Génération de contraintes pour le clustering à partir d'une ontologie - Application à la classification d'images satellites

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. L'utilisation des connaissances a priori peut fortement améliorer la classification non-supervisée. L'injection de ces connaissances sous forme de contraintes sur les données figure parmi les techniques les plus efficaces de la littérature. Cependant, la génération des contraintes est très coûteuse et demande l'intervention de l'expert ; la sémantique apportée par l'étiquetage de l'expert est aussi perdue dans ce type de techniques, seuls les contraintes sont retenues par le clustering. Dans cet article, nous proposons une nouvelle approche hybride exploitant le raisonnement à base d'ontologie pour générer automatiquement des contraintes permettant de guider et améliorer le clustering. L'utilisation d'une ontologie comme connaissance *a priori* a plusieurs avantages. Elle permet l'interprétation automatisée des connaissances, ajoute de la modularité dans la chaîne de traitement et améliore la qualité du clustering en prenant en compte la vision de l'utilisateur. Pour évaluer notre approche, nous l'avons appliquée à la classification d'images satellites et les résultats obtenus démontrent des améliorations notables à la fois au niveau de la qualité du clustering et au niveau de l'étiquetage sémantique des clusters sans intervention de l'expert.

1 Introduction

Ces dernières années, une grande quantité d'images satellites a été rendue disponible par les gouvernements et différents acteurs publics. L'analyse de ces images peut apporter des réponses tangibles à des problématiques d'ordre environnemental et sociétal. La segmentation est parmi les étapes les plus importantes du processus d'analyse d'images satellites et s'appuie essentiellement sur des algorithmes de clustering. 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 plusieurs cas, ce partitionnement est fortement lié à l'intérêt que porte l'utilisateur

final à certaines données. Deux experts qui n'ont pas le même intérêt thématique évalueront de manière très différente les résultats d'un même clustering. Un autre point important, l'interprétation des résultats –*i.e.* attribuer une sémantique aux clusters dégagés dans un contexte thématique défini, peut s'avérer problématique. Dans ce cadre, l'introduction des connaissances dans le processus d'apprentissage devient primordiale, que ce soit pour guider la phase du clustering ou pour aider à l'interprétation des clusters obtenus.

Par ailleurs, dans le domaine des connaissances, les ontologies ont démontré leur efficacité pour en particulier faciliter l'ancrage des symboles et exprimer des connaissances de plus en plus complexes. Les avancées dans l'interprétation automatisée des connaissances et le raisonnement déductif en permettent des exploitations de plus en plus efficaces. Cependant, la formalisation de la connaissance reste un verrou important, et malgré les importantes avancées en la matière, des concepts restent difficiles voire même impossibles à définir de manière exacte. Dans le contexte des images satellites par exemple, les experts ont plus de facilités à définir des concepts liés à la végétation ou à l'eau que des concepts liés au bâtiment.

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 de données. Les apports d'une démarche sont multiples, et permettent de pallier le manque et la relative incertitude des connaissances ainsi que de guider automatiquement le clustering. L'utilisation du raisonnement exploitera la connaissance disponible en étiquetant les données avec les concepts définis, les résultats sont ensuite transmis pour générer automatiquement les contraintes pour le clustering. Le clustering vient ensuite classer les données en plusieurs groupes homogènes. Ainsi, les instances qui n'ont pu être classées par le raisonnement par manque de connaissance – du fait de la complexité des concepts ou d'une définition n'englobant pas toutes les instances – sont classées par induction avec le clustering. Les points forts d'une telle approche sont multiples :

— La démarche est générique et peut être adaptée à tout domaine tirant parti de données expérimentales quantitatives ;
— Les experts du domaine ne sont sollicités que pour la construction du modèle des connaissances en amont de l'approche ;
— Les contraintes sont générées automatiquement ;
— La démarche permet de pallier l'incomplétude et l'incertitude des connaissances ;
— La clustering est adapté automatiquement pour s'approcher le plus possible de la vision de l'expert.

2 État de l'art

Plusieurs travaux ont été menés pour bénéficier des connaissances du domaine. On retrouve dans la littérature des approches orientées vers des systèmes à base de connaissances et des approches à base de contraintes appliquées au clustering. Bien que partageant le souhait de bénéficier des connaissances disponibles, ces approches abordent différemment la problématique d'intégration des connaissances. Ils interviennent le plus souvent à des niveaux différents dans le processus d'extraction des connaissances et ne partagent pas les mêmes objectifs. Une différence essentielle entre ces deux types d'approche réside dans le type de raisonnement

effectué. La plupart des systèmes à base de connaissances s'appuient essentiellement sur un processus déductif, tandis que les systèmes d'apprentissage artificiel se basent essentiellement sur un processus inductif.

2.1 Systèmes à base de connaissances

Les travaux menés dans le cadre des systèmes à base de connaissances (Andres et al., 2012; Falomir et al., 2011; Forestier et al., 2012) ont montré l'efficacité de ces systèmes pour réduire le fossé sémantique et fournir une interprétation de haut niveau à partir des connaissances expertes. Les approches proposées dans la littérature s'intéressent en général à l'interprétation des résultats du clustering à partir des connaissances ontologiques et utilisent rarement des systèmes à base de connaissances pour guider directement le clustering.
Forestier et al. (Forestier et al., 2012) ont proposé une méthode pour étiqueter les objets d'une image satellite avec les concepts d'une ontologie. Ces objets sont obtenus à partir d'une segmentation préalable des images satellites. Le processus de matching proposé se base sur des mesures de similarité entre chaque objet et les concepts de la base de connaissances. Un score est attribué par la suite évaluant la similarité entre l'objet et chaque concept de l'ontologie. Le concept ayant le score le plus élevé vient étiqueter l'objet.
Falomir (Falomir et al., 2011) et Andres (Andres et al., 2012) ont utilisé le raisonnement à base des logiques de description pour étiqueter les objets extraits de l'image avec les concepts définis dans l'ontologie. Ces objets sont préalablement extraits à l'aide d'un algorithme de segmentation paramétré manuellement.
Les travaux cités plus haut utilisent de différentes manières la connaissance experte, mais aucune de ces approches n'intervient pendant une phase de clustering ou bien une phase de segmentation. La principale utilisation reste l'interprétation sémantique des objets extraits. Forestier et al. se sont intéressés à l'application de mesures de similarité sur des descripteurs sémantiques d'objets extraits de l'image satellite mais n'ont pas exploité le raisonnement. Falomir et Andres ont, de leur coté, utilisé le raisonnement, mais sur des objets d'images préalablement extraits et toujours pour l'interprétation et non pour guider la phase du clustering. Globalement, peu de travaux ont appliqué le raisonnement aux images satellites et à notre connaissance, aucune approche proposée dans la littérature n'applique le raisonnement pour renforcer directement la phase du clustering.

2.2 Clustering avec des connaissances a priori

L'introduction de la connaissance a priori dans le clustering a fait l'objet de plusieurs travaux (Basu et al., 2004b)(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 : un ensemble réduit de données étiquetées, le nombre ou le volume des clusters attendus, la relation entre différentes instances ou l'importance (ou poids) de chaque variable... Plusieurs propositions sont disponibles pour permettre l'ajout de la supervision dans le clustering. Parmi elles, 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.

Dans le cadre de ce type de clustering, les connaissances sont exprimées sous-forme de liens *must-link* et *cannot-link*. Introduit initialement par Wagstaff (Wagstaff et Cardie, 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. Les contraintes *must-link* sont transitives, cela implique que si deux instances sont liées avec un *must-link*, et que d_j est en *must-link* avec une autre instance d_l , alors d_i et d_l sont liées aussi. De la même manière, si d_j est lié par un *cannot-link* à cette instance d_l, alors d_i est en *cannot-link* avec d_l.

Pour prendre en compte les contraintes, des modifications sont apportées aux algorithmes de clustering. Différentes techniques ont été proposées dans la littérature, notamment :

— Modification de la phase de mise à jour de l'affectation des instances aux clusters (Wagstaff et al., 2001; 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., 2004b).

L'algorithme COP-KMEANS (Wagstaff et al., 2001) est le premier algorithme qui a été proposé pour intégrer les contraintes. Une vérification a été ajoutée lors de la phase d'affectation. Son rôle est de s'assurer qu'aucune contrainte ne sera violée par l'ajout de l'instance au cluster. Ainsi, lors de l'affectation d'une instance d_i au cluster le plus proche C_i n'est validée que si toutes les contraintes déclarées sont satisfaites. Cela veut dire que l'algorithme vérifie à la fois qu'il n'existe pas une instance d_j liée par un *must-link* à d_i, et qui est déjà assignée à un autre cluster C_j, et qu'il n'existe aucune instance liée par un *cannot-link* avec d_i dans le cluster C_i. Si ces deux conditions ne sont pas satisfaites, l'algorithme passe au cluster suivant, jusqu'à la vérification de toutes les contraintes.
La seconde technique consiste à apporter une modification dans la phase d'initialisation de l'algorithme. Dans la variante du clustering ascendant hiérarchique par contraintes proposée par Davidson et Ravi (Davidson et Ravi, 2005), des fermetures transitives sont calculées à partir des contraintes pour produire des composants d'instances connectées qui vont être injectées dans l'algorithme par la suite.
Les algorithmes présentés par (Wagstaff et al., 2001; Davidson et Ravi, 2005; Shental et al., 2004) ont montré l'amélioration qui peut être apportée par l'utilisation des contraintes pour guider le clustering. Cependant, ces variantes adoptent une approche dite *strict enforcement* (ou *hard constrained*) indiquant que l'algorithme doit trouver le clustering respectant **toutes** les contraintes. Cela rend ces algorithmes très sensibles aux bruits et pose des problèmes de faisabilité en présence de contraintes incohérentes. Des expérimentations menées par Davidson et al. (Davidson et al., 2006) montrent même une dégradation des résultats dans certains cas.
D'autres travaux ont proposé des algorithmes tolérant la violation de quelques contraintes. Ces approches, dites *partial enforcement* (ou *soft constrained*), cherchent à trouver le meilleur clustering tout en respectant le maximum de contraintes. La principale idée dans ces approches est de faire en sorte que le clustering satisfasse le plus de contraintes possibles tout en pénalisant le non-respect de celles-ci quand leur faisabilité n'est pas possible ou trop coûteuse. La plupart de ces approches reposent sur une modification de la fonction objective des algorithmes. Les contraintes sont ainsi incorporées en ajoutant un terme de pénalité à la fonction objective. Comme c'est le cas de *PCK-Means* (Basu et al., 2004a) par exemple.
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.

2.2.1 Ontologie et clustering

Des travaux existants dans la littérature ont exploré l'exploitation des ontologies dans le cadre du clustering textuel. Jing (Jing et al., 2006) a proposé une mesure de distance prenant en compte les corrélations entre les termes en se basant sur les connaissances a priori contenues dans l'ontologie. La mesure développée a été implémentée dans le cadre de K-Means et les expérimentations ont montré une amélioration des résultats du clustering. Hotho et al. (Hotho et al., 2002) ont utilisé les ontologies à différents stades du clustering. L'ontologie est d'abord exploitée dans la phase du pré-traitement des données afin de lier les termes au concepts et filtrer les caractéristiques importantes pour le clustering, puis pour proposer des agrégations entre les clusters en se basant sur la hiérarchie des concepts. Ces travaux et d'autres dans le clustering de textes ont démontré l'utilité d'introduire les ontologies comme connaissance a priori. Cependant, comme ils traitent directement de texte, le problème de l'ancrage sémantique qui est essentiel pour exploiter sémantiquement l'information numérique des images n'est pas abordé.

3 Proposition

Nous présentons une méthode permettant d'exploiter les ontologies OWL comme support de connaissances pour guider et renforcer le clustering. 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. Nous proposons ainsi une approche hybride exploitant à la fois le raisonnement déductif et le clustering (inductif) afin de tirer le meilleur parti de ces deux méthodes. Cette approche permet à la fois d'exploiter les connaissances expertes disponibles de manière efficace, et de pallier les manques de cette même connaissance en utilisant le clustering par contraintes. La démarche proposée se déroule en plusieurs étapes :

1. Conceptualisation et formalisation des connaissances expertes et construction de la TBox de l'ontologie ;
2. Projection des données qui viennent peupler l'ABox de l'ontologie ;
3. Raisonnement à base de logiques de description pour une classification sémantique des données ;
4. Génération automatisée des contraintes à partir des données étiquetées sémantiquement ;
5. Clustering guidé par les contraintes générées ;
6. Capitalisation des résultats en exploitant la sémantique associée à chaque cluster.

Comme le montre la figure 1, les étapes s'enchaînent pour garantir une collaboration efficace et une capitalisation des résultats obtenus par le raisonnement et par le clustering. Nous allons

détailler dans la suite de cette section chaque étape de l'approche. Nous illustrerons aussi l'effet produit par chacune d'elles sur les données.

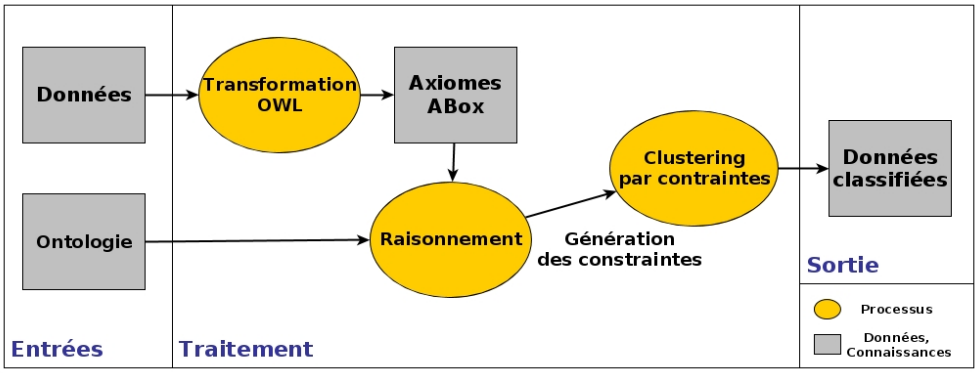

FIG. 1 – *Schéma global de l'approche hybride proposée*

Nous pouvons distinguer dans notre figure une double entrée. La première concerne les données, la deuxième porte sur les connaissances. Notre proposition est une démarche méthodologique valable dans toute problématique qui dispose de données matricielles, et de connaissance experte formalisée associée à ces données. 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. Les travaux menées sur les ontologies ont prouvé leur efficacité pour répondre à ce besoin. OWL propose dans 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.

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. Ce processus semi- automatique analyse les propriétés présentes dans la connaissance et les variables descriptives des données, et propose à l'utilisateur de faire correspondre les deux. Une fois que l'utilisateur indique les correspondances, le processus de transformation projette les données en instances OWL, et attache à chaque instance les propriétés adéquates. Après l'obtention de la représentation OWL des instances, nous injectons ces instances avec la connaissance experte dans la base de connaissances **KB** [1] du raisonneur. Dans la terminologie des logiques de description(Baader,

1. KB : Knowledge Base

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 de la TBox le plus précis auquel elle appartient. Cela revient à étiqueter sémantiquement les instances répondant parfaitement aux critères des concepts. Il faut noter que le langage OWL adopte l'hypothèse du monde ouvert. Ce qui a comme conséquence le non étiquetage de toutes les instances.

La figure 2 (b) montre un exemple de ce à quoi pourrait aboutir ces deux premières étapes. Nous allons utiliser cet exemple pour illustrer la démarche tout au long de cette section. Supposant que nous disposons d'un ensemble de données non-étiquetées, et d'une connaissance qui permet de discriminer les instances de deux concepts C1 et C2. L'objectif est de classifier les données en quatre classes thématiques qui intéressent l'expert. Nous illustrerons dans cet exemple l'apport de notre approche hybride, combinant les connaissances disponibles avec le clustering par contraintes. Aprés le raisonnement, un ensemble d'instances de C1 et de C2 seront identifiées comme le montre la figure 2 (b).

Une fois le raisonnement effectué, on procède à la génération des contraintes. Les instances appartenant au même concept sont ensuite liées entre elles par des *must-link*. Dans notre exemple, cela se traduit par la liaison de toutes les instances de C1 et de C2 par des liens de *must-link* (Figure 2 (c)). Ces *must-link* peuvent être assimilés à la définition de nouvelles contraintes posées sur les instances et vont dans ce sens, servir par la suite à guider le clustering.

A ce stade, nous obtenons au travers du raisonnement et du processus de génération des contraintes un ensemble d'instances étiquetées sémantiquement et liées entre elles. Nous utilisons cet ensemble pour alimenter le clustering par contraintes opéré sur l'ensemble des données. Comme nous l'avons déjà mentionné -section 2.2-, il existe deux variantes dans la prise en compte des contraintes. Une variante *strict enforcement*, et une autre dite *partial enforcement*. 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., 2004a) qui appartient à la variante *partial enforcement*. Si 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}.

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...

clustering par contraintes générées automatiquement d'après une ontologie

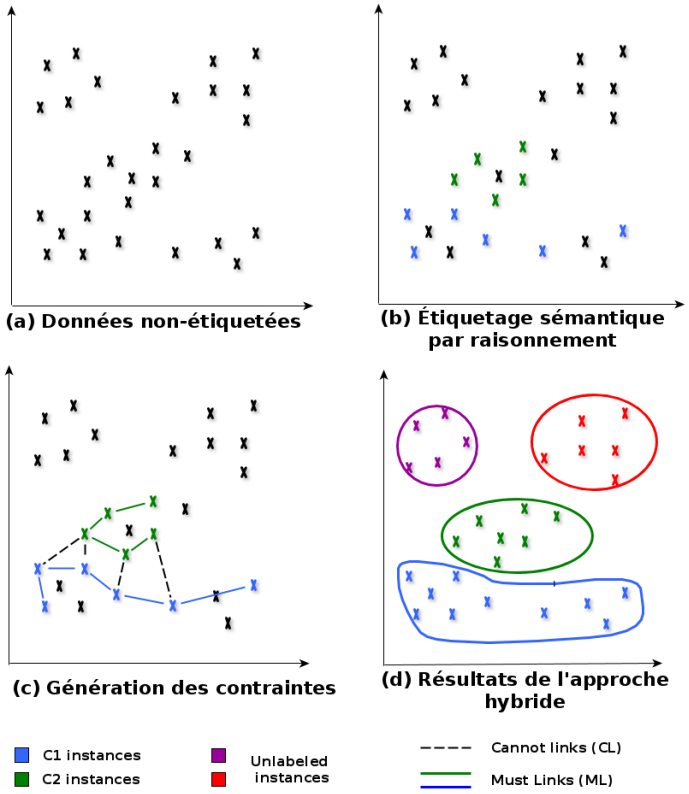

FIG. 2 – *Illustration du déroulement de l'approche sur des données à deux dimensions*

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. La figure 2(d) montre le résultat final obtenu sur notre exemple. Les clusters sont identifiés sémantiquement et contiennent les instances catégorisées par le raisonnement, mais aussi par le clustering. Les clusters non étiquetés représentent les classes identifiées seulement après l'étape du clustering par contraintes .

4 Expérimentations

Cette section décrit l'implémentation mise en place, les données utilisées ainsi que les expérimentations menées et les résultats obtenus. Nous avons appliqué notre approche à un cas d'étude réel de classification d'images satellites.

Les données classifiées sont des extraits d'images satellites de type Landsat 5 à 30m de résolution. Les extraits concernent la région de la rivière de rio Tapajós en Amazonie (Brésil). Chaque image est composée de 7 bandes spectrales et contient 468.000 pixels. Aucun échantillon de pixels étiquetés n'est utilisé dans notre approche. Les seules entrées sont les pixels de

l'image à classifier et la TBox de l'ontologie OWL contenant la formalisation de deux concepts thématiques : l'eau et la végétation.

Plusieurs frameworks ont été utilisés afin d'implémenter la plate-forme de test. Un processus dédié pour les pré-traitements des images satellites et leurs transformations a été mis en place avec la librairie *Orfeo ToolBox*[4]. Coté sémantique, la projection des données en instances OWL est implémentée en Java en s'appuyant sur la librairie *OWL API*. Pellet (Sirin et al., 2007) est le raisonneur OWL utilisé afin de matérialiser le type déduit des pixels à partir des définitions des concepts de l'ontologie, il a été notamment choisi pour son support complet des raisonnements sur les *datatypes*. La génération des contraintes ainsi que l'algorithme de clustering PCKmeans (Basu et al., 2004a) sont aussi implémentés en Java.

4.1 Résultats et discussions

L'objectif des expérimentations menées est de mettre en évidence les avantages de l'exploitation simultanée de l'ontologie pour l'étiquetage sémantique des pixels sans l'intervention d'expert et des contraintes générées automatiquement pour guider le clustering.

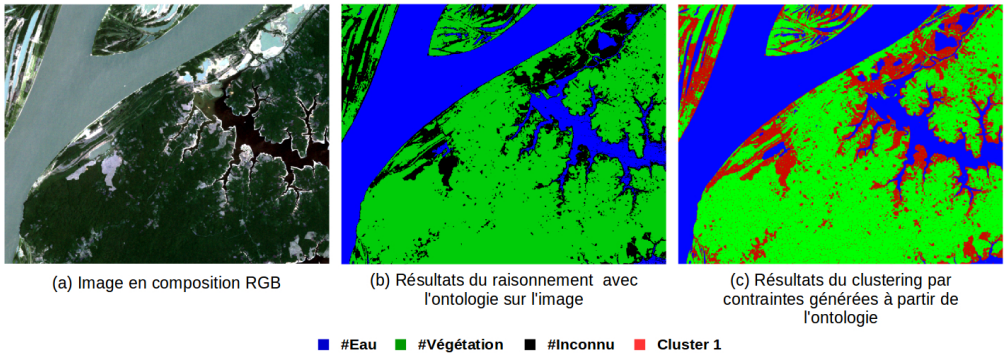

FIG. 3 – *Application de l'approche sur une image de la région d'Amazonie, Brésil*

La figure 3 montre les résultats de l'application de notre approche sur un extrait parmi les images d'Amazonie utilisées dans nos expérimentations. Nous pouvons remarquer visuellement une amélioration des résultats du raisonnement à base d'ontologie avec notre démarche, comme le montre la partie en haut à gauche des figures 3 (a) et 3 (b). Où on peut s'apercevoir que ces pixels n'ont pas été étiquetés par le raisonnement mais par le clustering par contraintes. Nous pouvons aussi observer dans cette figure la combinaison entre des clusters sémantiquement étiquetés par la connaissance experte (#Végétation et #Eau), et d'autres clusters induit par le clustering (Cluster 1). Les expérimentations montrent ainsi la capacité de notre approche à utiliser efficacement les connaissances, même quand ces connaissances sont incomplètes.

4. Orfeo Toolbox : www.orfeo-toolbox.org

clustering par contraintes générées automatiquement d'après une ontologie

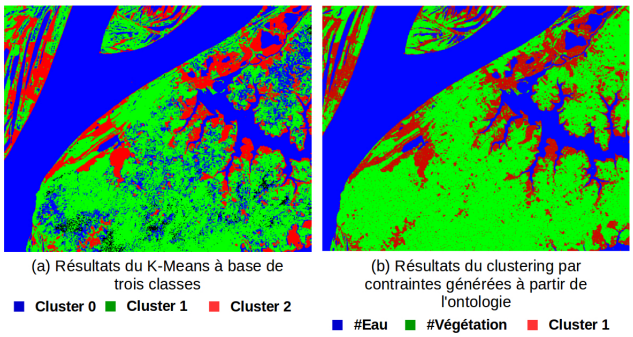

FIG. 4 – *Résultats de K-Means et de notre approche sur l'extrait 2 de la région d'Amazonie, Brésil*

La figure 4 présente les résultats d'un K-Means (fig.4 a) et ceux de notre approche (fig.4 b) sur le même extrait. Ces résultats montrent que l'injection des contraintes générées à partir de l'ontologie améliore le clustering. Dans cet extrait, nous pouvons remarqué que K-Means confond l'eau et la végétation. Selon l'expert, ces erreurs sont dues à la nature de la forêt Amazonienne, où des arbres poussent quelques fois sur des sols très humides. La prise en compte des connaissances dans notre approche permet de remédier à ces problèmes et aide le clustering à bien distinguer l'eau de la végétation.

Extraits	Clustering		Ontologie			Approche proposée	
	Prec.	F-Mes.	% Étiqueté	Prec.	F-Mes.	Prec.	F-Mes.
Extrait 1	0.8899	0.8764	83,6	0.9999	0.9999	**0.9445**	**0.9296**
Extrait 2	0.8701	0.8598	81,26	0.9999	0.9999	**0.9271**	**0.9181**
Extrait 3	0.8889	0.9241	90,33	0.9998	0.9986	**0.9299**	**0.9304**

TAB. 1 – *Résultats des expérimentations sur les extraits de la région d'Amazonie*

Pour évaluer la qualité des résultats obtenus, nous avons calculé la précision et la F-mesure par rapport à une classification de référence fournie par l ?expert. Nous avons aussi comparé notre approche à l ?étiquetage sémantique basé uniquement sur l ?ontologie, et à un clustering à base de K-Means. Notons ici que l ?évaluation de K-Means s ?est effectuée après l ?intervention de l ?expert pour étiqueter manuellement les clusters, ce qui représente une différence importante avec notre approche, où l ?étiquetage des pixels appartenant aux concepts définis dans l ?ontologie se fait automatiquement.

Le tableau présente les mesures obtenues sur les trois extraits fournis par l ?expert. Les valeurs les plus élevées sont obtenues par l ?ontologie. Cependant, ces mesures ne concernent que deux classes et sont calculées uniquement sur les pixels étiquetés (83,6 % pour l ?extrait 1). En effet, l ?ontologie ne permet pas une classification complète de l ?image, d ?où l ?intérêt d ?utiliser le clustering pour compléter la classification. En comparant les résultats obtenus par notre approche à ceux obtenus par K-Means, nous pouvons remarquer une amélioration de la précision et de la F-Mesure sur les différents extraits, ce qui prouve qu ?en plus de l ?apport sémantique, notre approche améliore globalement la qualité du clustering.

5 Concluion

Nous avons présenté dans cet article une nouvelle approche hybride combinant le raisonnement pour l'interprétation automatisée des connaissances expertes et le clustering guidé par des contraintes générées automatiquement à partir d'ontologie. En proposant une approche d'une nature à la fois déductive et inductive, nous avons pu obtenir une méthode insensible à l'incomplétude des connaissances et prenant en compte automatiquement la vision de l'utilisateur. Nous avons validé notre approche sur la classification d'images satellites et les résultats obtenus ont démontré des apports importants que ce soit au niveau l'étiquetage sémantique préalable des clusters ou au niveau de l'amélioration de la qualité du clustering. Dans nos futurs travaux, nous prévoyons d'explorer deux axes afin d'améliorer davantage notre approche. Le premier axe concerne le choix des contraintes à utiliser dans le cadre du clustering. Actuellement, les contraintes sont générées aléatoirement à partir des données étiquetées par l'ontologie. Nous pensons qu'une évaluation du comportement de notre approche suivant les contraintes utilisées nous permettra de définir un critère de choix permettant de sélectionner uniquement les contraintes qui amélioreront les résultats de notre approche. Le deuxième axe correspond à l'enrichissement de la connaissance experte avec de nouveaux concepts, afin d'augmenter le nombre de classes thématiques qui pourraient être détectées par l'ontologie.

Acknowledgements

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

Références

Andres, S., D. Arvor, et C. Pierkot (2012). Towards an ontological approach for classifying remote sensing images. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pp. 825–832. IEEE.

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

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

Basu, S., M. Bilenko, et R. J. Mooney (2004b). A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 59–68. ACM.

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.

Davidson, I., K. L. Wagstaff, et S. Basu (2006). *Measuring constraint-set utility for partitional clustering algorithms*. Springer.

Falomir, Z., E. Jiménez-Ruiz, M. T. Escrig, et L. Museros (2011). Describing images using qualitative models and description logics. *Spatial Cognition & Computation 11*(1), 45–74.

Forestier, G., A. Puissant, C. Wemmert, et P. Gançarski (2012). Knowledge-based region labeling for remote sensing image interpretation. *Computers, Environment and Urban Systems 36*(5), 470–480.

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

Hotho, A., A. Maedche, et S. Staab (2002). Ontology-based text document clustering. *KI 16*(4), 48–54.

Jing, L., L. Zhou, M. K. Ng, et J. Z. Huang (2006). Ontology-based distance measure for text clustering. In *Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*.

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.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur, et Y. Katz (2007). Pellet : A practical owl-dl reasoner. *Web Semantics : science, services and agents on the World Wide Web 5*(2), 51–53.

Wagstaff, K. et C. Cardie (2000). Clustering with instance-level constraints.

Wagstaff, K., C. Cardie, S. Rogers, S. Schrödl, et al. (2001). Constrained k-means clustering with background knowledge. In *ICML*, Volume 1, pp. 577–584.

Summary

Recent studies have shown that the use of *a priori* knowledge can significantly improve the results of unsupervised classification. However, capturing and formatting such knowledge as constraints is not only very expensive requiring the sustained involvement of an expert but it is also very difficult because some valuable information can be lost when it cannot be encoded as constraints.

In this paper, we propose a novel constraint-based clustering approach based on description logics and reasoning for automatically generating constraints from OWL ontology. We apply our approach to classify satellite images. The results have shown that our approach improves the quality of the clustering, while bridging the semantic gap and automating the process of image labeling.