niveaux d'expression, l'indice de dissimilarité D_k proposé dans Douzal-Chouakria et al. (2010, 2009) est considéré :

$$D_k(g_1, g_2) = f(\text{CORT}(g_1, g_2))\ \delta_E(g_1, g_2), \quad (4)$$

$$f(x) = \frac{2}{1 + exp(k\ x)},\ \ k \geq 0. \quad (5)$$

Cet indice couvre à la fois la distance euclidienne (Eq. (1)) pour la proximité en valeurs, et la corrélation temporelle (Eq. (3)) pour la proximité en forme. Il est basé sur une fonction de reglage $f(x)$ qui module la proximité en valeurs en fonction de la proximité en forme. Une fonction exponentielle $f(x)$ est préférable à une forme linéaire afin d'assurer un effet de réglage approximativement égal pour les valeurs extrêmes (i.e., CORT=-1, +1 et 0) et leurs plus proches voisins. La figure 1 montre l'effet de réglage pour plusieurs valeurs du paramètre $k \geq 0$. Dans le cas de gènes de formes différentes (i.e., CORT voisin de 0), $f(x)$ est voisin

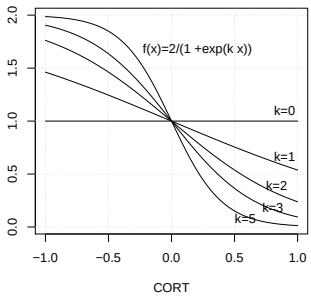


FIG. 1 – *L'effet du réglage en fonction de* k

de 1 quelles que soient les valeurs de k, et D_k est approximativemnt égal à δ_E. En revanche,

si CORT $\neq 0$ (forme non différentes) le paramètre k module les contributions des deux types de proximité (valeurs et forme) à l'indice de dissimilarité D_k. Lorsque k augmente, la contribution de la proximité en forme $1 - 2/(1 + exp(k\ |\text{CORT}|))$ augmente, tandis que celle de la proximité en valeurs $2/(1 + exp(k\ |\text{CORT}|))$ diminue. Par exemple, pour $k = 0$ et $|\text{CORT}| = 1$ (forme similaire ou opposée), la proximité en forme contribue 0% à D_k tandis que la proximité en valeurs contribue 100% à D_k (la valeur de D_k est totalement déterminée par δ_E). Pour $k = 2$ et $|\text{CORT}| = 1$, la proximité en forme contribue 76.2% à D_k tandis que celle en valeurs contribue 23.8% à D_k (23.8% de la valeur de D_k sont déterminés par δ_E et les 76.2% restants par CORT).
Notons que la dynamique time warping (e.g.,Kruskall et Liberman (1983), Shieh et Keogh (2008)), qui est largement utilisée, n'est pas abordée dans ce travail puisqu'elle n'est pas appropriée pour analyser des profils d'expressions de gènes au cours du cycle cellulaire. En effet, l'identification de gènes exprimés au cours cycle cellulaire repose principalement sur l'instant où les gènes sont fortement exprimés. Ainsi, les instants d'observation ne doivent pas subir de décalage lors de l'évaluation des proximités entre profils d'expression de gènes.

4 Comparison des métriques

Une étude de simulation est effectuée pour évaluer l'efficacité des métriques définies dans les équations (1) à (4). Pour la procédure de classification, nous proposons d'utiliser l'algorithme PAM (Partitioning Around Medoïds) afin de partitionner les gènes simulés en n classes (n étant le nombre de phases du cycle cellulaire ou de transitions de phases étudiées). L'algorithme PAM est préféré à l'approche classique des K-means pour plusieurs raisons. Il est plus robuste aux valeurs aberrantes qui sont nombreuses dans les données d'expression de gènes. PAM permet une analyse détaillée de la partition en fournissant des indices permettant d'apprécier la qualité des classes ainsi que celle des gènes. En effet, PAM mesure la *silhouette width* (sw) de chaque gène, qui est un indicateur de confiance quant à l'appartenance d'un gène à une classe. Pour plus de détails sur l'algorithme PAM voir Kaufman et Rousseeuw (1990). L'efficacité de chaque métrique est basée sur trois critères : *la silhouette width* moyenne d'une partition notée asw, le ratio standard $wbr = \frac{intra}{inter}$ et l'indice de Rand corrigé (RI). Pour la procédure de classement des gènes, l'algorithme 10-NN est utilisé, et les taux d'erreur de gènes mal classés sont retenus pour apprécier l'efficacité de chaque métrique.

5 Etude comparative

5.1 Modèle génératif de profils d'expression périodiques

Nous utilisons des profils simulés, générés sur la base du modèle de regression non-linéaire proposé par Liu et al. (2004). Ce modèle permet de simuler l'atténuation des amplitudes de l'expression des gènes liée aux variations stochastiques au cours des différentes phases du cycle cellulaire. La fonction sinusoïdale caractérisant la périodicité du profils d'expression d'un gène g au cours de différents cycles cellulaires est :

$$f(t, \theta_g) \quad = \quad a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} cos(\frac{2\pi t}{Texp(\sigma z)} + \Phi_g) exp(-\frac{z^2}{2}) dz. \qquad (6)$$

où $\theta_g = (K_g, T, \sigma, \Phi_g, a_g, b_g)$ est spécifique du gène g. Le paramètre K_g représente son amplitude initiale, T est la durée du cycle cellulaire. Le paramètre σ contrôle le taux d'atténuation des amplitudes au cours des différents cycles, Φ_g correspond à la phase du cycle cellulaire où le gène est le plus exprimé. Les paramètres a_g et b_g (l'ordonnée à l'origine et la pente, respectivement) contrôlent les tendances des profils. La figure 2 illustre la progression des expressions de gènes au cours des 5 phases et transitions de phase G_1/S, S, G_2, G_2/M et M/G_1. Nous allons utiliser le terme phase de manière générique pour parler de "phase" et "transition de phase" dans tout ce qui suit.

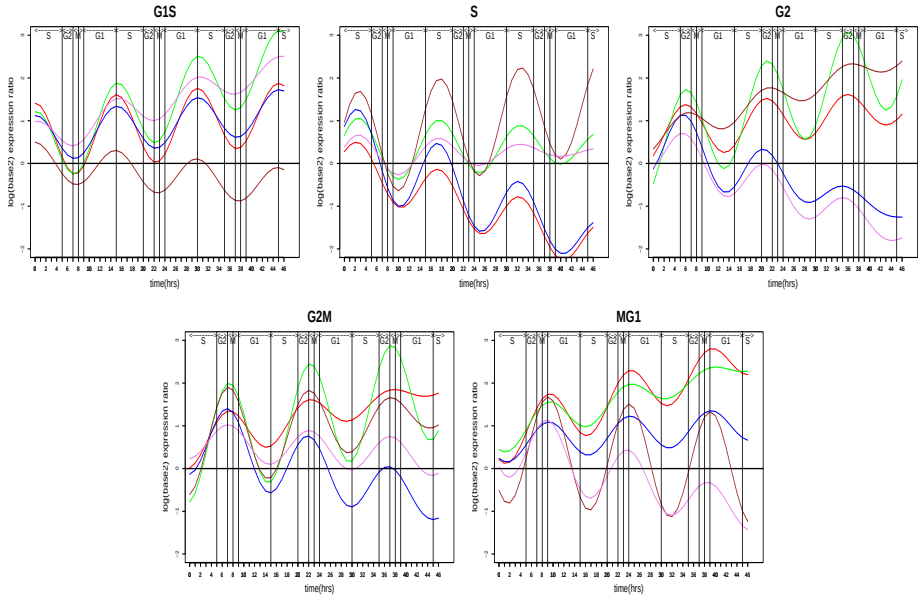


FIG. 2 – *Progression de l'expression des gènes durant les* 5 *phases* G_1/S, S, G_2, G_2/M *et* M/G_1.

5.2 Protocole de simulation

Sur la base de ce modèle et des valeurs de paramètres spécifiées dans Liu et al. (2004), quatre expériences sont menées pour étudier la façon dont chaque métrique considère les variations d'expression de gènes. La première expérience génère des profils avec une variation observée uniquement au niveau de l'amplitude initiale K_g variant dans $[0.34, 1.33]$. La seconde expérience inclut une atténuation des amplitudes σ évoluant dans $[0.054, 0.115]$. La troisième expérience inclut les effets de tendance $b_g \in [-0.05, 0.05]$ et $a_g \in [0, 0.8]$ et enlève les effets de σ. Enfin la quatrième expérience simule des profils avec une variation simultanée des paramètres K_g, σ, a_g, b_g dans les mêmes intervalles que précédemment. La valeur d'un paramètre est prise de manière aléatoire dans l'intervalle auquel il appartient. L'évolution des profils est suivi sur 3 cycles cellulaires, T est fixé à 15 heures pour toutes les simulations et Φ_g prend les valeurs 0, 5.190, 3.823, 3.278 et 2.459 pour la génération respective des 5 phases G_1/S, S,

G_2, G_2/M et M/G_1. La figure 3 montre les variations produites dans les quatre expériences pour les gènes exprimés dans la phase G_1/S. La spécification des paramètres du modèle des quatre expériences est résumée dans le tableau 1. Pour chaque expérience $j \in \{1, ..., 4\}$, 10 échantillons S_{ij} $i \in \{1, ..., 10\}$ sont simulés. Chaque échantillon est composé de 500 profils d'expression (de longueur 47) de gènes avec 100 gènes pour chacune des 5 phases G_1/S, S, G_2, G_2/M et M/G_1. La comparaison est effectuée pour chaque expérience sur 5000 gènes simulés (c'est-à-dire 10 échantillons de 500 gènes chacune)

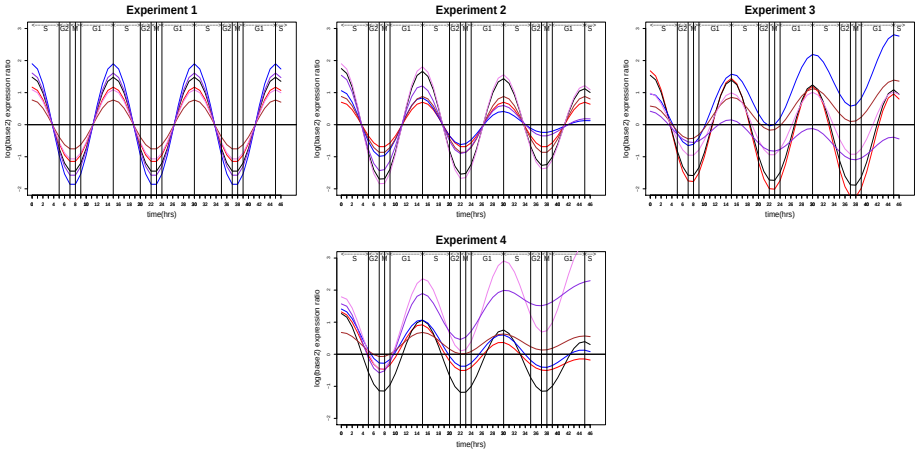


FIG. 3 – *Profils des gènes de la phase G_1/S suivant les quatre expériences.*

Experiment number	K_g	σ	b_g	a_g
1	[0.34, 1.33]	0	0	0
2	[0.34, 1.33]	[0, 0.115]	0	0
3	[0.34, 1.33]	0	[-0.05, 0.05]	[0, 0.8]
4	[0.34, 1.33]	[0, 0.115]	[-0.05, 0.05]	[0, 0.8]

TAB. 1 – *Spécification des paramètres du modèle.*

5.3 Evaluation de l'efficacité des métriques pour la classification des gènes

Pour chaque expérience et pour chaque métrique δ_E (Eq. (1)), COR (Eq. (2)), et CORT (Eq. (3)), nous partitionnons l'ensemble des profils de chaque échantillon S_{ij} en 5 classes (correspondant aux 5 phases). Par exemple, pour l'expérience j et la métrique δ_E, l'algorithme PAM est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ afin d'extraire les 10 partitions $\mathcal{P}_{\delta_E}^{1j}$,..., $\mathcal{P}_{\delta_E}^{10j}$. Pour chaque partition $\mathcal{P}_{\delta_E}^{ij}$, les valeurs des trois critères asw, wbr, RI sont retenues. Ainsi, l'évaluation de l'efficacité de la métrique δ_E par rapport à l'expérience j est réalisée en considérant les valeurs moyennes des critères asw, RI et wbr sur les 10 partitions $\mathcal{P}_{\delta_E}^{1j}$,..., $\mathcal{P}_{\delta_E}^{10j}$. Une classification adaptative est appliquée pour l'indice de dissimilarité D_k (Eq. (4)). Elle consiste, sur un échantillon S_{ij}, à exécuter l'algorithme PAM pour k allant de 0 à 6 (avec un

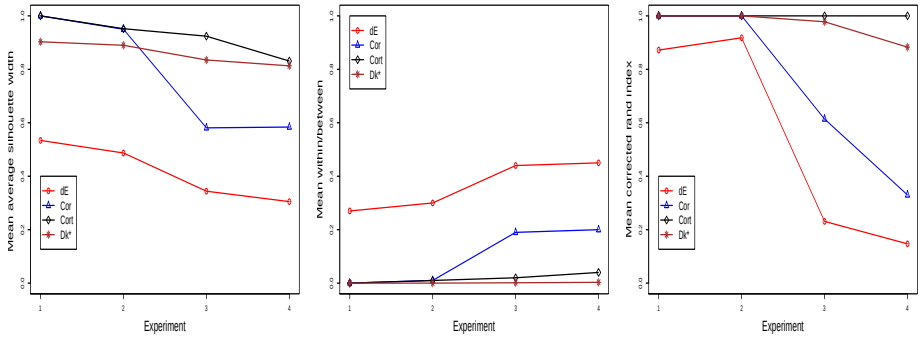


FIG. 4 – *Evaluation des métriques pour la classification des profils d'expression simulés. La progression des valeurs moyennes des critères asw (gauche), wbr (milieu) et RI (droite) est illustrée.*

pas égal à 0.01). Ceci permet d'apprendre la valeur k^* qui fournit la partition optimale $\mathcal{P}^{ij}_{D_{k*}}$ selon les critères asw et wbr. Notons que k^* fournit la meilleure contribution des proximités en valeurs et en forme à D_{k^*}. Par conséquent, D_{k*} appris est considéré comme le meilleur pour le partitionnement de S_{ij}. Le tableau 2 donne pour chaque expérience, la moyenne et la variance $(\overline{k*}, var(k*))$ de $k*$. Comme dans le cas des métriques δ_E, COR et CORT, l'évaluation de l'efficacité de la métrique D_k par rapport à l'expérience j, est résumée par les valeurs moyennes asw, RI et wbr sur les 10 partitions $\mathcal{P}^{1j}_{D_{k*}}$,..., $\mathcal{P}^{10j}_{D_{k*}}$. La figure 4 montre (pour chaque métrique) la progression des valeurs moyennes des critères asw (gauche), wbr (milieu) et RI (droite) suivant les quatre expériences.

Adaptive	Exp1	Exp2	Exp3	Exp4
Clustering	(6,0)	(6,0)	(6,0)	(5.85,0.06)
Classification	(3,3.53)	(3,3.53)	(4.55,1.18)	(4.84,0.98)

TAB. 2 – *k*(moyenne, variance)*

5.4 Evaluation de l'efficacité des métriques pour le classement des gènes

Pour chaque expérience et pour chaque métrique δ_E, COR et CORT, nous éxécutons l'algorithme 10-NN, pour chaque échantillon S_{ij}. Par exemple, pour l'expérience j et la métrique δ_E, l'algorithme 10-NN est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ pour générer les 10 classes $\mathcal{C}^{1j}_{\delta_E}$,..., $\mathcal{C}^{10j}_{\delta_E}$. Pour chaque classe $\mathcal{C}^{ij}_{\delta_E}$, le taux de profils de gènes mal classifiés est retenu. L'évaluation de la métrique δ_E dans l'expérience j est résumée par le taux d'erreur moyen sur les 10 classes $\mathcal{C}^{1j}_{\delta_E}$,..., $\mathcal{C}^{10j}_{\delta_E}$. Pour l'indice de dissimilarité D_k, un classement adaptatif est appliqué. Il consiste à exécuter l'algorithme 10-NN sur l'échantillon S_{ij} avec des valeurs de k allant de 0 à 6 (avec un pas égal à 0.01). Ceci permet d'estimer la valeur k^* minimisant le taux d'erreur de profils mal classifiés de la classe $C^{ij}_{D_k}$. L'évaluation de la métrique D_k, pour le classement des profils de gènes dans l'expérience j, se résume par le taux d'erreur moyen

calculé sur les 10 classes $\mathcal{C}_{D_k}^{1j}$,..., $\mathcal{C}_{D_k}^{10j}$. La figure 5 montre la progression suivant les quatres expériences du taux d'erreur moyen lié à chaque métrique.

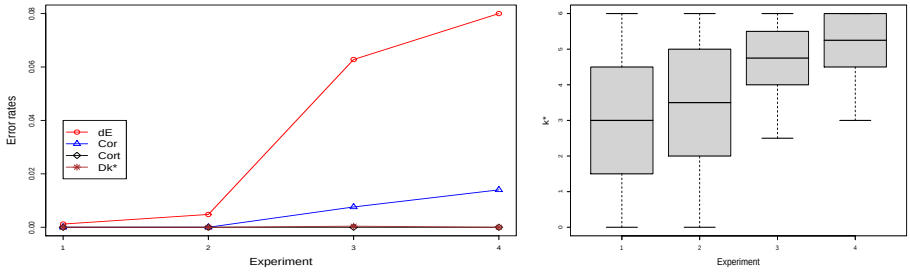


FIG. 5 – *Progression des taux d'erreur moyens des classements suivant les expériences (gauche). Distribution de $k*$ pour les classements adaptatifs suivant les expériences (droite)*

5.5 Discussion

Nous discutons, d'abord, sur l'intérêt de considérer un modèle génératif pour l'évaluation de la classification ou du classement des gènes exprimés au cours du cycle cellulaire. Le modèle périodique aléatoire est d'une grande importance en biologie : il permet de simuler des trajectoires périodiques d'expressions de gènes semblables à celles observées expérimentalement. Ces trajectoires peuvent varier considérablement en forme d'un gène à un autre dans une même classe (i.e. pour les gènes exprimés dans la même phase du cycle cellulaire). Ce modèle permet aussi de simuler des variations observées expérimentalement : principalement, la reproduction des variations sur l'atténuation en amplitude des valeurs d'expression et sur la tendance des trajectoires par rapport à des modifications de périodicité des cycles cellulaires. Le modèle périodique aléatoire peut être utile pour étudier séparément et avec précision les effets de chaque condition expérimentale (i.e. variation) sur l'efficacité de métriques, de résultats de classification ou de classement. Il existe aussi d'autres modèles dans la littérature pour évaluer des métriques, des résultats de classification ou de classement de gènes exprimés au cours du cycle cellulaire. On peut distinguer au moins deux principales techniques d'évaluation. D'une part, les données d'expression sont simulées à partir de modèles paramétriques, incluant : les modèles autorégressifs (Ramoni et al., 2002)), B-splines (Luan et Li, 2003), la décomposition en valeurs singulières (Alter et al. (2000), Holter et al. (2001)), ou les approches des moindres carrés partiels (Johansson et al., 2003). Ces modèles fournissent une estimation assez bonne des trajectoires exprimées. Cependant, les paramètres estimés ne sont pas biologiquement interprétables et fournissent des périodicités anormales sur toute la durée du cycle cellulaire. D'autre part, les données d'expression sont extraites à partir de données réelles (e.g.,Spellman et al. (1998), Whitfield et al. (2002)) où les techniques d'évaluation sont a priori basées sur un petit ensemble de gènes connus dont les trajectoires sont cycliques (dits gènes de référence), et qui sont supposés caractéristiques des phases du cycle cellulaire. Ici, les évaluations sont fortement dépendantes des gènes de référence choisis, et la littérature ne fournit pas de consensus clair entre les biologistes sur l'ensemble approprié de gènes de référence à tenir en considération. Pour les raisons citées ci-dessus, nous avons opté pour l'utilisation

d'un modèle génératif pour la simulation des données d'expression. Nous avons évalué des métriques avec des algorithmes classiques de classification (Kmeans, PAM, méthodes hiérarchiques, etc.) ou de classement (KNN, arbres de décisions, etc.). Ces métriques ont été évaluées par des critères de qualité : indice de Rand corrigé, taux de mauvais classement, etc.
Examinons d'abord les résultats de la classification. Notons quelques informations supplémentaires sur les critères en question. La valeur asw indique une forte structure (asw proche de 1) ou une faible structure ($asw < 0.5$) de classes. Le critère wbr mesure la compacité (variabilité au sein d'une classe) et la séparabilité (variabilité entre les classes) des classes. Une bonne partition est caractérisée par une faible valeur de wbr. Enfin, l'indice de Rand corrigé (RI) permet de comparer deux partitions. Une valeur $RI = 0$ correspond à une absence totale de correspondance entre la vraie partition et celle obtenue, alors qu'une valeur $RI = 1$ traduit une correspondance parfaite. La figure 4 montre que la classification basée sur δ_E donne, pour les expériences 1 à 4, les partitions les plus faibles comparée à celles fondées sur COR, CORT, ou D_k. En effet, les partitions fondées sur δ_E ont les plus faibles valeurs des critères asw et RI, et les valeurs les plus élevées pour wbr. En plus, les valeurs moyennes des critères asw, wbr et RI de la classification baseé sur δ_E se dégradent (diminution des asw et RI et augmentation de wbr) de l'expérience 1 à 4, montrant l'inadéquation de la distance euclidienne face aux variations complexes des profils de gènes cyliques. La classification basée sur COR donne, pour les expériences 1 et 2, de bonnes structures de partitions avec de très bonnes valeurs des critères asw, wbr et RI. Toutefois, cette qualité diminue de façon drastique dans les expériences 3 et 4 (Figure 4). Comme expliqué dans la section 3, ces résultats affirment la limite du coefficient de corrélation de $Pearson$ face aux variations de tendance. Enfin, les meilleures classifications et les plus fortes structures de partitions sont produites par CORT et D_k sur toutes les quatre expériences, avec une asw variant dans $[0.8, 1]$, une valeur wbr autour de 0, RI dans $[0.83, 1]$. Notons que la qualité de la classification basée sur D_k est légèrement inférieure à celle qui est fondée sur CORT, révélant que les profils d'expression de gènes sont naturellement plus différenciés par leur forme que par leurs valeurs. Cette hypothèse est soutenue par les fortes valeurs de $k*$ (proche de 6, avec une variabilité de 0) obtenues dans la classification adaptative pour les quatre expériences (Tableau 2).
Considérons les résultats du classement, la figure 5 (gauche) montre que, pour les expériences 1 et 2, les quatre métriques sont toutes aussi efficaces, avec des taux d'erreurs de classement autour 0. Toutefois, pour les expériences 3 et 4, nous notons une forte augmentation du taux d'erreur pour les classements basés sur δ_E, une légère augmentation du taux d'erreur pour les classements fondés sur COR, une augmentation négligeable pour D_k. Le tableau 2 et la figure 5 (droite) illustrent la distribution des valeurs de $k*$ dans les classements adaptatifs. Pour les expériences 1 et 2, nous notons une distribution uniforme de $k*$ dans $[0, 6]$. Ce cas se présente lorsque un bon classement peut être obtenu avec une métrique fondée sur des valeurs ($k*$ proche de 0) et avec une métrique basée sur la forme ($k*$ proche de 6). En effet, dans les deux premières expériences, la figure 5 (gauche) montre que les quatre métriques sont toutes aussi efficaces pour le classement des gènes avec des taux d'erreur négligeables. Pour les expériences 3 et 4, $k*$ prend des valeurs plus élevées indiquant que les mesures fondées sur la forme (c-à-d CORT et D_k) sont les plus efficaces pour le classement des profils d'expression de gènes, avec de très faibles taux d'erreur (figure 5 (gauche)). Enfin, selon les résultats des quatre expériences, les mesures CORT et D_k peuvent être considérées comme les plus efficaces pour le classement des profils d'expression de gènes.

6 Conclusion

En conclusion, pour la classification ou le classement des gènes exprimés au cours du cycle cellulaire, il est souhaitable de considérer la corrélation temporelle comme mesure de proximité. Toutefois, notons l'efficacité de la dissimilarité apprise D_k, qui fournit également une bonne classification et classement des gènes. La dissimilarité proposée D_k est particulièrement recommandée lorsque les instants d'observation ne doivent pas subir de décalage lors de l'évaluation des proximités (ce qui est le cas des profils d'expression de gènes du cycle cellulaire). Notons que le dissimilarité D_k généralise les métriques conventionnelles ; elle correspond à la corrélation temporelle pour k^* voisin de 6, à la distance euclidienne pour k^* voisin de 0 et plus généralement à une métrique couvrant à la fois des proximités portant sur les valeurs et sur les formes.

Références

Alter, O., P. Brown, et D. Bostein (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. 97*, 10101–10106.

Anagnostopoulos, A., M. Vlachos, M. Hadjieleftheriou, E. Keogh, et P. Yu (2006). Global distance-based segmentation of trajectories. *In Proc. of ACM SIGKDD*, 34–43.

Bar-Joseph, Z., G. Gerber, D. Gifford, T. Jaakkola, et I. Simon (2003). Continuous representations of time-series gene expression data. *Journal of Computational Biology 10*, 341–356.

Caiado, J., N. Crato, et D. Pena (2006). A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis 50*, 2668–2684.

Douzal-Chouakria, A., A. Diallo, et F. Giroud (2009). Adaptive clustering for time series: application for identifying cell cycle expressed genes. *Computational Statistics and Data Analysis 53*, 1414–1426. Elsevier.

Douzal-Chouakria, A., A. Diallo, et F. Giroud (2010). A random-periods model for the comparison of a metrics efficiency to classify cell-cycle expressed genes. *Pattern Recognition Letters 31*, 1601–1617. Elsevier.

Eisen, M. et P. Brown (1999). Dna arrays for analysis of gene expression. *Methods Enzymol 303*, 179–205.

Garcia-Escudero, L. A. et A. Gordaliza (2005). A proposal for robust curve clustering. *Journal of Classification 22*, 185–201.

Heckman, N. E. et R. Zamar (2000). Comparing the shapes of regression functions. *Biometrika 22*, 135–144.

Holter, N., A. Maritan, M. Cieplak, N. Fedoroff, et J. Banavar (2001). Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America 98*, 1693–1698.

Johansson, D., P. Lindgren, et A. Berglund (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics 19*, 467–473.

Kaufman, L. et P. Rousseeuw (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. New York: John Wiley and Sons.

Keller, K. et K. Wittfeld (2004). Distances of time series components by means of symbolic dynamics. *International Journal of Bifurcation Chaos 14*, 693–704.

Kruskall, J. et M. Liberman (1983). The symmetric time warping algorithm: From continuous to discrete. *In Time Warps, String Edits and Macromolecules*.

Liu, D., D. Umbach, S. Peddada, L. Li, P. Crockett, et C. Weinberg (2004). A random-periods model for expression of cell-cycle genes. *Proc Natl Acad Sci USA 101*, 7240–7245.

Liu, X., S. Lee, G. Casella, et G. Peter (2008). Assessing agreement of clustering methods with gene expression microarray data. *Computational Statistics and Data Analysis 52*, 5356–5366.

Luan, Y. et H. Li (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics 19*, 474–482.

Park, C., J. Koo, S. Kim, I. Sohn, et J. Lee (2008). Classification of gene functions using support vector machine for time-course gene expression data. *Computational Statistics and Data Analysis 52*, 2578–2587.

Ramoni, M. F., P. Sebastiani, et I. Kohane (2002). Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA 99*, 9121–9126.

Scrucca, L. (2007). Class prediction and gene selection for dna microarrays using regularized sliced inverse regression. *Computational Statistics and Data Analysis 52*, 438–451.

Shieh, J. et E. Keogh (2008). isax: Indexing and mining terabyte sized time series. *In Proc. of ACM SIGKDD*, 623–631.

Spellman, P., G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, et B. Futcher (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell 9*, 3273–3297.

Whitfield, M., G. Sherlock, J. Murray, C. Ball, K. Alexander, J. Matese, C. Perou, M. Hurt, P. Brown, et D. Botstein (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors molecular. *Biology of the Cell 13*, 1977–2000.

Summary

This paper addresses the clustering and classification of active genes during the process of cell division. Cell division ensures the proliferation of cells, but it becomes increasingly abnormal in cancer cells. The genes studied here are described by their expression profiles (i.e. time series) during the cell division cycle. This work focuses on evaluating the efficiency of four major metrics for clustering and classifying genes expression profiles and is based on a random-periods model for the expression of cell-cycle genes. The model accounts for the observed attenuation in cycle amplitude or duration, variations in the initial amplitude, and drift in the expression profiles.

Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol

El Ghali Lazrak* Noémie Schaller**
Jean-François Mari***

* Inra, UR 055 SAD ASTER, domaine du Joly, F-88500 Mirecourt
lazrak@mirecourt.inra.fr
** Inra/AgroParisTech, UMR 1048 SAD-APT, F-78850 Thiverval-Grignon
noemie.schaller@grignon.inra.fr
*** Loria, UMR CNRS 7503 et INRIA-Grand Est, F-54506 Vandœuvre-lès-Nancy
jfmari@loria.fr,
http://www.loria.fr/

Résumé. Nous modélisons la dynamique d'organisation spatiale et temporelle des paysages agricoles en articulant les échelles de l'exploitation agricole et du paysage. Nous développons une approche combinant deux méthodes : la modélisation des règles de décisions d'agriculteurs obtenues par enquêtes d'une part et, d'autre part, la modélisation de régularités stochastiques sur les proximités des occupations du sol.

1 Introduction

Le paysage agricole peut être perçu comme un assemblage de polygones de tailles différentes – les parcelles – où chaque parcelle porte une occupation du sol (OCS). A l'échelle de l'exploitation agricole, la façon dont chaque agriculteur organise son territoire est un processus à la fois temporel et spatial qui modèle le paysage dans son ensemble. D'une façon symétrique, les changements temporels dans la mosaïque paysagère rendent compte de décisions des différents agriculteurs qui, sans concertation systématique mais de manière souvent convergente, mettent en valeur un territoire agricole en répondant à un ensemble de contraintes et opportunités. Nous présentons une méthode de fouille de données à l'aide de modèles stochastiques pour représenter la mosaïque agricole et comprendre son évolution temporelle et spatiale. Cette fouille s'appuie sur des enquêtes en exploitations agricoles qui alimentent les interactions entre l'analyste et les experts du domaine d'étude. Pour analyser les dépendances temporelles et spatiales entre OCS, nous nous appuyons sur 2 hypothèses :

hypothèse de champ de Markov : l'OCS d'une parcelle dépend de l'OCS des parcelles voisines ;

hypothèse de chaîne de Markov : l'OCS d'une parcelle une année donnée dépend des OCS trouvées sur cette parcelle les années précédentes.

La fouille est dirigée par des experts agronomes qui, dans un premier temps à l'aide d'enquêtes effectuées dans les exploitations agricoles, retrouvent les traces dans la mosaïque paysagère des décisions prises à l'échelle de l'exploitation. Dans un second temps, une approche ascendante constate des régularités stochastiques dans la mosaïque et tente de les généraliser pour extraire des règles de décisions qui n'avaient pas été préalablement énoncées. Cette article est structuré de la façon suivante : nous présentons dans une première partie les données de la fouille constituées d'enquêtes en exploitations agricoles et par un relevé systématique des occupations des parcelles agricoles. Dans une deuxième partie, nous présentons les méthodes mises en œuvre pour nettoyer ces données et réduire la dimension de l'espace de représentation. Nous présentons alors la mesure utilisée pour capturer la dynamique des voisinages entre cultures associées à des parcelles voisines. Les résultats, tirés d'un cas d'étude situé dans la plaine de Niort sont donnés dans la section 4. Enfin, nous discutons de l'intérêt de cette approche hybride qui mêle enquêtes et modèles stochastiques pour fouiller les territoires agricoles et comprendre comment ceux-ci rendent compte des décisions prises, à une autre échelle, dans les exploitations agricoles.

2 Présentation des données

2.1 Constitution d'un corpus d'OCS

Le paysage agricole étudié s'étend sur 350 km^2 dans la Plaine de Niort. Depuis plus de 12 ans, la localisation et les occupations des parcelles sont renseignées grâce à des relevés de terrain annuels. Pendant cette période d'étude, le territoire enquêté s'est étendu. Les parcelles nouvellement enquêtées – principalement proches des prairies – ont été étiquetées *indéterminé* les premières années avant de l'être par leur véritable OCS. Ces relevés d'OCS annuels sont stockés dans un système d'information géographique sous format vectoriel et constituent une couche d'informations temporelles et spatiales.

Les frontières des parcelles changent chaque année en fonction des choix des agriculteurs (cf. Fig. 1 et 2). Pour tenir compte de ce changement, les enquêteurs définissent l'ensemble des micro-parcelles comme étant constituées de l'union de toutes les frontières de parcelles pendant la période d'étude. Il y a environ 20000 micro-parcelles dans le territoire étudié. Tous les points d'une micro-parcelle n'ont hébergé qu'une succession de cultures pendant la période d'étude. Dans la mosaïque parcellaire, le système de voisinage est irrégulier. Une parcelle a un nombre quelconque de parcelles avec lesquelles elle partage une frontière commune. Afin d'éviter la complexité due à l'irrégularité du système de voisinage et à sa variabilité temporelle (cf. Fig. 2), nous avons choisi – dans un premier temps – de rastériser cette couche d'information vectorielle avec une grille de points régulièrement espacés (10m x 10m) dans les 4 directions cardinales. Le corpus résultant est une matrice où les colonnes représentent les OCS année par année et les lignes, les différents points d'échantillonnage localisés. Le corpus compte au total 47 modalités d'OCS que nous avons regroupées, dans un travail antérieur en 11 OCS (Lazrak et al., 2010) suivant une démarche tenant compte des fréquences des OCS et de la similitude des conduites culturales. Dans le présent travail, nous nous intéressons aux prairies en tant que voisins du tournesol et du maïs. Nous avons modifié le regroupement en individualisant les prairies et en classant l'orge d'hiver – non central pour cette étude – avec le blé afin de maintenir le même nombre des modalités (Tab. 1).

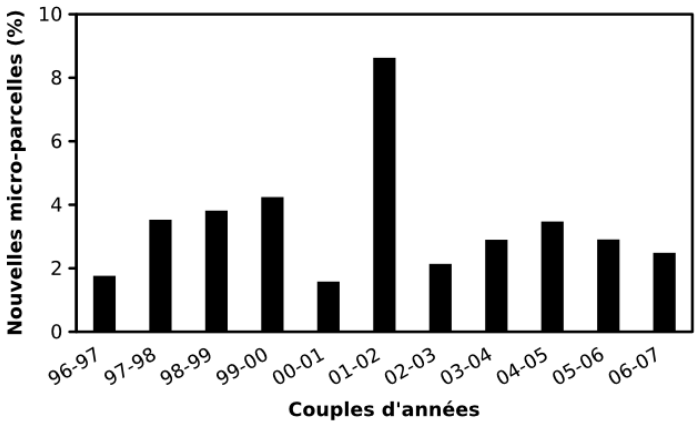


FIG. 1 – *Dynamiques inter-annuelles des frontières parcellaires dans la zone d'étude. Ces dynamiques sont exprimées en nombre de micro-parcelles nouvellement créées par rapport au nombre de parcelles de l'année précédente. Entre 2001 et 2002 plus de* 8% *des parcelles ont été redécoupées*

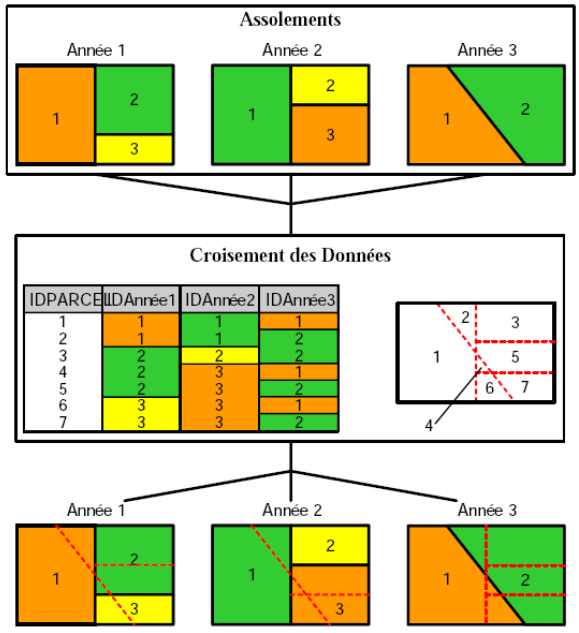


FIG. 2 – *Exemple d'évolution des limites de parcelles pendant trois années successives. L'union spatiale des frontières des parcelles pendant cette période aboutit à la définition de sept micro-parcelles*

OCS initiales	Fréq. cumulée
Blé (B)	0.337
Tournesol (T)	0.476
Colza (C)	0.600
Urbain (U)	0.696
Prairies et Luzernes(P)	0.774
Maïs (M)	0.850
Forêts et friches (F)	0.884
Orge d'hiver (O)	0.918
Ray-grass (R)	0.942
Pois (S)	0.964
Autres (A)	1.000

OCS dans les enquêtes	Fréq. cumulée
Blé (B)	0.372
Tournesol (T)	0.511
Colza (C)	0.635
Urbain (U)	0.730
Maïs (M)	0.806
Prairies (P)	0.861
Forêts et friches (F)	0.896
Luzernes (L)	0.922
Ray-grass (R)	0.946
Pois (S)	0.968
Autres (A)	1.000

TAB. 1 – *Les OCS du paysage et leurs fréquences moyennes sur la période d'étude. Les enquêtes en exploitations ont nécessité de revoir les regroupements. A gauche le regroupement selon Lazrak et al. (2010). A droite le regroupement revu.*

2.2 Les enquêtes en exploitations agricoles

Afin d'analyser les logiques individuelles des agriculteurs, nous avons combiné deux modélisations conceptuelles : le "modèle pour l'action" d'une part (Sebillotte et Soler, 1990) et le modèle d'utilisation des ressources dans l'exploitation d'autre part (Aubry et al., 1998). A partir de ces modèles, nous avons construit un cadre conceptuel générique pour modéliser les décisions des agriculteurs à travers les variables de décision, les déterminants et les règles de décisions (Schaller et al., 2010b).

Les variables de décision permettent de décrire le contenu de la décision et donner une réponse à la question : "En quoi consiste la décision ?"

Les déterminants sont tous les éléments qui influencent les variables de décision. Ils peuvent être de différentes natures : quantitatifs ou qualitatifs, internes (par exemple les ressources de l'exploitation agricole) ou externes à l'exploitation agricole (par exemple les conditions du marché, le climat, ...).

Les règles de décision sont les règles qu'un agriculteur définit et suit, en fonction des déterminants, pour faire son choix et donner une valeur à chacune des variables de décision.

Pour une exploitation, les variables de décision relatives à l'allocation des cultures aux parcelles sont : (i) la zone cultivable de la culture définie par l'ensemble des parcelles adaptées à cette culture, (ii) la taille de la sole définie comme la surface totale d'une culture une année donnée sur l'exploitation, (iii) le délai de retour défini comme le temps minimum à attendre avant de replanter la même culture sur la même parcelle – et (iv) les couples de cultures précédent / suivant acceptables (Maxime et al., 1995; Navarrete et Le Bail, 2007; Merot et al., 2008).

Entre 2006 et 2010, nous avons réalisé 67 enquêtes parmi les 185 exploitations ayant toute leur surface dans la zone d'étude. Les enquêtes visaient à comprendre le fonctionnement global de l'exploitation agricole. Nous avons distingué quatre objectifs spécifiques, qui ont été atteints grâce à quatre sessions successives d'enquêtes :

– 22 enquêtes en 2006 et 19 enquêtes en 2007 ont porté respectivement sur les stratégies des éleveurs et des agriculteurs pour faire face aux sécheresses estivales et aux interdictions d'irrigation (Havet et al., 2010) ;
– 12 enquêtes en 2009 ont porté sur les décisions des agriculteurs relatives à l'allocation des cultures dans les parcelles et au découpage des parcelles ;
– 14 enquêtes en 2010 ont porté sur l'évolution dans le temps des assolements annuels choisis par l'exploitant.

Les enquêtes étaient semi-structurées pour encourager l'agriculteur à expliciter les raisons de ses choix et leur évolution au fil du temps, notamment la façon d'allouer les successions de cultures dans les parcelles.

3 Méthodes

3.1 Choix de l'observation élémentaire

Le choix de l'observation élémentaire permet de définir les modalités d'un pixel de l'image représentant la mosaïque agricole. Plusieurs observations élémentaires sont envisageables :

1. l'OCS en un point d'une parcelle représentant son occupation ;
2. la succession d'OCS en un point d'une parcelle sur deux ou plusieurs années successives. L'observation élémentaire est un n-uplet d'OCS se chevauchant temporellement ;
3. l'OCS en un point d'une parcelle augmenté de ses 4 voisins du premier ordre : Nord (N), Sud (S), Est (E), Ouest (W). Les observations élémentaires sont des quintuplés d'OCS se chevauchant spatialement ;
4. le couple (OCS, OCS d'une parcelle voisine). Les observations élémentaires sont les configurations des cliques – deux sites voisins – se chevauchant spatialement.

La première observation est pratique pour calculer l'assolement moyen et l'évolution temporelle de celui-ci (Mari et Le Ber, 2006; Mignolet et al., 2007)

La seconde observation est utile pour retrouver les successions dominantes selon la méthode de fouille développée par (Le Ber et al., 2006; Lazrak et al., 2010).

La troisième observation permet de calculer l'information mutuelle spécifique entre OCS voisines (cf. 3.3.1), de tester l'isotropie du paysage et de déterminer la résolution spatiale optimale en fonction de la diversité des observations (Figure 3).

Enfin, l'utilisation de couples (OCS, OCS d'une parcelle voisine) permet de fouiller les voisinages entre OCS et leurs évolutions d'une façon efficace lorsque le milieu est isotrope. Cette information élémentaire permet de réduire significativement le nombre d'observations différentes, et de diminuer l'encombrement mémoire nécessaire pour représenter les distributions d'observations dans les modèles stochastiques de fouille élaborés.

3.2 Segmentation temporelle par modélisation stochastique à l'aide de HMM2

Afin d'éviter le biais introduit par la modalité *indéterminé* au voisinage des prairies pendant les premières années de l'étude, nous effectuons une segmentation des séquences d'OCS

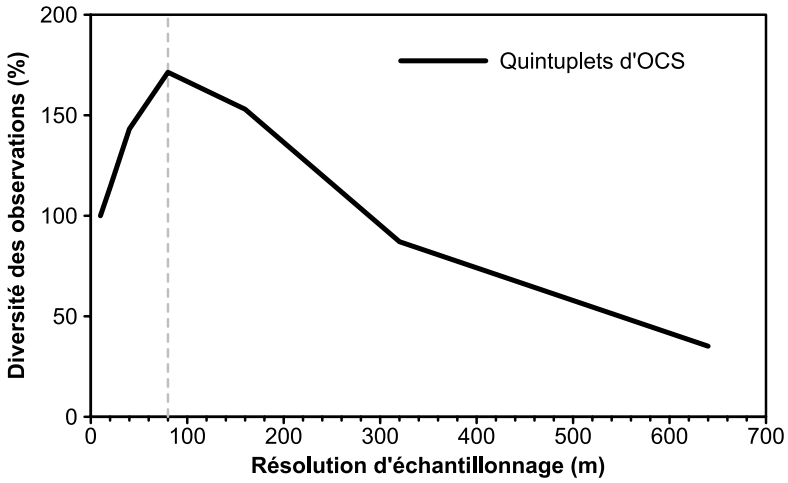


FIG. 3 – *Nombre de quintuplés (la configuration d'un site augmenté de ses 4 voisins) suivant la résolution spatiale. Le nombre de quintuplés à 10 m est considéré comme référence (100%). La courbe montre les mêmes propriétés qu'en analyse de textures d'images numériques . Quatre-vingts mètres est la résolution spatiale donnant la plus grande diversité de voisinages et sera retenue dans la suite de cette étude*

par un HMM2 afin d'isoler ce segment temporel d'*indéterminé*. Nous effectuons un alignement élastique de la séquence des 12 OCS avec un HMM2 linéaire chargé de capturer les OCS *indéterminé* dans ses premiers états. Nous utilisons des modèles de Markov cachés du second ordre HMM2 (Mari et Le Ber, 2006) pour représenter la dynamique temporelle des voisinages représentés par des quintuplés d'OCS. Chaque année t, pendant une période de 12 ans, un site S_t et ses 4 voisins (Nord, Sud, Est, Ouest) prennent 12 valeurs de quintuplés différentes représentées par les 5 variables aléatoires : $S_t, No_t, So_t, Es_t, We_t, \quad t = 1, 12$. Nous modélisons cette suite à l'aide d'un HMM2 linéaire estimé sur tous les sites du territoire étudié. La modélisation reprend les principes donnés dans (Le Ber et al., 2006; Mari et Le Ber, 2006) et cherche à segmenter la période en autant de classes que d'états. Nous cherchons des segments temporels non chevauchants pendant lesquels la distribution des quintuplés est stationnaire. L'estimation se fait selon le maximum de vraisemblance en utilisant l'algorithme forward-backward. La figure 4 montre les différentes associations des années avec les états du HMM2 suivant le nombre d'états. Par exemple, cette figure montre qu'un modèle de 7 états permet une association biunivoque entre les états et les segments et d'associer chaque état à un seul segment temporel d'une durée moyenne de 2 ans. La proportion d'*indéterminé* est maximale dans les premiers états. Les segments temporels associés sont ignorés.

3.3 Voisinages et cliques

Pour représenter la relation de voisinage entre sites, nous estimons la probabilité conditionnelle $P(V/S)$ représentant la probabilité d'avoir l'OCS x sur le site voisin – $(V = x)$ – sachant le site actuel occupé par l'OCS y : $(S = y)$. Ces probabilités sont estimées à partir des

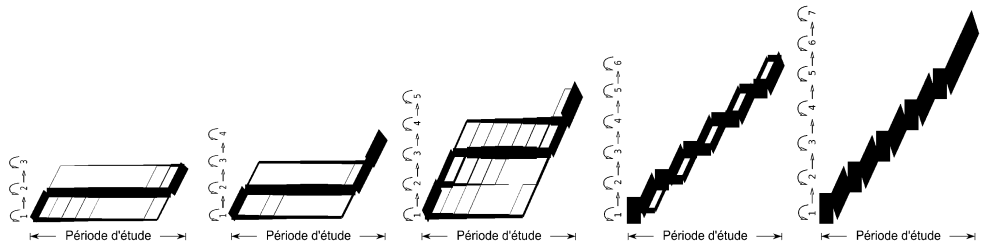


FIG. 4 – *Recherche d'une segmentation satisfaisante de la période d'étude avec des HMM2 linéaires ayant un nombre croissant d'état. L'épaisseur des traits est proportionnelle à la probabilité a posteriori des états du HMM2. La segmentation recherchée comporte au moins 2 périodes de plusieurs années ne se recouvrant pas et ne contenant pas l'OCS* indéterminé. *Dans un HMM2 de 6 états, les états 2 et 5 identifient deux périodes disjointes*

lois marginales des distributions des quintuplés. Si les distributions jointes $P(S, V)$ sont les mêmes quelle que soit la direction de voisinages – No, So, Es, We – la mosaïque agricole est dite isotrope. La distance entre deux distributions est calculée à l'aide de la divergence (Tou et Gonzales, 1974)

$$div(f, g) = \frac{1}{2} \sum_{x} (f(x) - g(x)) \log \frac{f(x)}{g(x)} \tag{1}$$

quand f et g sont deux distributions discrètes sur le même espace décrit par x.

A la dernière itération de l'algorithme *Forward-backward*, les comptes de quintuplés sont calculés sur chaque état et permettent le calcul des comptes des cliques Nord (S, No), Sud (S, So), Est (S, Es) et Ouest (S, We). A partir de ces comptes, on estime les lois marginales $P(S)$ et $P(V, S)$.

Les seules cliques que nous considérons sont constituées de deux sites voisins – soit horizontalement, soit verticalement – de configurations différentes : on ignore les cliques "plein champ" dont la configuration est faite de deux OCS identiques. Cela revient à n'échantillonner le territoire que le long des frontières des parcelles occupées par des OCS différentes. Une fois l'isotropie du paysage démontrée, nous considérons que l'orientation ne porte plus d'information et nous utilisons la clique d'OCS comme observation élémentaire pour fouiller les relations de voisinages entre cultures et leur évolution.

3.3.1 Information mutuelle spécifique

La probabilité du voisinage $P(V/S)$ n'est pas une bonne mesure pour évaluer la co-localisation de deux OCS car elle dépend des probabilités marginales des OCS. Nous utilisons l'information mutuelle spécifique (PMI comme Pointwise Mutual Information) (Novovičová et al., 2004) définie de la façon suivante :

$$PMI(x, y) = \log(\frac{P(V = x/S = y)}{P(V = x)}) = \log(\frac{P(V = x, S = y)}{P(V = x) \times P(S = y)}) \tag{2}$$

Cette quantité représente l'information apportée par la connaissance d'une variable sur l'autre. Une valeur positive signifie que le couple d'OCS (x, y) est co-localisé : les OCS x et y s'attirent.

L'expert du domaine (l'agronome) peut dans ce cas rechercher les règles de décisions des agriculteurs qui expliquent cette co-localisation. Une valeur nulle signifie que les variables V et S sont indépendantes. Une valeur négative signifie que les OCS se repoussent, l'agronome peut dans ce cas expliquer ou rechercher par enquêtes auprès des agriculteurs la (les) raison(s) d'éviter de mettre ces OCS côte à côte.

La PMI est une mesure qui se rencontre dans d'autres domaines : dans l'analyse du texte écrit (Schneider, 2005) pour la recherche des couples de mots co-localisés, et aussi en analyse d'images (Mounir Ait kerroum et Aboutajdine, 2010) quand il est question d'étude des voisinages ou des textures.

4 Résultats

4.1 Segmentation temporelle

En fonction des résultats donnés Fig. 3, nous choisissons une résolution de 80 m qui donne la plus grande diversité de voisinages, représentées par environ 40000 quintuplés différents d'OCS. A cette résolution, l'ensemble des sites est utilisé pour l'apprentissage de différents HMM2, comme le montre la figure 4. Sur chacun des états, nous calculons les lois marginales $P(V, S)$ dans chaque direction. La matrice des divergences obtenue est nulle (de l'ordre de 10^{-2}) sur chacun des états et confirme l'hypothèse d'isotropie de la mosaïque agricole. Par la suite, nous estimons un HMM2 à l'aide de cliques d'OCS comme observations élémentaires sans tenir compte de leur orientation. Nous calculons la PMI à partir des comptes des cliques sur les états sélectionnés. Le HMM2 linéaire retenu comporte 6 états et permet de définir 6 périodes différentes parmi lesquelles les périodes correspondants aux états 2 et 5 ne se recouvrent pas. L'état 2 correspond à la période 1998 à 2000, et l'état 5 correspond à la période de 2004 à 2006. Nous avons choisi ces états pour comparer les relations de voisinages du tournesol et du maïs sur deux périodes distinctes encadrant une période de sécheresse qui a influencé le raisonnement des agriculteurs.

4.2 Impact des décisions prises au niveau des exploitations sur le paysage agricole

Dans cette étude, le cadre conceptuel des enquêtes (variable, déterminant, règle) et le cadre Markovien sont liés. Le cadre formel des enquêtes a fait apparaître des règles de décisions chez les agriculteurs dont l'impact dans le paysage est évalué par la modélisation markovienne.

- Les variables "couples précédent/suivant" et "délai de retour" permettent d'explorer la dimension temporelle des décisions, et donc des régularités en termes de successions de cultures ;
- la variable "zone cultivable" permet d'explorer la dimension spatiale des décisions, et donc des régularités en termes de voisinage de cultures ;
- la variable "taille de sole", permet, le cas échéant, d'explorer l'évolution au cours du temps des surfaces des catégories de cultures.

Dans un premier temps, les méthodes d'enquêtes ont révélé une règle de décision commune entre exploitations agricoles concernant la localisation de la culture de tournesol : les agriculteurs évitent de cultiver le tournesol à proximité des forêts et bosquets en raison des

dégâts plus fréquents causés par les ravageurs (lapins, corbeaux). La figure 5 montre que le voisinage entre tournesol et forêts (T-F) est moins fréquent au cours de la période 2004-2006 (état 5 du HMM2) qu'en 1998-2000 (état 3 du HMM2) contrairement aux voisinages avec les cultures de vente telles que blé (T-B) ou colza (T-C).

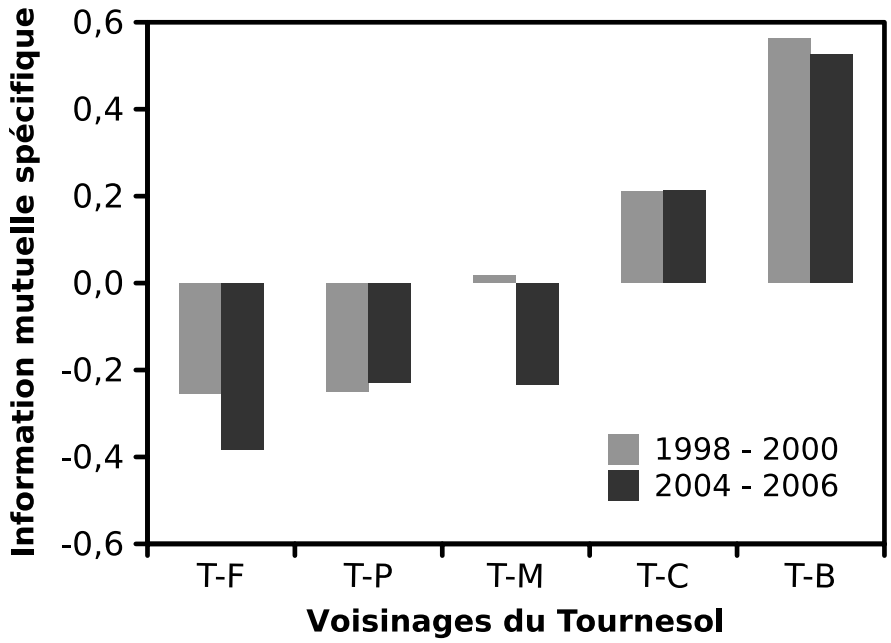


FIG. 5 – *Évolution de l'information mutuelle spécifique entre le tournesol (T), les forêts (F), les prairies (P), le colza (C) et le blé (B). Plus la PMI est élevée, plus les cultures sont fréquemment voisines*

4.3 Apparition de nouvelles règles de décisions par fouille des voisinages

La modélisation stochastique a également permis d'identifier une régularité d'évolution du voisinage entre maïs et prairies. Ces deux OCS ont tendance à être de plus en plus fréquemment voisines tandis que maïs et colza ou tournesol sont de moins en moins fréquemment voisins (Figure 6).

Cette régularité semble cohérente avec une règle de décision commune identifiée concernant le maïs : les agriculteurs réduisent les surfaces en maïs en raison des risques de sécheresse estivale et le concentrent dans les terrains les plus humides, fréquemment à proximité de prairies. Les éleveurs étendent même la surface des prairies pour sécuriser la production de fourrages dans le cas où la production de maïs serait insuffisante, d'où la co-localisation de ces deux OCS.

Ainsi, les résultats obtenus par enquêtes et modélisation stochastique apparaissent cohérents, ce qui suggère une bonne complémentarité entre les deux méthodes pour modéliser les

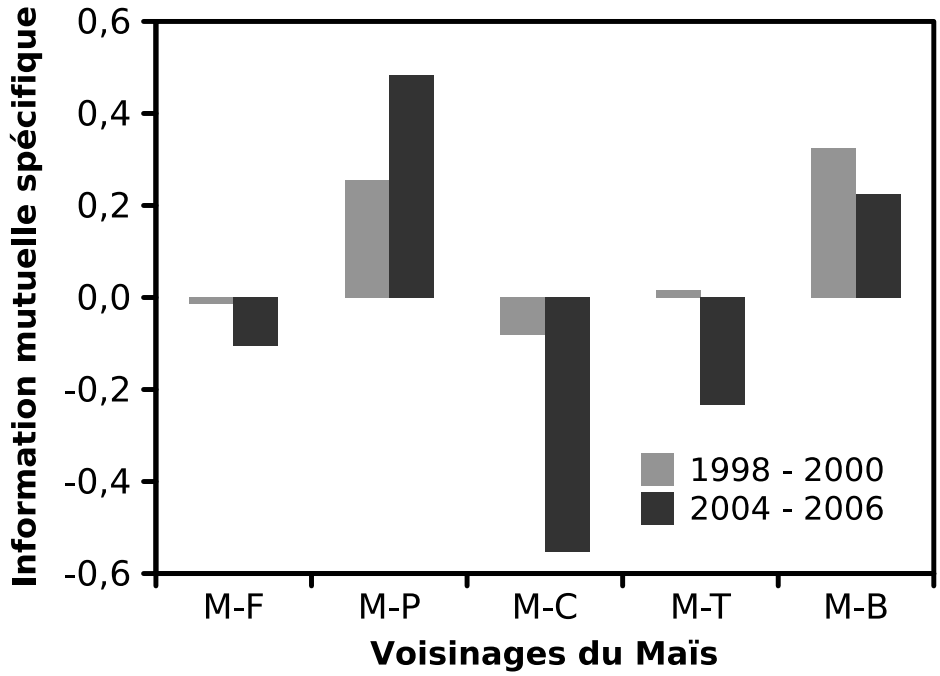


FIG. 6 – *Évolution de l'information mutuelle spécifique du maïs avec les autres OCS. Celui ci s'éloigne des forêts qui abritent les ravageurs. On remarque aussi l'évolution vers une co-localisation avec les prairies (P) et sa disparition des terres "à colza" (C). Ces dernières sont des terres de plaine, séchantes, et peu adaptées à la culture de maïs sans irrigation non restreinte*

dynamiques d'organisation spatiale des paysages. Les règles de décisions identifiées à l'échelle de l'exploitation peuvent être évaluées à l'échelle du paysage, tandis que les régularités stochastiques du paysage pourraient être en partie expliquées par des règles de décisions d'agriculteurs.

5 Conclusions

Nous avons présenté une méthode de fouille de données complexes pour identifier et modéliser des règles de décisions d'agriculteurs à l'échelle de l'exploitation agricole concernant l'assolement et retrouver leurs impacts dans le paysage sous forme de régularités stochastiques.

Les données de la fouille provenaient de deux sources différentes : une source est constituée d'enquêtes effectuées sur un échantillon d'exploitations agricoles portant sur des variables, déterminants et règles représentant les décisions d'assolement et l'autre source constituée de relevés exhaustifs d'occupation du territoire qui rendent compte de la variabilité dans le temps

et l'espace des parcelles agricoles et leurs occupations. L'information traitée se situait à deux niveaux d'échelle.

Pour contrer le fait que le territoire d'étude n'a pas été enquêté uniformément dans le temps, nous avons segmenté la période d'étude en plusieurs sous périodes par un HMM2 qui permet d'isoler les OCS "indéterminées" et de déterminer deux périodes d'étude non chevauchantes. Après avoir montré que le territoire est isotrope vis à vis des OCS, le calcul de la PMI sur les configurations des cliques a permis de montrer les tendances aux rapprochements entre OCS (phénomène de co-localisation) ou d'éloignement, que des enquêtes dans les exploitations ont expliquées en partie.

Les dynamiques d'organisation spatiale des paysages agricoles impactent de nombreux processus environnementaux. Modéliser les paysages agricoles est donc une étape clé pour pouvoir décrire et comprendre ces dynamiques d'organisation spatiale des paysages, ainsi que leurs conséquences environnementales. En plus du travail d'extraction de connaissances, une perspective importante de ce travail est ainsi d'utiliser les règles de décisions d'agriculteurs et les régularités stochastiques pour générer des paysages agricoles et tester des scénarios. A terme, cette perspective pourrait permettre aux gestionnaires des territoires agricoles d'agir sur les décisions des agriculteurs afin d'orienter favorablement les dynamiques d'organisation spatiale des paysages pour des questions environnementales locales.

Remerciements

Nous remercions le Centre d'étude biologique de Chizé (CEBC UPR 1934 CNRS), les régions Lorraine et Île de France, l'ANR BiodivAgrim et l'API Ecoger pour leurs supports.

Références

Aubry, C., A. Biarnes, F. Maxime, et F. Papy (1998). Modélisation de l'organisation technique de la production dans l'exploitation agricole : la constitution de système de culture. *Etud Rech Syst Agraires Dév* (31), 25–43.

Havet, A., P. Martin, M. Laurent, et B. Lelaure (2010). Adaptation des exploitations laitières aux incertitudes climatiques et aux nouvelles réglementations. le cas des productions bovines et caprines en plaine de niort. *Fourrages 202*, 145–151.

Lazrak, E., J.-F. Mari, et M. Benoît (2010). Landscape regularity modelling for environmental challenges in agriculture. *Landscape Ecology 25*(2), 169 – 183. http ://hal.inria.fr/inria-00419952/en/.

Le Ber, F., M. Benoît, C. Schott, J.-F. Mari, et C. Mignolet (2006). Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software. *Ecological Modelling 191*(1), 170 – 185. http ://hal.archives-ouvertes.fr/hal-00017169/fr/.

Mari, J.-F. et F. Le Ber (2006). Temporal and Spatial Data Mining with Second-Order Hidden Markov Models. *Soft Computing 10*(5), 406 – 414. http ://hal.inria.fr/inria-00000197.

Maxime, F., J. Mollet, et F. Papy (1995). Aide au raisonnement de l'assolement en grande culture. *Cah Agri* (4), 351–362.

Merot, A., J. Bergez, A. Capillon, et J. Wery (2008). Analysing farming practices to develop a numerical, operational model of farmers' decision-making processes : An irrigated hay cropping system in France. *Agricultural Systems 98*(2), 108–118.

Mignolet, C., C. Schott, et M. Benoît (2007). Spatial dynamics of farming practices in the Seine basin : Methods for agronomic approaches on a regional scale. *Science of The Total Environment 375*(1–3), 13–32. http ://www.sciencedirect.com/science/article/B6V78-4N3P539-2/2/562034987911fb9545be7fda6dd914a8.

Mounir Ait kerroum, A. H. et D. Aboutajdine (2010). Input Textural Feature Selection By Mutual Information For Multispectral Image Classification. *International Journal of Information and Communication Engineering 6*(1).

Navarrete, M. et M. Le Bail (2007). Saladplan : a model of the decision-making process in lettuce and endive cropping. *Agron Sust Dev 3*(27), 209–221.

Novovičová, J., A. Malik, et P. Pudil (2004). Feature selection using improved mutual information for text classification. In A. Fred, T. Caelli, R. P. W. Duin, A. Campilho, et D. d. Ridder (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition*, Volume 3138 of *Lecture Notes in Computer Science*, pp. 1010–1017. Springer Berlin / Heidelberg. $10.1007/978-3-540-27868-9_111$.

Schaller, N., C. Aubry, A. Havet, et P. Martin (2010b). Diversity of farmers' adaptation strategies in a context of changes and consequences on land-use dynamics : a methodological approach. In *Proceedings of the 1st Latin American and European congress on co-innovation of sustainable rural livelihood systems (Eulacias project)*, Uruguay, pp. 189–192.

Schaller, N., C. Aubry, et P. Martin (2010a). Modelling farmers' decisions of splitting agricultural plots at different time scales : a contribution for modelling landscape spatial configuration. In *Proceedings of 'Agro2010 the XIth ESA Congress*, Montpellier, France, pp. 879–880.

Schneider, K.-M. (2005). Weighted Average Pointwise Mutual Information for Feature Selection in Text Categorization. pp. 252–263.

Sebillotte, M. et Soler (1990). *Modélisation systémique et systèmes agraires*, Chapter Les processus de décision des agriculteurs : acquis et questions vives, pp. 93–102. INRA Paris.

Tou, J. T. et R. Gonzales (1974). *Pattern Recognition Principles*. Addison-Wesley.

Summary

We model the dynamics of spatial and temporal organization of agricultural landscapes by articulating the farm and landscape levels. We develop an approach combining two methods: the modeling of the decision rules of farmers acquired by on farm surveys and the stochastic modeling of neighborhood regularities.

Complexité liée à la variabilité sémantique des statistiques socio-économiques

Christine Plumejeaud*, Jérôme Gensel*

*Laboratoire d'Informatique de Grenoble, équipe STEAMER,
681 rue de la Passerelle, 38400 St Martin d'Hères
{prenom.nom}@imag.fr,
http://steamer.imag.fr/

Résumé. Cette recherche concerne le développement de méthodes pour l'analyse de similarité entre valeurs statistiques, issues de sources multiples, sur des échelles géographiques et des périodes temporelles variables. Les valeurs des indicateurs statistiques, comme les effectifs des différentes catégories socio-professionelles, mesurés en un certain lieu à une certaine période, représentent des réalités complexes et difficilement comparables. Les éléments textuels (mots clés, thèmes et résumés) présents dans les métadonnées (profil ISO19115 ou SDMX) de l'indicateur définissent une part de la signification de chaque valeur. Cependant, les catégories qu'une valeur représente sont fortement hétérogènes dans l'espace et le temps, et la valeur fait l'objet de processus d'estimation qui rendent difficile l'analyse de similarité d'une valeur avec une autre. Ce travail expose en détail cette problématique, en expliquant pourquoi les métadonnées sont nécessaires mais encore insuffisantes ou sous-exploitées, et explore les solutions existantes pour faciliter la comparaison de valeurs.

1 Introduction

L'information statistique disponible aujourd'hui sur tous les espaces géographiques est un potentiel de richesse encore largement inexploité du fait de l'hétérogénéité et de l'évolutivité des sources d'information. En Europe, particulièrement, la quantité d'indicateurs socio-économiques produits, comme le chômage, le Produit Intérieur Brut (PIB), mais également la pyramide des âges, n'a cessé d'augmenter depuis 1950 jusqu'à nos jours, pour couvrir les différentes échelles géographiques (des communes aux États) et proposer une vue plus exhaustive de l'espace européen. Par ailleurs, les méthodes de production de ces indicateurs se sont rationalisées et perfectionnées : les collectes de données, issues de comptage sur des zonages territoriaux, se sont régularisées, et la publication de métadonnées accompagnant ces données est devenue systématique, suivant en cela les recommandations d'INSPIRE, directive publiée par le Parlement européen (2007), mais aussi de différents travaux de recherche, tels ceux de Shoshani (1982); Dean et Sundgren (1996), relayés par les instances internationales sous forme de guides méthodologiques à l'intention des producteurs de données, UN/ECE (1995).

Dans ces conditions, il serait envisageable d'exploiter cette richesse pour, par exemple, estimer des valeurs manquantes pour un certain indicateur à une certaine échelle et période en

utilisant la valeur d'autres indicateurs, identiques ou équivalents, connus à d'autres échelles, d'autres périodes. Il apparaît cependant qu'un même indicateur peut avoir une sémantique différente suivant le producteur de données, et les méthodes de mesure et réajustement employées. Par exemple, les données du chômage produites par l'Institut National des Statistiques et des Études Économiques (INSEE) en France diffèrent par leur valeur de celles produites par Eurostat, alors que les deux instituts déclarent s'appuyer sur la même définition du chômage, celle du Bureau International du Travail (BIT). Ce problème de cohérence des données statistiques dans le domaine social, économique ou agricole est connu depuis longtemps des statisticiens qui manipulent ces données, qui, comme Wilks (1939), soulignent le manque d'homogénéité des échantillons, la non-comparabilité et l'inexactitude de ces données. La question est de savoir si les métadonnées *dans leur format actuel* peuvent expliquer ces divergences, mais également être utilisées pour rendre comparables des données qui ne le sont pas, et réaliser ainsi une véritable *intégration statistique*, selon Colledge (1998).

Les métadonnés sont définies comme « des données sur les données ». Selon l'ONU, UN/ECE (1995), qui complète cette définition dans le cas des données statistiques, les métadonnées doivent répondre à deux besoins. Il s'agit, d'une part de définir le contenu des données (en fournissant des définitions, mais également en décrivant le processus de production des données), et, d'autre part, d'expliquer pour quel usage les données ont été produites. Il apparaît que les métadonnées sont nécessaires, mais aujourd'hui insuffisantes pour réaliser complètement l'intégration statistique. Parmi les standards utilisés pour décrire des données statistiques (le Dublin-Core, la norme ISO 19115, le standard SDMX), aucun ne semble capable de restituer, dans un formalisme exploitable par des automates, les processus de transformation agissant sur les données. En effet, si ces informations sont généralement décrites dans des rubriques structurées, le cœur de l'information est encore décrit par du texte écrit en langage naturel. C'est un problème qui a déjà été décrit par Comber et al. (2005) pour l'analyse de données d'occupation du sol. Aujourd'hui, un nombre croissant de travaux s'orientent vers l'usage d'ontologies, tels ceux de Pattuelli et al. (2003); Comber et al. (2010), pour l'extraction semi-automatisée de métadonnées des textes, mais également pour l'alignement sémantique de données issues de sources hétérogènes.

Dans cet article, nous exposons plus en détail le problème que pose la recherche d'équivalence entre valeurs statistiques, qui s'inscrit dans la problématique plus générale de l'intégration statistique. La première section décrit quelles sont les spécificités de l'information statistique, et les causes de la variabilité sémantique. La seconde section formule des critiques à l'endroit des standards de métadonnées actuels vis à vis du problème exposé. La troisième section expose les nouvelles pistes qui sont explorées pour remédier à ce problème. La dernière section délivre nos conclusions et les perspectives qui sont entrevues.

2 Les causes de la variabilité sémantique

L'information statistique se présente sous la forme d'une série de nombres ou de variables qualitatives, collectées sur des découpages territoriaux, via des recensements ou des enquêtes datées, et qui donnent, par exemple, la mesure de la population, du nombre de ventes d'une marque de voiture, ou des préférences électorales. Ainsi, chaque valeur est associée au moins à une unité territoriale et une date. La production de ces données est opérée par différents acteurs, qui sont les instituts statistiques nationaux, des agences européennes comme Eurostat

et l'Agence Européenne de l'Environnement, ou encore des groupes de recherche, et il s'en suit une certaine hétérogénéité dans les définitions, classifications, et méthodes de production de ces indicateurs.

2.1 L'aspect multi-dimensionnel de l'information statistique

Un nombre conséquent d'indicateurs statistiques socio-économiques se présentent sous la forme de tableaux de contingence, qui associent des valeurs à des catégories croisées. C'est le cas notamment des données démographiques, qui sont publiées par sexe et par tranche d'âge, ou de la population active telle que les publie l'INSEE sur son site[1] et comme reproduit dans le tableau 1.

Population active (en milliers)	hommes	femmes	ensemble
15 ans ou plus	14 806	13 463	28 269
15-64 ans	14 702	13 394	28 096
15-24 ans	1 487	1 224	2 712
25-49 ans	9 576	8 756	18 332
50-64 ans	3 639	3 413	7 052
dont : 55-64 ans	1 801	1 677	3 478
65 ans ou plus	104	69	173

TAB. 1 – *Population active selon le sexe et l'âge en 2009.*

Ce caractère multi-dimensionnel de l'information statistique est exposé dans Rafanelli et Shoshani (1990), qui présentent l'objet statistique comme étant un quadruplet $< N, C, S, f >$ où :

- N est le nom de l'indicateur statistique ;
- C est un ensemble fini de catégories (ou dimensions) $C_1, C_2, ..., C_n$, qui ont chacune leur unité de mesure, et un domaine spécifique ;
- S est un attribut résumant la variable quantitative mesurée, qui possède un domaine de valeur, et une unité de mesure ;
- f est la fonction d'agrégation utilisée pour résumer les valeurs (la somme, le compte, le minimum, le maximum ou la moyenne).

L'objet statistique défini dans l'exemple précédent porte sur N, la « population active » utilise le compte comme fonction d'agrégation f, et S correspond au décompte de personnes actives, qui sont classées suivant deux dimensions, e.g. par tranche d'âge (C_1), et sexe (C_2).

Cependant, ces catégories sont rarement homogènes sur l'ensemble des données collectées, car elles sont spécifiques au lieu et à la période étudiés. Les catégories socio-professionnelles, les recensements à caractère ethnique ou les tranches d'âge sont des exemples de classifications instables. Les recensements à caractère ethnique font débat, et l'histoire du recensement de la population aux États-Unis (déroulée par Gauthier (2002)) démontre comment les transformations politiques et sociales d'un pays peuvent amener à reconsidérer les classifications

[1] Les données sont disponibles sur http ://www.insee.fr/fr/themes/tableau.asp ?reg_id=0&ref_id= NATCCF03170.

officielles employées. De même, en France, les actifs sont classés en fonction de leur statut professionnel (salarié, chef d'entreprise, indépendant), de la taille de l'entreprise dans laquelle ils travaillent, du secteur de l'activité (primaire, secondaire ou bien tertiaire), du niveau d'études requis pour pratiquer leur profession, etc. Mais ce mode de classification de la population en catégories socio-professionnelles n'a pas d'équivalent européen comme l'explique Kieffer et al. (2002), parce que chaque pays construit ces catégories en fonction de son histoire et de théories spécifiques. Ce problème d'instabilité des catégories, déjà relevé dès 1982 par Shoshani (1982), n'a toujours pas trouvé de réponse, et rend difficile la comparaison de valeurs à travers l'espace et le temps.

2.2 Les processus de transformation à l'oeuvre

L'exemple du chômage montre à quel point ces processus de transformation peuvent faire diverger les résultats produits, même lorsque les recensements se basent sur des définitions communes. Ainsi, en dépit d'une tentative d'harmonisation européenne symbolisée par le partage d'une définition commune définie par l'Organisation Internationale du Travail, l'INSEE et Eurostat publient des chiffres de chômage différents pour la même unité (la France) à la même date : ainsi, le taux de chômage publié par l'INSEE en Février 2008 (8,4 %) diffère de celui estimé par Eurostat (8,8 %). Pour les deux instituts, un chômeur est une personne qui n'a pas eu d'activité rémunérée supérieure à une heure pendant une semaine, et qui peut prouver sa recherche d'emploi. Cependant, les méthodes de calcul, de pondération et de correction des chiffres à partir de l'enquête emploi trimestrielle diffèrent entre l'INSEE, et Eurostat. Une explication détaillée de ces méthodes est publiée dans des documents publics [2], mais le format de ces documents ne permet d'exploiter leur contenu par un système informatique. Par exemple, Eurostat explique qu'il s'appuie sur les chiffres de l'enquête emploi publiée par l'INSEE, mais qu'il les réajuste ensuite :

> Les séries mensuelles sur le chômage et l'emploi sont calculées dans un premier temps au niveau de quatre catégories (hommes et femmes de 15 à 24 ans, hommes et femmes de 25 à 74 ans) pour chaque État membre. Ces séries sont ensuite corrigées des variations saisonnières et tous les agrégats au niveau national et européen sont calculés. [...] Pour la Suède et la Finlande, la tendance-cycle a été utilisée à la place des données corrigées des variations saisonnières jugées trop volatiles. [...]

Il faut noter ici que le texte explique comment sont transformées les données (par l'usage de la tendance-cycle en Suède par exemple), sans pour autant produire la formule exacte permettant de revenir au niveau des données brutes. Le chômage illustre aussi l'évolution des méthodes de calcul et de mesure dont sont l'objet les indicateurs. Goux (2003) explique que l'INSEE fait évoluer régulièrement sa méthodologie de calcul du chômage, décrite et justifiée dans des documents accessibles en ligne [3].

Par ailleurs, les indicateurs statistiques sont parfois le produit de réflexions théoriques qui visent à produire une représentation synthétique d'un ensemble de facteurs mesurés par des données brutes. Une étude récente de l'UMS 2414 RIATE et al. (2008) portant sur le déclin démographique en Europe, et présentée devant le parlement européen en 2008, donne un

[2] INSEE : http ://www.insee.fr/fr/methodes/sources/pdf/eeencontinu.pdf
EUROSTAT : http ://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-29012010-AP/FR/3-29012010-AP-FR.PDF

[3] http ://www.insee.fr/fr/methodes/sources/pdf/estimations_chomageBIT_enquete_emploi.pdf

exemple de ce type d'indicateur. Dans le texte de l'étude se trouve la définition d'un indicateur synthétique de vieillissement :

> Il suffit alors d'effectuer le rapport entre l'âge moyen d'une population et son espérance de vie en bonne santé pour en déduire un *indicateur synthétique de vieillissement* exprimé sous la forme d'un pourcentage du potentiel d'activité de la population qui a été consommé.

A travers cet exemple, il apparaît que la représentation du calcul d'un indicateur composite à partir d'indicateurs de base est possible (ici, c'est le ratio de deux indicateurs de base qui fournit l'indicateur). En effet, si la relation mathématique entre l'indicateur synthétique de vieillissement (noté V) et l'espérance de vie en bonne santé est connue, il devient possible d'estimer des valeurs manquantes de V en employant des indicateurs bruts (espérance de vie, âge moyen). Par ailleurs, la connaissance de la formule de calcul permettrait de déduire que l'indicateur V est un bon *proxy*[4] pour calculer l'espérance de vie d'une population, lorsque son âge moyen est connu. Dans ces conditions, si les transformations sont clairement exprimées, ainsi que les éléments sur lesquels elles opèrent, l'analyse de similarité est quasiment directe.

Cependant, les formules permettant de fabriquer des indicateurs composites sont rarement si simples. Un guide méthodologique résumant les bonnes pratiques est publiée par l' Organisation pour la Coopération et le Développement Economique (2008), pour l'ensemble des opérations qui doivent être mises en oeuvre : sélection, normalisation, estimation de données manquantes, pondération et agrégation des données. Bien qu'effectivement les opérations soient plus complexes, l'étude de ce document fait apparaître l'usage récurrent d'une certaine terminologie pour ces opérations : les mots « correlation », « moyenne », « variance », « hypothèse » reviennent très fréquemment. La description des transformations recourt généralement à un vocabulaire statistique, qu'il s'agit de référencer, afin de construire un dictionnaire des transformations possibles (comme simplement une pondération, une moyenne, une différence). Ce dictionnaire pourrait servir de support à une structuration des métadonnées, dont l'usage est fortement recommandé car, de plus en plus, la production de ce type d'indicateur vise un certain niveau de qualité, non seulement en termes de méthode de construction, mais également en termes de documentation.

3 Usage des métadonnées

Dans cette section, nous examinons les différents standards de métadonnées en profondeur, pour vérifier s'ils répondent complètement à la problématique de l'intégration statistique.

3.1 Les normes de l'information géographique

Les données statistiques étant à références spatiale et temporelle, il semble naturel d'étudier les standards de métadonnées dédiés aux données géographiques. Le premier, le Dublin-Core, créé à l'initiative des États-Unis d'Amériques en 1995, insiste particulièrement sur les aspects légaux (droits de propriété et d'usage) concernant les données. Le standard ISO 19115, promu

[4] *proxy* est un des termes employé en statistiques comme synonyme de variable auxiliaire aidant à retrouver une variable inconnue, qui serait ici l'espérance de vie d'une population.

par la norme INSPIRE pour la diffusion de données géographiques, inclut les 15 éléments proposés dans le Dublin-Core. Les éléments communs au Dublin-Core concernent :

- la description du contenu : le titre, et la date de publication, le sujet et une description des données ;
- la désignation des responsables : l'identification des contributeurs et responsables de cette publication ;
- les informations nécessaires pour l'utilisation des données : les droits de propriété, le format de distribution ;
- les informations relatives à la qualité des données : mode et fréquence de maintenance, couverture spatio-temporelle.

La norme ISO 19115 intègre des éléments spécifiques à l'information géographique permettant de reconnaître leur extension spatiale et le mode de représentation utilisé. Notre intérêt se concentre sur les éléments disponibles pour décrire la *qualité interne* de données spatio-temporelles, regroupés dans l'élément $DQ\_Quality$, qui se mesure à l'aune de sept critères (le lignage, l'exactitude de la position spatiale, et temporelle, la précision des attributs thématiques, la complétude, la cohérence logique et sémantique), d'après Servigne et al. (2006). Tous ces critères sont quantifiables, à l'exception du lignage qui doit permettre de retracer les procédures d'acquisition, les sources, et les méthodes employées pour transformer les données brutes, et obtenir par dérivation la donnée décrite. A travers ce critère, deux objectifs sont visés : assurer que les méthodes de production respectent les normes en vigueur, et donc s'assurer que les données sont comparables, mais également rassurer l'utilisateur sur la nature des sources employées. La figure 1 représente la structure de l'élément $DQ\_Lineage$ de la norme, qui utilise de façon conjointe un élément $LI\_ProcessStep$ pour décrire une étape de transformation, et LI_Source pour décrire les « ingrédients » de la transformation. L'élément $LI\_Lineage$ s'applique à tout ou partie du jeu de données : l'emprise spatio-temporelle des éléments concernés par cette transformation est décrite dans l'élément EX_Extent.

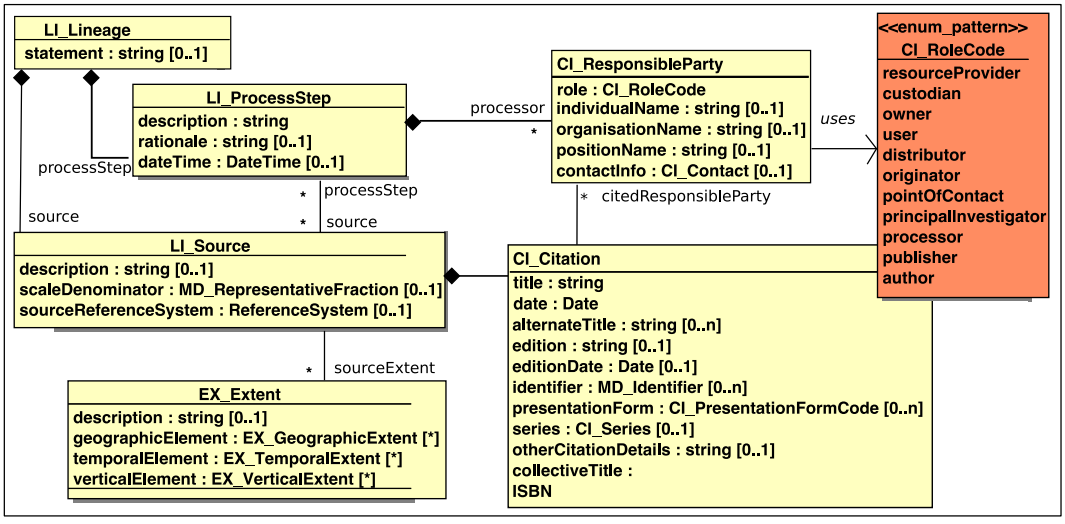


FIG. 1 – *L'élément Lineage défini par la norme ISO 19115.*

Cette norme extensible, a fait l'objet d'une adaptation pour les données statistiques afin de la rendre plus opérationnelle, Plumejeaud et al. (2010). Il s'agissait de simplifier la saisie de métadonnées pour des données statistiques. Par exemple, la référence spatiale correspond à un code d'unité territoriale, et donc le mode de définition des éléments EX_Extent devait être modifié. De même, l'identification de chaque indicateur a été enrichie avec son code et son unité de mesure. Enfin, contrairement à un jeu de données géographiques plus classique, chaque valeur du jeu de données possède un lignage spécifique, puisque les valeurs de certaines unités territoriales du jeu de données sont issues de sources spécifiques. La spécialisation de l'élément $LI\_Lineage$ de la norme simplifie la représentation du lignage. En effet, la saisie de chaque processus de transformation et de la source associée est trop fastidieuse pour les utilisateurs, et donc il a été convenu de décrire soit par un champs textuel, soit par l'adjonction d'un document multi-média les processus de transformation. Mais, en réalité, à moins d'extraire les informations depuis les documents, (ce qui suppose de créer un corpus spécialisé, et d'inventer un format pour structurer ces informations), cet élément dérivé de $LI\_Lineage$ ne propose que l'URL de la source originelle des données, et ne permet pas de retranscrire les transformations des données. Par ailleurs, dans cette extension, l'aspect multi-dimensionnel de l'information statistique a été négligé, et il reste à créer des éléments de type *MD_Catégorie* décrivant la nature et le domaine des catégories associée aux indicateurs.

3.2 SDMX, un modèle pour l'échange de données statistiques

Les langages semi-structurés ont été avancés depuis quelques années comme une solution au problème de non-interopérabilité entre Systèmes d'Information Statistiques (SIS), Meyer et al. (2004). La raison étant que ces langages (basés sur XML et l'emploi de schémas XSD) permettent d'embarquer une description du format des données dans les données. Ainsi, la flexibilité et la souplesse s'en trouveraient accrue. C'est la raison pour laquelle le Statistical Data Model eXchange (SDMX) est aujourd'hui promu par l'OCDE, mais également Eurostat, ou d'autres organismes de production de données statistiques. Les auteurs de SDMX se sont concertés pour définir une liste de concepts[5], termes[6], et codes[7] identifiés et partagés dans un registre, en vue de faciliter l'intégration statistique. Les concepts, référencés par un identifiant unique, quelque soit la langue, peuvent être valués soit par des codes, indépendants de la langue, soit par du texte. Par exemple, le statut (OBS_STATUS) d'une valeur statistique est décrit par une liste de codes (A, B, E, F, I, M, P, S) qui signale si la valeur est normale (A), manquante (B), estimée (E), etc., et le registre définit exactement la signification de ces codes. L'échange de données statistiques nécessite la définition du format des données dans un fichier externe, le *Data Structure Definition (DSD) file*, qui structure l'information à l'aide de ces codes et ces concepts harmonisés. La liste des concepts et des codes qui leurs sont associés peut également être étendue dans le fichier DSD. Par exemple une catégorie TRANCHE_AGE sera définie comme concept, avec la liste des différentes tranches d'âges codifiées, décrites chacune par un champs texte. Le fichier de données fait référence à cette description au niveau par exemple la valeur observée, lorsque sont présentées les données :

<Obs TIME_PERIOD="2009" OBS_VALUE="14 702" OBS_STATUS="A" TRANCHE_AGE="15-64" SEX="M"/>

[5] http ://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf
[6] http ://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf
[7] http ://sdmx.org/wp-content/uploads/2009/01/02_sdmx_cog_annex_2_cl_2009.pdf

La prise en charge dans ce standard de l'aspect multi-dimensionnel de l'information statistique est donc meilleure que dans le format ISO 19115. En ce qui concerne la définition des transformations, le concept DOC_METHOD s'utilise pour décrire les méthodes de mesure, la définition et les transformations opérées sur les données, et le concept COMPARABILITY pour fournir un commentaire sur l'équivalence de cette donnée avec une autre. Néanmoins, ces champs sont valués par du texte, et présentent donc la même difficulté d'analyse de leur contenu que celle décrite pour la norme ISO 19115. Ce format ne résout pas non plus le problème d'hétérogénéité des catégories, comme le rapporte l'expérience à grande échelle menée par Oakley et al. (2005), membres du Bureau Australien des Statistiques, producteur officiel en Australie, dans le cadre d'un projet visant à exporter au format SDMX les statistiques économiques nationales. Ils rapportent en particulier le cas de concepts de SDMX qui ne trouvent pas leur pendant dans leur base de données statistiques, parce qu'ABS utilise une classification nationale spécifique, qui, par exemple, codifie les activités de pêche, agriculture et d'exploitation forestière avec la lettre A, alors que SDMX classe séparément les activités de pêche (avec le code AYB) et les activités d'agriculture, chasse et exploitation forestière, codées AYA.

4 Pour aller plus loin que les métadonnées

Les problèmes qui viennent d'être exposés méritent a priori l'intervention d'experts pour leur résolution. En effet, comment réussir à aligner des catégories ou bien réaliser un graphe des transformations des données sans une acquisition manuelle de ces connaissances ? Les deux problèmes peuvent être traités séparément : dans le premier cas, il s'agit de tirer un meilleur parti des textes contenus dans les métadonnées, et dans le second d'étendre les métadonnées avec des formalismes peut-être plus adaptés à la retranscription fine d'un lignage. Nous examinons ici un certain nombre de travaux qui offrent des pistes intéressantes.

4.1 Alignement sémantique des indicateurs

Les travaux de Wadsworth et al. (2006) et de Comber et al. (2005, 2010) étudient plusieurs approches pour l'alignement des différentes catégories d'usage de sol, en vue de comparer des cartes d'occupation du sol produites par différents organismes, à dix ans d'intervalles. Ces catégories, qui sont qualitatives et non-ordonnées présentent le même niveau d'hétérogénéité que les catégories socio-professionnelles. Parmi les différentes approches étudiées pour leur alignement, il ressort qu'une analyse détaillée de la description de la catégorie (lorsqu'elle comporte plus de 100 mots) par une technique de fouille de données textuelle permet d'établir une matrice de recouvrement entre catégorie, que les auteurs démontrent supérieure aux autres techniques qui nécessitent l'intervention d'experts. La technique consiste à établir la liste des mots employées dans la description de chaque catégorie, puis à calculer leur poids sémantique dans chaque catégorie à l'aide d'une mesure de fréquence inverse dans le document (*Inverse Document Frequency*, IDF, telle que discutée par Robertson (2004)). Le niveau de recouvrement entre chaque catégorie est ensuite calculé via l'usage de la théorie de l'analyse sémantique latente probabiliste introduite par Hofmann (1999), qui stipule que la similarité sémantique entre deux concepts (ici les catégories) peut-être mesurée par la quantité d'information qu'ils partagent (les mots). Cette approche permet également d'identifier les concepts qui structurent une classification données, et les termes qui s'y rapportent.

Cette approche pourrait être appliquée à l'ensemble des métadonnées collectées sur les indicateurs statistiques, pour l'analyse de similarité entre indicateurs, mais également pour l'alignement des catégories associées aux valeurs des indicateurs, à la condition que les textes descriptifs produits soient suffisamment longs. Il est ainsi possible d'établir une méta-classification des classifications d'indicateurs ou de catégories, qui peut être structurée sous la forme d'une ontologie de concepts, liées par des relations de subsomption,« is-a »(ou héritage), ou bien des relations méréologiques (ou composition), « Is_part_of », mais également une relation sémantique de domaine indiquant dans quelle mesure une catégorie recouvre l'autre, telle que « intersects », valuée par l'indice de recouvrement. Par exemple, les concepts de chômeur, population active et population peuvent être identifiés via l'analyse des définitions des indicateurs statistiques, qui sont des instances de cette ontologie. Il apparaît que le nombre de chômeurs défini et mesuré par l'INSEE ou l'EUROSTAT se rapporte à un concept de « chômeur » identique, et qu'un chômeur est un membre de la population active, qui elle-même fait partie de la population totale :

- *chômeur_INSEE* is-a *chômeur*
- *chômeur_Eurostat* is-a *chômeur*
- *chômeur* is_part_of *population-active*
- *population-active* is_part_of *population*

L'analyse des métadonnées peut donc servir à établir puis peupler une ontologie. Pattuelli et al. (2003) montrent que l'existence d'une ontologie statistique facilite la compréhension des termes statistiques, et la création d'un glossaire interactif. Cependant, identifier que les chômeurs suivant l'INSEE ou EUROSTAT se rattachent à un même concept "chômeur" ne résoud pas le problème d'équivalence des valeurs que nous avons pointé. L'ontologie sert essentiellement à résoudre et à raisonner sur les problèmes d'équivalences entre catégories. Il s'agit aussi de pouvoir ensuite raisonner sur les valeurs au niveau de leur transformations (modalités de calcul, réajustement, estimation).

4.2 La description des transformations

Pour raisonner au niveau des transformations, il faut se doter de formalismes de représentation. La description des transformations en vue de retrouver des données originelles à partir de données transformées est une ambition affichée par Woodruff et Stonebraker (1997), qui proposent un formalisme sous forme de graphe (direct et acyclique) du flot de données dans une base de données, chaque nœud modélisant une fonction de transformation, et chaque arc correspondant à une donnée particulière. Le graphe est défini par l'utilisateur via une interface graphique. L'utilisateur spécifie chaque fonction f de transformation par son nom et son type, ainsi que les paramètres d'entrée (des attributs de tuples), par leur type et leur nom, et l'indicateur produit par leur type. Brilhante et al. (2006) structurent les indicateurs dans une ontologie et associent les indicateurs à leurs formules de calcul dans une base de connaissance. Cependant les modalités de construction et d'acquisition de ces opérateurs ne sont pas mentionnées, et les formules ne sont pas rédigées suivant un formalisme mathématique standardisé.

L'usage d'un graphe de flots de données semble adapté à la problématique du lignage, mais si la description des transformations nécessite la contribution d'un utilisateur, il faut absolument proposer des outils pour faciliter cette saisie. L'interface graphique pourrait par exemple être enrichie avec un dictionnaire des fonctions de transformation usuelles dans le domaine statistique, et un dictionnaire des données utilisables comme entrée de la transformation. Ainsi,

l'utilisateur éviterait la saisie du type, et du nom de l'indicateur par un simple « *drag and drop* » depuis une liste d'indicateurs recensés dans l'ontologie. L'ontologie des indicateurs précédemment évoquée est réutilisable pour retrouver rapidement l'instance d'un indicateur utilisé dans la formule comme ingrédient, et ceci a déjà été souligné dans le travail de Brilhante et al. (2006). Il s'agit aussi d'établir le dictionnaire des transformations, le plus souvent des opérandes mathématiques, qui peuvent être simples comme la division, ou plus complexes comme une formule de normalisation. La saisie des formules peut être envisagée à l'aide de langages comme MathML[8], (pour Mathématiques Markup Language), ou bien OpenMath[9]. Par exemple, la formule de l'indicateur synthétique de vieillissement pourrait s'écrire (en utilisant Amaya[10] comme éditeur WYSIWYG de MathML) comme dans la figure 2 :

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE math PUBLIC "-//W3C//DTD MathML 2.0//EN" "http://www.w3.org/TR/MathML2/dtd/mathml2.dtd">
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mi>IndicateurSynthétiqueVieillisement</mi>
  <mo>=</mo>
  <mfrac bevelled="true">
    <mi>ageMoyenPopulation</mi>
    <mi>espéranceVieBonneSanté</mi>
  </mfrac>
</math>
```

FIG. 2 – *Formule d'un indicateur composite, exprimée en MathML.*

Ces formules peuvent être rattachées aux indicateurs dans l'ontologie, ou à leurs instances. Par exemple, les formules d'ajustement des variations saisonnières pourraient être accolées aux instances de l'indicateur *chômeur* : *chômeur_INSEE* et *chômeur_Eurostat*.

5 Conclusion et perspectives

Cet article expose en détail les problèmes que pose la recherche d'équivalences entre valeurs d'indicateurs socio-économiques lorsque ces statistiques sont produites par différentes sources, en différents lieux et dates. Les métadonnées, qui avaient été recommandées comme une solution à l'intégration statistique, permettent effectivement de donner du sens aux données, mais cette solution *a minima* ne résout pas les conflits de catégorie, ni les changements de sémantique liées aux transformations de certaines données. En effet, les métadonnées restent essentiellement une description textuelle qui n'autorise pas un traitement automatisé de l'information. Il apparaît néanmoins que l'emploi de techniques de fouille de texte appliquées sur les métadonnées peut faciliter le calcul d'alignement des classifications, mais également aider à construire une ontologie statistique, en mettant en relation dans un graphe les concepts représentés par les indicateurs. Cette ontologie serait utilisable également pour établir une base de connaissance des transformations.

Pour exploiter cette base de connaissances, et par exemple calculer l'équivalence entre deux valeurs, il faudrait l'associer à un raisonneur capable de résoudre automatiquement les liens entre valeurs d'indicateurs par l'interprétation des formules de transformation. Dans cette perspective, nous allons travailler sur la définition d'un formalisme plus avancé des transforma-

[8] http ://www.w3.org/TR/MathML3/
[9] http ://www.openmath.org/documents/bibliography.html
[10] http ://www.w3.org/Amaya/Overview.html

tions, tel qu'un Langage de Modélisation Algébrique (LMA), et à la construction de l'ontologie des indicateurs statistiques associée.

Références

Brilhante, V., A. Ferreira, J. Marinho, et J. S. Pereira (2006). Information integration through ontology and metadata for sustainability analysis. In *International Environmental Modelling and Software Society (iEMSs) Third Biannual Meeting "Summit on Environmental Modelling and SoftwareThird Biennial Meeting"*, Burlington, USA.

Colledge, M. J. (1998). Statistical integration through metadata management. *International Statistical 67*(1), 79–98.

Comber, A., P. F. Fisher, et R. A. Wadsworth (2005). A comparison of statistical and expert approaches to data integration. *Journal of Environmental Management 77*, 47–55.

Comber, A., A. Lear, et R. Wadsworth (2010). A comparison of different semantic methods for integrating thematic geographical information : the example of land cover. In *AGILE'2010*.

Dean, P. et B. Sundgren (1996). Quality aspects of a modern database service (position paper). In P. Svensson et J. C. French (Eds.), *SSDBM*, pp. 156–161. IEEE Computer Society.

Gauthier, J. G. (2002). *Measuring America : the decennial censuses from 1790 to 2000*. U.S. Census Bureau.

Goux, D. (2003). Une histoire de l'enquête emploi. *Economie et Statistique* (362).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In H. M., G. F., et T. R. (Eds.), *Proceedings of 22nd International Conference on Research and Development in Information Retrieval*, Berkeley, Californie, US, pp. 50–57.

Kieffer, A., M. Oberti, et E. Preteceille (2002). Enjeux et usages des catégories socio-professionnelles en europe. *Sociétés contemporaines 45-46*, 157–185.

Meyer, D., T. Hothorn, F. Leisch, et K. Hornik (2004). Statdataml : An xml format for statistical data. *Journal of Computational Statistics 19*(3), 493–509.

Oakley, G., A. Hamilton, et J. Michel (2005). Experiences and plans of the australian bureau of statistics related to data and metadata exchange. Technical report, Australian Bureau of Statistics.

OCDE, O. (2008). Handbook on constructing composite indicators : Methodology and user guide. Technical Report ISBN 978-92-64-04345-9, OCDE.

Parlement européen (2007). directive 2007/2/ce établissant une infrastructure d'information géographique dans la communauté européenne (inspire) (http ://inspire.jrc.ec.europa.eu/).

Pattuelli, M. C., S. W. Haas, S. W. Haas, et J. Wilbur (2003). The govstat ontology. In *Proceedings of the 2003 annual national conference on Digital Government research (DG.O)*.

Plumejeaud, C., J. Gensel, et M. Villanova-Oliver (2010). Opérationnalisation d'un profil iso 19115 pour des métadonnées socio-économiques. In *Actes du XXVIIIème congrès INFORSID*, pp. 25–41.

Rafanelli, M. et A. Shoshani (1990). Storm : A statistical object representation model. In Z. Michalewicz (Ed.), *In Proc. of Statistical and Scientific Database Management, 5th In-*

ternational Conference SSDBM, Charlotte, NC, USA, pp. 14–29.

Robertson, S. (2004). Understanding inverse document frequency : On theoretical arguments for idf. *Journal of Documentation 60*(503-520), 2004.

Servigne, S., N. Lesage, et T. Libourel (2006). Spatial data quality components, standards and metadata. In *Fundamentals of Spatial Data Quality*, Number ISBN 1905209568, pp. 179–210. International Scientific and Technical Encyclopedia.

Shoshani, A. (1982). Statistical databases : Characteristics, problems, and some solutions. In *VLDB '82 : Proceedings of the 8th International Conference on Very Large Data Bases*, San Francisco, CA, USA, pp. 208–222. Morgan Kaufmann Publishers Inc.

UMS 2414 RIATE, UMR 8504 Géographie-cités, LIG, IGEAT, Umeå University, Università l'Orientale, et Université "Alexandru Ioan Cuza" Iaşi. (2008). Régions en déclin : un nouveau paradigme démographique et territorial. Etude pour le Parlement Européen - Structural and cohesion policies - Juillet 2008 IP/B/REGI/IC/2007-044, Parlement européen.

UN/ECE (1995). Guidelines for the modelling of statistical data and metadata. Technical report, UN/ECE, New York, Geneva.

Wadsworth, R. A., A. Comber, et P. F. Fisher (2006). Expert knowledge and embedded knowledge or why long rambling class descriptions are useful. In G. E. Andreas Riedl, Wolfgang Kainz (Ed.), *Progress in Spatial Data Handling, Proceedings of SDH*, pp. 197 – 213. Springer Berlin / Heidelberg.

Wilks, S. S. (1939). The rise of modern statistical science. In *MIT Industrial Statistics Conference*, New York, NY, USA, pp. 283–310. Pitman Publ. Corp.

Woodruff, A. et M. Stonebraker (1997). Supporting fine-grained data lineage in a database visualization environment. In *Proceedings of the Thirteenth International Conference on Data Engineering*, Number Report n °UCB/CSD-97-932, Birmingham, U.K., pp. 91–102.

Summary

This research aims at developing new methods for similarity analysis of statistical values, coming from various sources, at different geographical scales and validity periods. The values associated to statistical indicators, such as "social and occupational group workforce", measured on given locations and for given dates, have not an homogeneous semantic. The indicator metadata (using a ISO 19115 profile, or inside the SDMX format) provide a part of the value' meaning through textual elements (keywords, themes, abstract). However, a value is often a combination of various categories (sex, age, group) that change over time and space; besides, it can be the result of an estimation process. In those conditions, values are not equivalent: this is the semantic variability of statistical indicators. We explain why the metadata, although necessary, are not sufficient, and we dress the list of the various existing solutions for this problem.