Approche hybride à base d'ontologie pour le clustering par contraintes

Hatim Chahdi[*,**] Nistor Grozavu[**], Isabelle Mougenot[*],
Laure Berti-Equille[*,***], Younès Bennani[**]

[*]UMR U228 Espace-Dev, IRD - Université de Montpellier
Maison de la télédétection - 500 Rue J.F. Breton, 34093 Montpellier
prenom.nom@ird.fr
[**]UMR 7030 LIPN, CNRS - Université Paris 13
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
prenom.nom@lipn.univ-paris13.fr
[***]Qatar Computing Research Institute, Hamad bin Khalifa
University Doha, Qatar

Résumé. Les méthodes de clustering semi supervisé font très souvent usage de contraintes afin de répondre au mieux aux besoins des utilisateurs. Cependant, dans des domaines comme le traitement et l'interprétation d'images satellites, la génération de ces contraintes est coûteuse et demande une grande expertise du domaine. Dans cet article, nous présentons une approche hybride qui exploite en amont les ontologies pour automatiser la génération des contraintes pour le clustering et introduire une labellisation sémantique des clusters. L'objectif est de disposer d'une brique ontologique qui va permettre d'étiqueter sémantiquement une partie du jeu de données en se basant sur des mécanismes de raisonnement déductif, puis d'utiliser les résultats de ce raisonnement pour (i) générer des contraintes qui vont guider le clustering, (ii) étiqueter sémantiquement les clusters obtenus avec les concepts de l'ontologie. Nous avons appliqué notre approche à la classification d'images satellites et les résultats obtenus valident la capacité de notre approche à améliorer le clustering sans l'utilisation de données étiquetées ni l'intervention manuelle de l'expert.

1 Introduction et Motivations

Ces dernières années, différents portails et programmes ont facilité l'accès aux images satellites et ont ainsi contribué à les rendre incontournables dans diverses problématiques environnementales, sociétales et sanitaires. Cependant, l'exploitation de ces images requièrent leurs analyse et interprétation préalables, tâches qui sont en grande partie réalisées manuellement par les experts. Elles constituent le goulot d'étranglement empêchant l'exploitation de tout le potentiel qu'offrent ces différentes images satellites. Dans ce cadre, les techniques d'extraction de connaissances peuvent jouer un rôle facilitateur aidant à automatiser l'analyse et l'interprétation de ces images. Classiquement utilisé dans un cadre exploratoire et non supervisé, le clustering vise à partitionner de gros volumes de données non-étiquetées en un

ensemble de groupes de données au regard de leurs similarités. Cependant, dans le contexte des images satellites, ce partitionnement est fortement lié à l'intérêt que porte l'utilisateur final à ces images. Deux experts qui n'ont pas le même intérêt thématique évalueront de manière très différente les résultats du même clustering. Dans ce cadre, les techniques du clustering semi supervisées représentent une bonne solution pour la prise en compte de la vision de l'expert dans le cadre clustering. Ces techniques sont efficaces pour contraindre la formation de clusters en utilisant les contraintes de l'utilisateur. Cependant, elles nécessitent l'intervention de l'expert à chaque fois pour spécifier les contraintes. La sémantique apportée par l'étiquetage de l'expert est aussi perdue dans ce type de techniques, seules les contraintes sont retenues par le clustering. Dans cet article, nous proposons une approche hybride alliant, d'une part l'explicitation des connaissances du domaine d'intérêt pour, dans un premier temps, étiqueter sémantiquement le plus grand nombre de données, et d'autre part, le clustering semi supervisé à base de contraintes pour guider dans un deuxième temps les activités du clustering et étendre l'étiquetage sémantique des données. Les apports de notre approche sont multiples. Elle permet de générer automatiquement des contraintes à partir des connaissances et sans l'intervention manuelle de l'expert. L'utilisation du raisonnement à base d'ontologie apporte une modularité et une interprétation automatique du clustering. Finalement, l'explicitation des connaissances en ontologie permet leur exploitation pour l'analyse de plusieurs jeux de données.

2 Travaux connexes

L'introduction de la connaissance a priori dans le clustering a fait l'objet de plusieurs travaux (Davidson et Basu, 2007). Afin d'améliorer les résultats du clustering, l'intégration des connaissances vise à prendre en compte la vision de l'utilisateur et/ou à apporter une information supplémentaire sur les données ou le domaine des données à traiter. Cette connaissance a priori est souvent fournie par l'expert de façon non formalisée et se manifeste de plusieurs façons. Parmi les techniques proposées, on retrouve le clustering par contraintes au niveau des instances. Les travaux publiés dans ce domaine ont prouvé l'efficacité des contraintes pour guider directement la formation des clusters. Les connaissances sont ainsi exprimées sous forme de liens (transitifs) *must-link* et *cannot-link*. Introduit initialement par Wagstaff et Cardie en 2000, un *must-link* $ML(d_i, d_j)$ spécifie que deux instances, notées d_i et d_j, doivent se retrouver dans le même cluster final. Tandis qu'un *cannot-link* $CL(d_i, d_j)$ spécifie que les deux instances ne peuvent appartenir au même cluster. Pour prendre en compte les contraintes, des modifications sont apportées aux algorithmes de clustering :

— Modification de la phase de mise à jour de l'affectation des instances aux clusters (Wagstaff et Cardie, 2000; Shental et al., 2004) ;
— Modification de la phase d'initialisation des clusters (Davidson et Ravi, 2005) ;
— Modification de la fonction objective du clustering (Basu et al., 2004).

Bien que le domaine du clustering par contraintes a reçu un fort intérêt ces dernières années, peu de travaux se sont intéressés à la génération automatisée des contraintes. Les méthodes proposées utilisent toujours des contraintes obtenues manuellement, que ce soit directement sous formes de ML et CL de l'expert, ou à partir de données étiquetées manuellement. De plus, l'information sémantique apportée par la classe d'appartenance des instances liées ou préalablement étiquetées n'est pas exploitée.

3 Approche à base d'ontologie pour automatiser le clustering par contraintes

Nous présentons dans cet article une approche permettant d'exploiter les ontologies OWL comme support de connaissances pour guider et renforcer automatiquement le clustering. Notre proposition est une démarche méthodologique valable dans toute problématique qui dispose de données matricielles, et de connaissance experte formalisée associée à ces données. Notre démarche repose sur deux axes : le premier est l'utilisation du raisonnement pour automatiser l'interprétation des connaissances et l'étiquetage sémantique des données. Le deuxième propose une génération automatisée des contraintes pour guider le clustering. La démarche proposée se déroule en plusieurs étapes :

1. Raisonnement à base de logiques de description pour une classification sémantique des données ;
2. Génération automatisée des contraintes à partir des données étiquetées sémantiquement ;
3. Clustering guidé par les contraintes générées ;
4. Capitalisation des résultats et étiquetage sémantique des clusters.

Nous allons détailler dans la suite de cette section chaque étape de l'approche. La connaissance que nous considérons est reprśentée au travers du langage OWL(Group et al., 2009). OWL s'appuie sur les logiques de description pour proposer différents services d'inférence relevant du raisonnement déductif. En plus des éléments descriptifs des concepts, la connaissance experte doit contenir des éléments permettant de réduire le fossé sémantique. Ce fossé, bien connu dans la littérature, est du à la difficulté du passage des valeurs numériques exprimées dans les données à des concepts de haut niveau relevant d'une représentation symbolique. Nous utilisons dans notre ontologie un ensemble d'opérateurs logiques et d'éléments restrictifs comme les intervalles de valeurs ou des conditions existentielles sur des propriétés. Cet ensemble permet de faire le lien entre les concepts et les instances. Comme exemple d'utilisation des éléments du langage OWL, nous pouvons considerer la définition d'une partie du concept Eau : $Water\_Pixel \equiv Pixel \wedge ((\exists TM4. < 0.05 \wedge \exists ndvi. < 0.01) \vee (\exists TM4. < 0.11 \wedge \exists ndvi. < 0.001))$. Cet exemple illustre l'utilisation des primitives OWL pour exprimer une connaissance riche et formalisée de l'expert.

Une fois la connaissance experte formalisée, la première étape est la projection des instances dans la base des connaissances. Pour ce faire, une représentation OWL des données est nécessaire. Nous avons développé à cet effet un processus de transformation semi-automatisé. Ce processus transforme les données matricielles en instances OWL décrites par les propriétés présentes dans l'ontologie. Nous injectons ensuite ces instances avec la connaissance experte dans la base de connaissances **KB** [1] du raisonneur. Dans la terminologie des logiques de description(Baader, 2003), la connaissance experte est appelée **TBox** [2], et les instances l'**ABox** [3]. L'utilisation du raisonnement à base des logiques de description permet l'exploitation d'un certain nombre de services d'inférence, parmi lesquels on retrouve la réalisation. C'est un service fourni par le raisonneur qui consiste à retrouver pour une instance de l'ABox, le concept

1. KB : Knowledge Base
2. Terminological Box : Assertions sur les concepts et les relations entre ces concepts...
3. Assertional Box : Déclarations concernant les instances et les propriétés de ces instances...

de la TBox le plus précis auquel elle appartient. Cela revient à étiqueter sémantiquement les instances répondant parfaitement aux critères des concepts. A cause de l'hypothèse du monde ouvert adoptée par les logiques de description, et donc par OWL, les instances ne répondant pas parfaitement aux définitions des concepts ne seront pas étiquetés.

Une fois le raisonnement effectué, nous procédons à la génération des contraintes. Les instances appartenant au même concept sont ensuite liées entre elles par des *must-link*, celles appartenant à des concepts différents sont liées par des contraintes *cannot-link*. Ces contraintes sont utilisées par la suite pour alimenter le clustering opéré sur l'ensemble des données. Dans notre approche, la génération des contraintes est basée sur le raisonnement, le processus proposé est complétement automatisé. Pour éviter que les erreurs liées à la connaissance experte ne se propagent dans le clustering, nous utilisons l'algorithme PCKMeans (Basu et al., 2004), qui a l'avantage de tolérer la violation de quelques contraintes. On note k le nombre de clusters, M l'ensemble des contraintes $ml(x_i, x_j)$ générées et C l'ensemble des contraintes $cl(x_i, x_j)$ générées ; Avec $W = w_{ij}$ et $\overline{W} = \overline{w}_{ij}$ les poids attribués respectivement aux contraintes M et C. Le problème du clustering par contraintes PCKmeans est formulé par la minimisation de la fonction objective suivante :

$$R_{pckm} = \frac{1}{2} \sum_{x_i \in \chi} \|x_i - \mu_{li}\|^2 + \sum_{(x_i,x_j) \in M} w_{ij} 1[l_i \neq l_j] + \sum_{(x_i,x_j) \in C} \overline{w}_{ij} 1[l_i = l_j] \quad (1)$$

Où l_i ($l_i \in h_{h=1}^{k}$) est le cluster d'appartenance de l'instance x_i, et où $w_{ij} 1[l_i \neq l_j]$ et $\overline{w}_{ij} 1[l_i = l_j]$ correspondent respectivement aux coûts de la violation des contraintes $ml(x_i, x_j) \in M$ et $cl(x_i, x_j) \in C$. On note aussi que 1 est une fonction ayant comme valeur $1[true] = 1$ et $1[false] = 0$, et que x_i représente l'instance affectée à la partition χ_{li} ayant comme centroid μ_{li}. Une fois le clustering effectué, nous propageons l'étiquetage sémantique des instances obtenu avec le raisonnement à leurs clusters d'appartenance. Ainsi, nous bénéficions de l'induction du clustering pour retrouver les instances des autres clusters. Les clusters sont identifiés sémantiquement et contiennent les instances catégorisées par le raisonnement, mais aussi par le clustering. Notre approche produit aussi des clusters non étiquetés si k est supérieur au nombre de concepts présents dans l'ontologie, ces clusters représentent les classes, non définies par l'expert, identifiées seulement après l'étape du clustering par contraintes.

4 Expérimentations

Nous avons appliqué notre approche à un cas d'étude réel de classification d'images satellites provenant de deux régions différentes du monde(Amazonie, Brésil et Montpellier, France). L'objectif est de mettre en évidence les avantages de l'utilisation de l'ontologie dans notre approche pour l'étiquetage sémantique des pixels sans l'intervention de l'expert et la génération automatisée des contraintes pour guider le clustering. Les données classifiées sont des extraits d'images satellites de type Landsat 5 à 30m de résolution. Chaque image est composée de 7 bandes spectrales et contient 468.000 pixels. Aucun échantillon de pixels étiquetés n'est utilisé dans notre approche. L'ontologie utilisée contient la formalisation de deux concepts thématiques : l'eau et la végétation. Les figures 1 et 2 montrent les résultats obtenus en utilisant la même ontologie sur deux extraits d'images satellites. Tout d'abord, on doit noter qu'à la différence du clustering, les pixels appartenant à l'eau et la végétation sont sémantiquement étiquetés dans notre approche. Dans la figure 1, nous pouvons distinguer visuellement

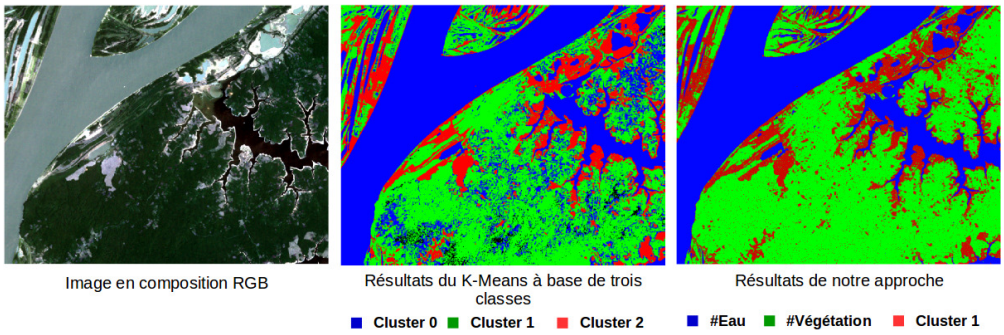

FIG. 1 – *Application de notre approche et d'un K-Means sur une image de la région d'Amazonie, Brésil*

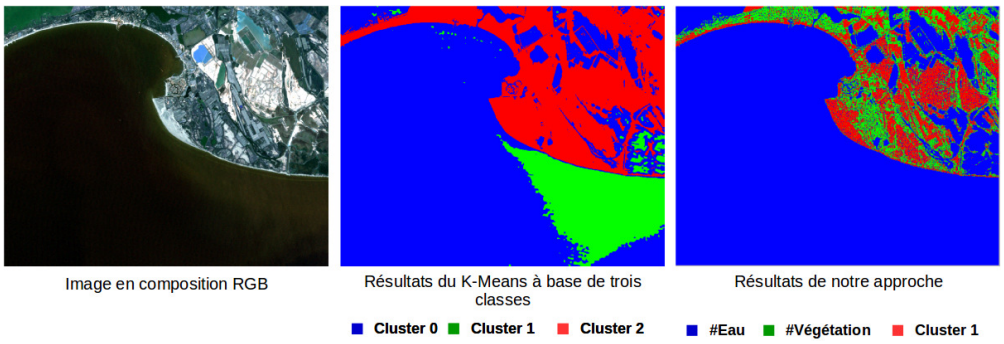

FIG. 2 – *Résultats de K-Means et de notre approche sur un extrait d'image satellite de la région de Montpellier, en utilisant les mêmes connaissances*

une amélioration des résultats par rapport au clustering. La végétation et l'eau sont bien séparés dans notre approche, tandis qu'on peut nettement voir sur les résultats du K-Means une confusion entre plusieurs pixels. Les résultats de la figure 2 montrent quand à eux un aspect très important de notre approche, celui permettant à l'utilisateur de contrôler efficacement la formation des clusters. En effet, nous pouvons remarquer qu'en appliquant un clustering classique à base de trois classes, les résultats de K-Means sont décevants. Les résultats montrent une formation de deux clusters contenant tous les deux des pixels *Eau*, et un troisième cluster contenant tout les autres types de pixels. En appliquant notre approche, nous obtenons de bons résultats reflétant le souhait de l'expert, avec la formation d'un unique cluster d'eau et un autre contenant uniquement de la végétation. Ces résultats montrent que l'injection des contraintes générées à partir de l'ontologie améliore le clustering et automatise la labélisation.

5 Concluion

Nous avons présenté dans cet article une nouvelle approche hybride se basant sur une ontologie comme référentiel de connaissance pour automatiser la génération des contraintes et faciliter l'interprétation sémantique des clusters. Nous avons validé notre approche sur la classification d'images satellites et les résultats obtenus mettent en évidence les apports importants que ce soit au niveau de l'étiquetage sémantique des clusters ou au niveau de l'amélioration de la qualité du clustering. Dans nos futurs travaux, nous prévoyons d'enrichir la connaissance experte dont nous disposons avec de nouveaux concepts, afin d'augmenter le nombre de classes thématiques qui pourraient être détectées par l'ontologie.

Remerciements

Ce travail a été réalisé dans le cadre du projet ANR COCLICO, ANR-12-MONU-0001

Références

Baader, F. (2003). *The description logic handbook : theory, implementation, and applications*. Cambridge university press.

Basu, S., A. Banerjee, et R. J. Mooney (2004). Active semi-supervision for pairwise constrained clustering. In *SDM*, Volume 4, pp. 333–344. SIAM.

Davidson, I. et S. Basu (2007). A survey of clustering with instance level constraints. *ACM Transactions on Knowledge Discovery from Data*, 1–41.

Davidson, I. et S. Ravi (2005). Agglomerative hierarchical clustering with constraints : Theoretical and empirical results. In *Knowledge Discovery in Databases : PKDD 2005*, pp. 59–70. Springer.

Group, W. O. W. et al. (2009). Owl 2 web ontology language document overview.

Shental, N., A. Bar-Hillel, T. Hertz, et D. Weinshall (2004). Computing gaussian mixture models with em using equivalence constraints. *Advances in neural information processing systems 16*(8), 465–472.

Wagstaff, K. et C. Cardie (2000). Clustering with instance-level constraints. *AAAI/IAAI 1097*.

Summary

This paper presents a new hybrid approach for automatic contraints generation based on ontology reasoning. Semi supervised clustering methods have shown their efficiency in incorporating user point of view. However, when dealing with complex data, such as satellite images, obtaining constraints is very expensive and requieres a deep knowledge about domain. In this context, we propose to use ontology reasoning to automate semantic labelling and constraints generation for clustering. We apply our approach to classify satellite images, and results show an improvement of clustering quality and a efficient semantic interpretation of the obtained clusters.