Apports des ontologies à l'analyse exploratoire des images satellitaires

Hatim Chahdi

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **I2S, Information Structures Systèmes**
De l'unité de recherche **UMR Espace-Dev - Université de Montpellier**
Et de l'unité de recherche **UMR LIPN - Université Paris 13**

Spécialité : **Informatique**

Par : **Hatim CHAHDI**

Apports des ontologies à l'analyse exploratoire des images satellitaires

Soutenue le 4 Juillet 2017 devant le jury composé de :

Mme. Nathalie AUSSENAC-GILLES, DR, IRIT - CNRS	Président du jury
Mme. Rosanna VERDE, PR, Univ. della Campania Luigi Vanvitelli	Rapporteur
M. Bernd AMANN, PR, LIP6 - Univ. Pierre et Marie Curie	Rapporteur
M. Mohamed QUAFAFOU, PR, LSIS - Aix-Marseille Univ.	Examinateur
Mme. Laure BERTI-EQUILLE, DR, Espace-Dev - IRD	Directeur de thèse
Mme. Isabelle MOUGENOT, MCF - HDR, Espace-Dev - Univ. de Montpellier	Co-Directeur de thèse
M. Younès BENNANI, PR, LIPN - Université Paris 13	Co-Directeur de thèse
M. Nistor GROZAVU, MCF, LIPN - Université Paris 13	Encadrant de thèse

Résumé

A l'heure actuelle, les images satellites constituent une source d'information incontournable face à de nombreux enjeux environnementaux (déforestation, caractérisation des paysages, aménagement du territoire, etc).

En raison de leur complexité, de leur volume important et des besoins propres à chaque communauté, l'analyse et l'interprétation des images satellites imposent de nouveaux défis aux méthodes de fouille de données.

Le parti-pris de cette thèse est d'explorer de nouvelles approches, que nous situons à mi-chemin entre représentation des connaissances et apprentissage statistique, dans le but de faciliter et d'automatiser l'extraction d'informations pertinentes du contenu de ces images. Nous avons, pour cela, proposé deux nouvelles méthodes qui considèrent les images comme des données quantitatives massives dépourvues de labels sémantiques et qui les traitent en se basant sur les connaissances disponibles.

Notre première contribution est une approche hybride, qui exploite conjointement le raisonnement à base d'ontologie et le clustering semi-supervisé. Le raisonnement permet l'étiquetage sémantique des pixels à partir de connaissances issues du domaine concerné. Les labels générés guident ensuite la tâche de clustering, qui permet de découvrir de nouvelles classes tout en enrichissant l'étiquetage initial.

Notre deuxième contribution procède de manière inverse. Dans un premier temps, l'approche s'appuie sur un clustering topographique pour résumer les données en entrée et réduire de ce fait le nombre de futures instances à traiter par le raisonnement. Celui ci n'est alors appliqué que sur les prototypes résultant du clustering, l'étiquetage est ensuite propagé automatiquement à l'ensemble des données de départ. Dans ce cas, l'importance est portée sur l'optimisation du temps de raisonnement et à son passage à l'échelle.

Nos deux approches ont été testées et évaluées dans le cadre de la classification et de l'interprétation d'images satellites. Les résultats obtenus sont prometteurs et montrent d'une part, que la qualité de la classification peut être améliorée par une prise en compte automatique des connaissances et que l'implication des experts peut être allégée, et d'autre part, que le recours au clustering topographique en amont permet d'éviter le calcul des inférences sur la totalité des pixels de l'image.

Abstract

Satellite images have become a valuable source of information for Earth observation. They are used to address and analyze multiple environmental issues such as landscapes characterization, urban planning or biodiversity conservation to cite a few.

Despite of the large number of existing knowledge extraction techniques, the complexity of satellite images, their large volume, and the specific needs of each scientific community, give rise to new challenges and require the development of new efficient approaches.

In this thesis, we investigate the potential of intelligent combination of knowledge representation systems with statistical learning. Our goal is to develop novel methods which allow automatic analysis of remote sensing images. We elaborate, in this context, two new approaches that consider the images as unlabeled quantitative data and examine the possible use of the available domain knowledge.

Our first contribution is a hybrid approach, that successfully combines ontology-based reasoning and semi-supervised clustering for semantic classification. An inference engine first reasons over the available domain knowledge in order to obtain semantically labeled instances. These instances are then used to generate constraints that will guide and enhance the clustering. In this way, our method allows the improvement of the labeling of existing classes while discovering new ones.

Our second contribution focuses on scaling ontology reasoning over large datasets. We propose a two step approach where topological clustering is first applied in order to summarize the data, in term of a set of prototypes, and reduces by this way the number of future instances to be treated by the reasoner. The representative prototypes are then labeled using the ontology and the labels automatically propagated to all the input data.

We applied our methods to the real-word problem of satellite images classification and interpretation and the obtained results are very promising. They showed, on the one hand, that the quality of the classification can be improved by automatic knowledge integration and that the involvement of experts can be reduced. On the other hand, the upstream exploitation of topographic clustering avoids the calculation of the inferences on all the pixels of the image.

Remerciements

En premier lieu, je tiens à remercier ma directrice de thèse Laure BERTI-EQUILLE d'avoir accepté de diriger cette thèse. Elle m'a fourni un cadre de travail idéal qui m'a permis de développer mes idées et de mener à bien mes travaux.

Je remercie très sincèrement ma co-directrive de thèse Isabelle MOUGENOT pour son soutien, son grand investissement et sa constante disponibilité durant toute cette période. J'ai découvert le monde des ontologies pendant mon stage de Master avec elle et c'est grâce à elle que j'ai commencé cette thèse.

Un remerciement spécial à mon co-directeur de thèse Younès BENNANI. Nos échanges m'ont aidé à cerner le domaine d'apprentissage numérique et les défis relatifs au clutering. Il a toujours su m'orienter vers des pistes de recherches pertinentes.

Je remercie également mon encadrant Nistor GROZAVU pour sa contribution importante dans l'encadrement de cette thèse. Ses encouragements m'ont donné la motivation nécessaire pour surmonter mes jours de doute.

Merci à Rosanna VERDE, Bernd AMANN, Nathalie AUSSENAC-GILLES et Mohamed QUAFAFOU pour l'intérêt qu'ils ont porté à mon travail et pour avoir accepté d'évaluer ma thèse.

Merci à tous mes amis et collègues d'Espace-Dev pour leur accueil et la très bonne ambiance au travail.

Merci également à tous mes amis et collègues du LIPN pour leur accueil et leur aide. Ils ont rendu mes séjours à Paris très agréables.

Merci à toute l'équipe du département informatique de l'université de Montpellier pour leur accueil et leur confiance. J'ai pris beaucoup de plaisir à enseigner à leurs cotés.

Un grand Merci également à Maha pour sa présence, son soutien et ses encouragements durant toutes ces années.

Finalement, je tiens à remercier du plus profond de mon coeur mes parents et toute ma famille. Merci pour votre infaillible soutien, je vous dois tout. Je vous aime.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Le premier chapitre se consacre à présenter le cadre institutionnel de la thèse et à introduire les objectifs généraux qui ont motivé les travaux de recherche que nous avons menés. Nous donnons également une vue d'ensemble des chapitres qui composent le mémoire.

Sommaire

"The only true wisdom is in knowing you know nothing",
Socrates

1.1 Motivations et objectifs de la thèse

1.1.1 Contexte

Les images satellites se sont imposées comme des données incontournables pour l'observation de la Terre depuis les dernières décennies. Plusieurs programmes gouvernementaux et industriels ont consisté à placer en orbite des satellites permettant l'acquisition d'images à différentes résolutions spatiales et spectrales et répétivités temporelles. De manière complémentaire, des portails Web voire des infrastructures de données spatiales ont été mis en place. Ces systèmes sont à destination d'usagers finaux et donc des communautés scientifiques pour leur faciliter la découverte, la consultation et l'accès aux images issues de différents satellites, parfois de manière gracieuse.

L'ensemble des dispositifs rendus disponibles a donc favorisé l'adoption des images satellites comme des sources de données à même d'apporter de nouveaux savoirs dans l'étude de différentes problématiques environnementales, sociétales et sanitaires.

Chaque communauté ayant des besoins différents, l'acquisition de l'image n'est qu'une première étape, son intérêt thématique ne peut se révéler qu'après une analyse et une interprétation précise de l'information codifiée numériquement sous forme de zéro et un. En effet, la valeur de ces énormes quantités de données acquises par les satellites réside dans les connaissances qui en sont extraites.

1.1.2 Problématique

Actuellement, le processus d'extraction des connaissances repose en grande partie sur des tâches effectuées manuellement par les experts. Donner un sens à ces données quantitatives constitue aujourd'hui le goulot d'étranglement empêchant l'exploitation de tout le potentiel qu'offrent ces différentes images satellites acquises tout autour de la terre par des centaines de satellites.

À cet effet, les techniques d'analyse de données et de représentation des connaissances peuvent jouer un rôle facilitateur et permettre l'automatisation de l'analyse et de l'interprétation de ces images. Cependant la complexité inhérente aux images et à leur interprétation, le coût très important de la constitution des jeux de données d'apprentissage et la volumétrie des images imposent de nouveaux défis aux méthodes classiques de fouille de données.

En analyse exploratoire des données, le *clustering* ou encore classification non supervisée permet de partitionner de gros volumes de données non-étiquetées (à savoir dépourvues de labels sémantiques visant à les typer) en des sous-groupes homogènes au regard de leurs similarités. L'objectif est alors la découverte de la structure sous-jacente des données pour en extraire de nouvelles connaissances. Un des intérêts du *clustering* par rapport à de la classification dite supervisée est de s'affranchir d'un travail

préalable de constitution d'un jeu de données de référence qui est chronophage et qui ne peut être mené que par un expert.

Cependant, l'analyse souhaitée d'une image satellite dépend fortement du besoin des usagers finaux. Pour la même image, deux experts-thématiciens jugeront d'une manière radicalement différente les résultats d'une même activité de *clustering*. Prenons l'exemple de deux thématiciens, le premier est intéressé par les différents types de végétation retrouvés dans l'image et le deuxième par l'évolution des surfaces bâties. Leurs attentes sont en conséquence complètement différentes. L'interprétation du visuel obtenu en sortie du *clustering*, nécessite donc également l'implication de l'expert ayant une connaissance précise des classes thématiques recherchées, pour identifier la sémantique de chaque *cluster* obtenu et lui attribuer une éventuelle classe d'appartenance. Le travail d'expertise est donc non plus en amont de l'apprentissage mais en aval et bien souvent réalisé par l'expert qui a commandité le clustering. Dans ce cadre, il nous semble d'importance de prendre en charge les connaissances du domaine dès l'étape de *clustering*, afin de faciliter voire de s'affanchir du travail d'expertise à mener sur les résultats de la fouille de données.

L'émergence du Web sémantique a contribué à faire évoluer la recherche autour des systèmes de représentation de connaissances. Les principales avancées ont permis de dégager des standards de représentation de connaissances et de développer les raisonneurs capables d'en automatiser l'interprétation.

Ces développements permettent aujourd'hui d'envisager les ontologies comme des supports à même de permettre de formaliser des connaissances de plus en plus complexes et en particulier de faciliter le passage de l'image satellite en tant qu'agrégat de données numériques aux concepts thématiques du domaine de la télédétection.

Cependant, l'acquisition et la formalisation des connaissances d'une image reste une tâche délicate et difficile. En dépit des avancées rapides en la matière, les ontologies construites présenteront toujours des incomplétudes en raison de la difficulté à lever toute ambiguïté dans la définition des concepts à expliciter. Cela n'enlève en rien à l'intérêt de l'utilisation des ontologies, mais mène à réfléchir aux moyens de pallier cette incomplétude dans l'analyse et l'interprétation des données et des connaissances.

A partir de ces constats, nous explorons dans notre travail de thèse deux axes que nous croyons essentiels à l'amélioration du processus général d'extraction de connaissances. Le premier concerne l'automatisation de l'interprétation des connaissances expertes pour la réduction du fossé sémantique [HARE et al., 2006] entre les concepts de haut niveau et les données numériques représentant les images satellites, tout en introduisant de la modularité dans la prise en compte de la vision de l'expert par une séparation des connaissances des parties procédurales. Le deuxième axe concerne l'utilisation des ontologies pour guider et améliorer le processus d'analyse exploratoire des données, notamment au travers l'utilisation de techniques de *clustering* semi-supervisées ainsi que leur exploitation pour

pallier le manque des connaissances au niveau des concepts formalisés de l'ontologie.

La problématique principale étudiée par nos travaux est l'analyse et l'interprétation automatisée de données quantitatives non-étiquetées en présence des connaissances expertes, même incomplètes. Répondre à cette problématique complexe exige l'élaboration d'approches capables d'intervenir à plusieurs niveaux du processus d'extraction et d'interprétation des connaissances et de répondre aux verrous sous-jacents suivants :

— Modélisation des connaissances pour la réduction du fossé sémantique ;

— Intégration et prise en compte des connaissances formalisées dans le cadre du *clustering* ;

— Gestion de l'incomplétude des connaissances ;

— Automatisation de l'interprétation des données en l'absence d'exemples étiquetés.

1.1.3 Objectifs

Le parti pris de nos travaux est d'explorer les moyens d'exploiter efficacement les connaissances expertes disponibles en amont du *clustering*, de s'assurer de leur prise en compte pendant le *clustering* tout en permettant la découverte de nouvelles connaissances non-explicitées jusqu'alors et de permettre finalement l'interprétation sémantique des résultats.

Dans cette perspective, notre objectif est de proposer des approches originales alliant à la fois une exploitation automatisée des connaissances formalisées en ontologie et des méthodes de *clustering* prenant en compte la vision de l'utilisateur, pour améliorer et automatiser l'interprétation des images satellites, ou éventuellement d'autres types de données quantitatives.

Il s'agit donc de développer des méthodes hybrides, alliant des mécanismes de raisonnement déductifs et des mécanismes de raisonnement inductifs pour un apprentissage semi-supervisé à partir des données.

1.2 Organisation du mémoire

La bonne compréhension des travaux présentés nécessite la connaissance des grands principes relevant de l'ingénierie des connaissances et de l'apprentissage artificiel. Les quatre premiers chapitres, qui constituent la partie état de l'art de la thèse, vont s'attarder à introduire dans un premier temps ces sous-disciplines de l'intelligence artificielle concernées par nos axes de recherche, puis à détailler les travaux connexes à notre sujet.

Le deuxième chapitre commence ainsi par présenter les ontologies en tant que moyens de modélisation et de formalisation des connaissances. Les

différents fragments des logiques de description qui sont les formalismes de représentation retenus ainsi que les langages de description standardisés du Web sémantique sont présentés et les services d'inférence offerts par les raisonneurs associés illustrés. Des généralités sur l'apprentissage artificiel ainsi que sur différentes techniques de *clustering* sont introduites dans le troisième chapitre. Les limites de la classification non-supervisée sont aussi détaillées et discutées dans au sein de ce même chapitre et l'intérêt de la prise en compte des connaissances y est souligné. Les spécificités des images satellites seront aussi présentées dans un chapitre dédié afin de mieux rapprocher le lecteur des difficultés propres à leurs analyse et interprétation. Le quatrième chapitre, nommé travaux connexes, passe en revue les travaux disponibles dans la littérature liés à notre problématique de recherche. Il spécifie également les limites des approches existantes et met en exergue leurs différences avec nos propositions.

Le cinquième chapitre, intitulé **Classification à base d'ontologie et de *clustering* semi-supervisé** [CHAHDI et al., 2016c ; CHAHDI et al., 2016b ; CHAHDI et al., 2016a] détaille la première contribution de notre thèse. Nous y introduisons une nouvelle approche hybride exploitant le raisonnement à base d'ontologie, pour générer automatiquement des contraintes permettant de guider et améliorer le *clustering*. Nous montrons que l'utilisation mutuelle d'une ontologie comme connaissance a priori et des contraintes sur le *clustering* offre plusieurs avantages. En permettant l'interprétation automatisée des connaissances, notre approche ajoute de la modularité dans la chaîne de traitement et améliore la qualité du *clustering*. L'utilisation du *clustering* semi-supervisé dans notre approche permet aussi de compléter les connaissances et découvrir de nouveaux concepts, tout en respectant la vision de l'utilisateur. Nous présentons, dans la suite du chapitre, les expérimentations menées pour évaluer notre approche, dans le cadre de la classification d'images satellites. Les résultats obtenus montrent 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 et sans l'utilisation de données étiquetées.

Le sixième chapitre présente notre deuxième contribution [CHAHDI et al., 2016d ; CHAHDI et al., 2014]. En nous appuyant sur les caractéristiques des cartes topologiques SOM, nous proposons une approche originale permettant le passage à l'échelle du raisonnement à base des logiques de description. Notre méthode s'articule autour de deux étapes. La première étape s'appuie sur les cartes de Kohonen pour réduire la taille des données à étiqueter par raisonnement, les données d'entrées (ABox de l'ontologie) sont représentées par les prototypes des neurones, résultats du *clustering*. La deuxième étape de notre proposition consiste à étiqueter les prototypes en utilisant le raisonnement à base des logiques de description, et à propager cet étiquetage aux instances appartenant aux neurones des prototypes labélisés. Nous appliquons notre approche au problème de classification d'images satellites. Les résultats ont montré l'apport de l'approche, à la fois pour l'étiquetage sémantique des cartes SOM et pour le passage à l'échelle et l'optimisation du raisonnement sur de grandes quantités de données.

Finalement, le dernier chapitre du manuscrit résume les différentes propositions de nos travaux de thèse. Il dresse, dans ce sens, un bilan de nos contributions et discute les possibles améliorations qui peuvent y être apportées. Nous concluons le chapitre par la présentation de pistes de recherche, que nous pensons intéressantes à explorer afin de bénéficier pleinement de ces deux champs de recherche que sont l'ingénierie des connaissances et l'apprentissage semi-supervisé.

1.3 Cadre institutionnel

Une présentation du projet et des structures d'accueils qui ont permis le bon déroulement des travaux de recherche nous semble nécessaire.

1.3.1 COCLICO

La thèse s'intègre dans le cadre du projet COCLICO [1] (**CO**llaboration, **CL**assification, **I**ncrémentalité et **Co**nnaissances). COCLICO est un projet de recherche financé par l'ANR [2] qui s'achève en 2017 et qui s'est consacré à la définition de méthodes génériques à même de permettre une analyse multi-échelle de grands volumes de données spatio-temporelles de qualité très variable. L'idée générale était de mettre en œuvre une approche multistratégie et incrémentale dans laquelle la collaboration entre les différentes méthodes de fouille de données serait guidée par des connaissances. Ces connaissances déclinées sous la forme d'ontologies concernent à la fois le domaine thématique (à l'exemple des géosciences ou de la géographie) et le domaine de l'analyse (connaissances sur les méthodes d'apprentissage) afin de garantir des résultats de qualité prenant en compte à la fois la qualité des données et celles des connaissances.

La travaux présentés dans ce mémoire ont été menés au sein de l'équipe **MICADO** de l'UMR [3] **ESPACE-DEV** et de l'équipe **A3** de l'UMR **LIPN**.

1.3.2 UMR ESPACE-DEV

L'UMR **ESPACE-DEV** développe ses recherches sur les dynamiques spatiales caractérisant les éco-socio-systèmes. Ses objectifs concernent la définition d'indicateurs de ces dynamiques : bio-géophysiques, évolutions des sociétés, risques liés aux maladies émergentes en fonction de paramètres environnementaux, changements et vulnérabilité des territoires aux changements globaux. À cet effet, elle met au point des méthodologies en télédétection spatiale et en intégration des connaissances multidisciplinaires.

L'équipe **MICADO** est spécialisée dans la modélisation, l'analyse, le contrôle et la validation des systèmes spatialisées complexes (non stationnaires, non linéaires...) afin de répondre aux problématiques de l'UMR et de fournir des outils permettant de caractériser et suivre les éco-socio-systèmes

1. COCLICO : http ://icube-coclico.unistra.fr
2. ANR : Agence Nationale de Recherche
3. UMR : Unité Mixte de Recherche

étudiés en utilisant l'information apportée par les images satellites. Il s'agit notamment d'apporter des contributions fondamentales dans le domaine des données spatialisées et des approches symboliques et numériques. L'équipe exploite notamment les thématiques liées aux ontologies comme leviers favorisant l'interdisciplinarité au sein de l'UMR **ESPACE-DEV**.

1.3.3 UMR LIPN

L'UMR **LIPN** est de son coté une unité de recherche associant l'Université Paris XIII au CNRS [4]. Le LIPN poursuit des recherches théoriques et appliquées en combinatoire, optimisation combinatoire, algorithmique, logique, génie logiciel, langage naturel et apprentissage.

Au sein du LIPN, l'équipe *Apprentissage Artificiel et Applications* **A3**, traite les problèmes liés à l'apprentissage numérique et symbolique par des méthodes supervisées, non supervisées ou hybrides. Ces recherches sont alimentées, coordonnées et évaluées grâce à diverses applications dans les domaines de la fouille de données et de la reconnaissance des formes.

1.3.4 Co-Encadrement de la thèse

Cette thèse a bénéficié d'un co-encadrement entre l'équipe MICADO d'ESPACE-DEV et l'équipe A3 du LIPN avec un hébergement principal au sein de l'UMR ESPACE-DEV. Elle a été supervisée par **Laure Berti-Equille** et **Isabelle Mougenot** du coté d'ESPACE-DEV et de **Younès Bennani** et **Nistor Grozavu** du coté du LIPN.

Les compétences complémentaires des deux équipes ont permis le bon déroulement des travaux et une compréhension commune de la problématique étudiée et notamment des approches hybrides à définir et à mettre en œuvre.

Les activités autour du pré-traitement des images, de la formalisation des connaissances et de la définition de l'ossature générale des approches hybrides ont en particulier été menées sur le long terme dans les locaux d'ESPACE-DEV.

Tout au long de notre thèse, des déplacements réguliers ont été effectués au LIPN, avec en particulier trois séjours d'une durée de deux mois chacun. Ces séjours ont permis d'approfondir nos connaissances en fouille de données et d'enrichir nos travaux par des réflexions autour des méthodes d'apprentissage les plus à même d'être combinées à de la représentation de connaissances.

En complément des compétences propres à chaque équipe, l'étroite collaboration entre nos différents encadrants a été un élément prépondérant dans la bonne orientation des travaux présentés dans ce manuscrit. Elle a permis l'élaboration d'approches pluridisciplinaires et de leur justes positionnements scientifiques. Cette collaboration a également donné lieu à plusieurs publications dans des conférences nationales et internationales.

4. CNRS : Centre National de la Recherche Scientifique

Chapitre 2

Ontologie et Logiques de Description

Dans ce chapitre, nous définissons les ontologies globalement envisagées comme des supports à la connaissance. Nous introduisons également les logiques de description qui sont retenues dans notre travail en tant que langages de formalisation des ontologies. Une présentation des langages du web sémantique, et notamment du format OWL (Ontology Web Language) qui s'adosse aux logiques de description, est donnée par la suite. Un élément important porte sur l'explicitation des connaissances associées aux images. En ce sens, nous dressons un état de l'art sur le raisonnement comme moyen d'interprétation automatisée des faits de l'image.

Sommaire

"La connaissance des mots conduit à la connaissance des choses", **Platon**

2.1 Introduction

En Science de l'information, une ontologie correspond à un ensemble de termes et de concepts qui vient structurer les connaissances d'un domaine. Gruber [GRUBER, 1993] a ainsi défini une ontologie comme étant une conceptualisation formelle, explicite et partagée d'un domaine. Une ontologie permet ainsi de formaliser les connaissances disponibles en définissant les concepts, les relations et les instances d'un domaine. Elle permet de les représenter explicitement de manière claire et concise et de les formaliser afin de faciliter leur partage et exploitation.

Les connaissances peuvent prendre différentes formes. Elles sont par exemple à la source du savoir et de l'expérience détenus par des thématiciens. Elles sont aussi implicitement codées dans les outils et les logiciels informatiques que manipulent ces experts, et sont parfois obtenues et extraites au moyen d'analyses de données et de capitalisation de résultats d'expérimentation. Cette dispersion et cette diversité des connaissances les rendent difficilement exploitables. Le partage et la formalisation des connaissances revêtent donc une importance capitale. Cette capitalisation des connaissances les rend exploitables sur le long terme et les ouvre à d'autres disciplines. Dans ce cadre, les ontologies constituent un moyen efficace de formalisation du savoir. A cet effet, elles permettent d'expliciter les connaissances acquises par les experts et désambiguïser celles implicitement incluses dans les codes informatiques. Les ontologies peuvent aussi servir de support au partage, à la réutilisation et à la manipulation des connaissances, à la fois par des humains ou par des machines.

Dans cette perspective de formalisation et de partage des connaissances, le domaine de l'ingénierie des connaissances est un sous-domaine de l'informatique qui étudie la mise en œuvre des ontologies au travers l'acquisition des connaissances du domaine, la mise en place de standards et de langages de représentation de ces connaissances et le développement de moyens permettant leur manipulation et leur interprétation. Avec l'avènement du Web sémantique [BERNERS-LEE, HENDLER et LASSILA, 2001], le domaine a connu une forte progression. L'évolution du Web étant fortement liée aux ontologies et au domaine de l'ingénierie des connaissances. Le W3C [1] a ainsi contribué de manière significative à l'établissement de standards de représentation de l'information sémantique.

2.2 Les logiques de description

En ingénierie des connaissances, il existe plusieurs propositions permettant la modélisation et la manipulation des connaissances formelles. On retrouve entre autres les langages à base des frames (ou schémas) [MINSKY, 1975 ; HAYES, 1979 ; REITER, 1980 ; KIFER, LAUSEN et WU, 1995], les graphes conceptuels [SOWA, 1983 ; MUGNIER et CHEIN, 1992 ; WILLE, 1997] et les logiques de description [BAADER, 2003]. Ces solutions ont pour objectif de

1. World Wide Web Consortium : www.w3.org

fournir un moyen de représentation des connaissances et peuvent donc être utilisées pour modéliser les ontologies. Elles ne sont pas totalement disjointes et partagent des principes communs, elles appartiennent toutes à la famille des *systèmes à base de connaissances*.

Dans notre travail, nous nous intéressons aux logiques de description (DL) comme fondements formels de la représentation des connaissances. Elles sont largement adoptées et notamment très utilisées par les communautés scientifiques et industrielles. De plus, leur utilisation par le W3C comme fondements formels du langage ontologique du Web (OWL), dans le cadre du Web sémantique, a fortement contribué aux développements d'implémentations efficaces et fiables de leurs différents fragments.

Dans la suite de cette section, nous allons définir les notions essentielles des logiques des description et introduire ces différents fragments. Nous allons aussi illustrer notre propos par un exemple simple portant sur la description d'une image satellite, que nous allons enrichir au fur et à mesure de la progression de la discussion.

Les logiques de description sont une famille de langages qui correspondent à des fragments décidables de la logique du premier ordre. Elles trouvent leurs origines dans les réseaux sémantiques [QUILLIAN, 1968] et les frames [MINSKY, 1975]. En logiques de description, la représentation d'un domaine se fait au travers une base de connaissances $\mathcal{K}$. Une base de connaissances [2] (Eq. 2.1) est constituée d'un ensemble de concepts (classes) N_C, de rôles (propriétés) N_R et d'individus (instances) N_I.

$$\mathcal{K} = \quad < \mathcal{N_C}, \mathcal{N_R}, \mathcal{N_I} > \tag{2.1}$$

Nous reprenons ici les définitions de Nardi et Brachman [NARDI et BRACHMAN, 2003]. Un concept dénote un ensemble d'individus, et peut être atomique (A par exemple) ou composé (à savoir construit à partir d'un ensemble d'expressions). Un rôle est une relation binaire entre deux entités. Un individu, quant à lui, est une représentation d'une entité réelle du domaine modélisé. La base de connaissance est souvent divisée conceptuellement en deux parties, une partie terminologique et une partie assertionnelle (2.2). La partie terminologique, appelée TBox $\mathcal{T}$, contient les axiomes sur les définitions des concepts et des rôles. La TBox peut être vue comme la partie intensionnelle de la base de connaissances [NARDI et BRACHMAN, 2003]. La partie assertionnelle, ou encore ABox $\mathcal{A}$, contient les affirmations à propos de l'existant, c'est un ensemble d'assertions ou faits du domaine modélisé. Ces assertions peuvent être classées en deux types. Des assertions d'appartenance à un concept, ou des assertions de relation.

$$\mathcal{K} = \quad < \mathcal{T}, \mathcal{A} > \tag{2.2}$$

2. Dans ce mémoire, les termes "base de connaissances" et "ontologie" sont utilisés comme des synonymes

Dans le cadre des logiques de descriptions, la modélisation des connaissances se fait en utilisant des primitives permettant de formaliser et désambiguïser les connaissances. Ces primitives sont dotées d'une sémantique formelle basée sur la logique du premier ordre. Les différents langages correspondent à la combinaison ou l'empilement de différents fragments de cette logique. Chaque fragment ajoute des constructeurs qui apportent plus d'expressivité à la modélisation et offre donc un potentiel de représentation des connaissances accru. Plus les langages sont expressifs, plus la complexité des calculs des conséquences logiques augmente. Tout est donc affaire de compromis pour le modélisateur.

2.2.1 La logique de description minimale $\mathcal{AL}$

La logique $\mathcal{AL}$ (pour *Attribute Language*) est la logique de description dite minimale en raison du choix délibéré de la plus grande sobriété des expressions rendues possibles. Il est entendu qu'aller au delà et se priver de l'un des constructeurs présents dans $\mathcal{AL}$ limiterait par trop l'expressivité de l'ontologie en cours de construction. $\mathcal{AL}$ a été introduite par SCHMIDT-SCHAUSS et SMOLKA (1991), après l'ajout de la négation atomique à la logique $\mathcal{FL}$, proposée initialement par BRACHMAN et LEVESQUE (1984) dans le contexte des langages de frame. Le tableau 2.1 propose un résumé des constructeurs présents dans $\mathcal{AL}$. Il est ainsi possible de définir des concepts atomiques comme $Satellite$, $Image$ et $Radiometrique$ et de la même manière des rôles comme $produit\_par$ et $traitement\_reu$. $\mathcal{AL}$ permet aussi de disposer de la négation sur des concepts atomiques. Par exemple, $(\neg Image)$ désigne le concept de toute chose qui n'est pas image. Les constructeurs permettent aussi de modéliser des concepts plus complexes. On peut exprimer le concept d'image satellite, en employant la quantification universelle, par l'expression $(Image \sqcap \forall produit\_par.Satellite)$ qui serait équivalent à l'expression naturelle : "les images qui n'ont été produites que par des satellites". Finalement, le concept représentant les images ayant subi des traitements [3] peuvent être modélisées à l'aide de la quantification existentielle limitée : $Image \sqcap \exists traitement\_recu.\top$.

On notera que les relations d'équivalence et de subsomption sont prises en charge par défaut par les formalismes des logiques de description [NARDI et BRACHMAN, 2003]. Une équivalence $(C \equiv D)$ indique que les individus de C sont des individus de D et vice versa. On parle de *définition* quand l'équivalence est utilisée avec un concept atomique à sa gauche à l'exemple de :

$(Image\_Satellite \equiv Image \sqcap \forall produit\_par.Satellite)$.

Une subsomption $(C \sqsubseteq D)$ indique que les individus appartenant au concept C appartiennent au concept D. Par exemple $(Image\_Satellite \sqsubseteq Image)$.

3. Le chapitre 4 donnera des explications détaillées sur les spécificités des images satellites

Constructeur	Syntaxe DL
Concept atomique	A
Concept universel	$\top$
Concept vide	$\bot$
Négation atomique	$\neg A$
Conjonction	$C \sqcap D$
Quantification universelle	$\forall R.C$
Quantification existentielle limitée	$\exists R.\top$

TABLE 2.1 – Les constructeurs du langage $\mathcal{AL}$
Avec, A, un concept atomique ; C et D des concepts arbitraires, et R, un rôle atomique

2.2.2 Des logiques de description plus expressives

D'autres logiques peuvent être rendues plus expressives en ajoutant de nouveaux constructeurs à la logique $\mathcal{AL}$. La logique $\mathcal{ALC}$ [4] est ainsi construite en ajoutant la négation complète ($\mathcal{AL}$ permet uniquement la négation de concept atomique). L'intérêt est de disposer des constructeurs de disjonction ($C \sqcup D$) et de quantification existentielle complète ($\exists R.C$). Nous donnons deux exemples d'usage :

— le concept "des images ayant reçu des traitements radiométriques" avec l'expression ($Image \sqcap \exists traitement\_recu.Radiometrique$)
— le concept désignant "les images autres que les images satellites" avec l'expression ($Image \sqcap \neg\forall produit\_par.\ Satellite$).

Le tableau 2.2 recense les différents constructeurs possibles en logiques de description ainsi que les lettres qui les désignent. Le nom de la logique est un mnémonique qui fait état des différents constructeurs offerts par ce fragment de logiques de description [5]. $\mathcal{R}^+$ introduit la transitivité des rôles, $\mathcal{F}$ permet la définition de rôles fonctionnels. La restriction des cardinalités sur les rôles est rendue possible par le fragment $\mathcal{N}$, la restriction qualifiée est, quant à elle, introduite par le fragment $\mathcal{Q}$. $\mathcal{O}$ donne la possibilité d'utiliser les énumérations, $\mathcal{H}$ celle de définir des hiérarchies de rôles atomiques et $\mathcal{R}$ des hiérarchies de rôles non-atomiques. Finalement, le fragment $\mathcal{I}$ permet l'inversion des rôles.

2.2.3 L'interprétation des logiques de description

Tout l'intérêt des logiques de description est de rendre la représentation des connaissances d'un domaine, formelle et partageable, afin d'en dégager de nouveaux savoirs et d'ainsi résoudre des problèmes inhérents au domaine.

4. Les lettres $\mathcal{ALC}$ et $\mathcal{ALCUE}$ dénotent la même logique
5. Le fragment $\mathcal{ALCR}^+$ est parfois indiqué par la lettre $\mathcal{S}$, les deux notations étant équivalentes

Lettre	Constructeur	Syntaxe DL
$\mathcal{C}$	Négation complète	$\neg C$
$\mathcal{U}$	Disjonction	$C \sqcup D$
$\mathcal{E}$	Quantification existentielle complète	$\exists R.C$
$\mathcal{R}^+$	Transitivité des rôles	R^+
$\mathcal{F}$	Fonctionnalité	$\leq 1R$
$\mathcal{N}$	Restriction de cardinalité	$\leq nR ; \geq nR$
$\mathcal{Q}$	Restriction de cardinalité qualifiée	$\leq nR.C ; \geq nR.C$
$\mathcal{O}$	Énumération	$\{a_1 \dots a_n\}$
$\mathcal{H}$	Hiérarchie de rôles atomiques	$R_1 \sqsubseteq R_2$
$\mathcal{R}$	Hiérarchie de rôles non-atomiques	$R_1 \sqsubseteq R_2$
$\mathcal{I}$	Inversion de rôles	R^{-1}

TABLE 2.2 – Les constructeurs possibles en DL
Avec, C et D des concepts arbitraires et R un rôle atomique

Chaque fragment permet ainsi de structurer rigoureusement les connaissances au moyen des constructeurs présentés ci-dessus. La sémantique des connaissances modélisées est ensuite explicitée à l'aide d'interprétations. Une interprétation $\mathcal{I}$ est composée d'un domaine d'interprétation $\Delta^{\mathcal{I}}$ non vide et d'une fonction d'interprétation $.^{\mathcal{I}}$ qui associe chaque concept atomique A à un sous domaine de $\Delta^{\mathcal{I}}$ noté $A^{\mathcal{I}}$, chaque rôle atomique R à une relation binaire $R^{\mathcal{I}} \sqsubseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ et chaque individu a à un élément $a^{\mathcal{I}}$. La fonction d'interprétation est étendue aux constructeurs des concepts par les définitions induites rapportées dans le tableau 2.3, où à chaque constructeur est attaché une sémantique formellement définie.

2.2.4 Les domaines concrets

Les logiques de description permettent de modéliser efficacement des connaissances riches et complexes. Cependant, ces logiques souffrent d'une limitation importante. Elles ne permettent pas d'exprimer des connaissances sur des qualités concrètes. En effet, toutes les connaissances exprimées doivent correspondre à des éléments abstraits. Cela peut parfois poser des problèmes pour modéliser des entités au travers des descriptions quantitatives, telles que le temps, la distance, la température ou simplement des seuils de valeurs. Pour pallier ce manque, BAADER et HANSCHKE (1991) ont proposé d'étendre la logique $\mathcal{ALC}$ avec un domaine concret $_{(\mathcal{D})}$. Un domaine concret consiste en un ensemble d'éléments concrets, tels que l'ensemble des entiers, et un ensemble de prédicats, tels que les relations d'ordre "$<, >$", définis sur l'ensemble des éléments concrets. Les deux ensembles ont une interprétation prédéfinie et unique, et ce dans tous les domaines d'interprétation. L'ajout d'un domaine concret aux logiques de description se fait en ajoutant à la logique de description des constructeurs correspondant aux éléments du domaine concret, ainsi que des rôles permettant d'associer les

Lettre	Constructeur	Interprétation sémantique
$\mathcal{AL}$	$\top$	$\Delta^{\mathcal{I}}$
$\mathcal{AL}$	$\bot$	$\emptyset$
$\mathcal{AL}$	$\neg A$	$\Delta^{\mathcal{I}} \backslash A^{\mathcal{I}}$
$\mathcal{AL}$	$C \sqcap D$	$C^{\mathcal{I}} \sqcap D^{\mathcal{I}}$
$\mathcal{AL}$	$\forall R.C$	$\{a \in \Delta^{\mathcal{I}} \mid \forall b.(a,b) \in R^{\mathcal{I}} \Rightarrow b \in C^{\mathcal{I}}\}$
$\mathcal{AL}$	$\exists R.\top$	$\{a \in \Delta^{\mathcal{I}} \mid \exists b.(a,b) \in R^{\mathcal{I}}\}$
$\mathcal{C}$	$\neg C$	$\Delta^{\mathcal{I}} \backslash A^{\mathcal{I}}$
$\mathcal{U}$	$C \sqcup D$	$C^{\mathcal{I}} \sqcup D^{\mathcal{I}}$
$\mathcal{E}$	$\exists R.C$	$\{a \in \Delta^{\mathcal{I}} \mid \exists b.(a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}$
$\mathcal{R}^+$	R^+	$\bigcup_{n \geq 1} (R^{\mathcal{I}})^n$
$\mathcal{F}$	$\leq 1R$	$\{a \in \Delta^{\mathcal{I}} \mid \mid \{b \mid \in R^{\mathcal{I}}\}\mid \leq 1\}$
	$\leq nR$	$\{a \in \Delta^{\mathcal{I}} \mid \mid \{b \in \Delta^{\mathcal{I}} \mid (a,b) \in R^{\mathcal{I}}\}\mid \leq n\}$
$\mathcal{N}$	$\geq nR$	$\{a \in \Delta^{\mathcal{I}} \mid \mid \{b \in \Delta^{\mathcal{I}} \mid (a,b) \in R^{\mathcal{I}}\}\mid \geq n\}$
	$= nR$	$\{a \in \Delta^{\mathcal{I}} \mid \mid \{b \in \Delta^{\mathcal{I}} \mid (a,b) \in R^{\mathcal{I}}\}\mid = n\}$
	$\leq nR.C$	$\{a \in \Delta^{\mathcal{I}} \mid \mid \{b \in \Delta^{\mathcal{I}} \mid (a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}\mid \leq n\}$
$\mathcal{Q}$	$\geq nR.C$	$\{a \in \Delta^{\mathcal{I}} \mid \mid \{b \in \Delta^{\mathcal{I}} \mid (a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}\mid \geq n\}$
	$= nR.C$	$\{a \in \Delta^{\mathcal{I}} \mid \mid \{b \in \Delta^{\mathcal{I}} \mid (a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}\mid = n\}$
$\mathcal{O}$	$\{a_1 \ldots a_n\}$	$\{a_1^{\mathcal{I}} \ldots a_n^{\mathcal{I}}\}$
$\mathcal{H}$	$R_1 \sqsubseteq R_2$	$R_1^{\mathcal{I}} \sqsubseteq R_2^{\mathcal{I}}$
$\mathcal{R}$	$R_1 \sqsubseteq R_2$	$R_1^{\mathcal{I}} \sqsubseteq R_2^{\mathcal{I}}$
$\mathcal{I}$	R^{-1}	$\{(a,b) \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid (b,a) \in R^{\mathcal{I}}\}$

TABLE 2.3 – L'interprétation des constructeurs de concepts
Avec, A, C et D des concepts et R_i des rôles

valeurs des éléments du domaine concret aux éléments abstrait de la logique de description cible [LUTZ, 2002].

2.3 Les langages de description des données

L'evolution du Web vers une dimension sémantique [BERNERS-LEE, HENDLER et LASSILA, 2001 ; SHADBOLT, BERNERS-LEE et HALL, 2006] a conduit le W3C à établir des standards de données complémentaires permettant d'attacher aux ressources une description et une sémantique interprétables à la fois par les humains et par les ordinateurs. L'idée est de définir un cadre de traitement novateur de l'information, dans lequel les agents logiciels disposent d'une information qui leur est intelligible, et qu'ils vont pouvoir traiter et gérer efficacement. Dans cette perspective, le Web sémantique a naturellement placé les ontologies au centre de la proposition en tant que dispositifs chargés d'apporter cette nouvelle intelligibilité.

Le W3C propose un enrichissement progressif des ressources décrites, au travers de standards de description de données. Ces standards peuvent être vus comme des briques indépendantes, où chaque standard vient répondre à des besoins d'expressivité spécifiques et peut satisfaire ses propres

utilisations. Cependant, ces langages entrent dans le cadre d'une vision globale, chaque langage sert alors de support au langage du niveau supérieur, conduisant à une architecture sous forme d'empilement de couches, appelée "Semantic Web Stack" (figure 2.1). Pour un maximum d'efficacité, les couches doivent répondre à différentes préconisations, actuellement seuls les niveaux du bas de la pile s'étageant jusqu'au langage OWL sont standardisés. XML a été adopté comme standard de base pour la représentation des couches supérieures. RDF établit un cadre général autour de la ressource afin d'en standardiser la description. Le standard RDFS (RDF Schema) vient par la suite introduire quelques éléments fondateurs de modélisation des connaissances au travers en particulier de la notion de schéma et de classe. Finalement, OWL est le langage élaboré pour permettre une représentation ontologique complète des connaissances en se fondant sur les formalismes des logiques de description.

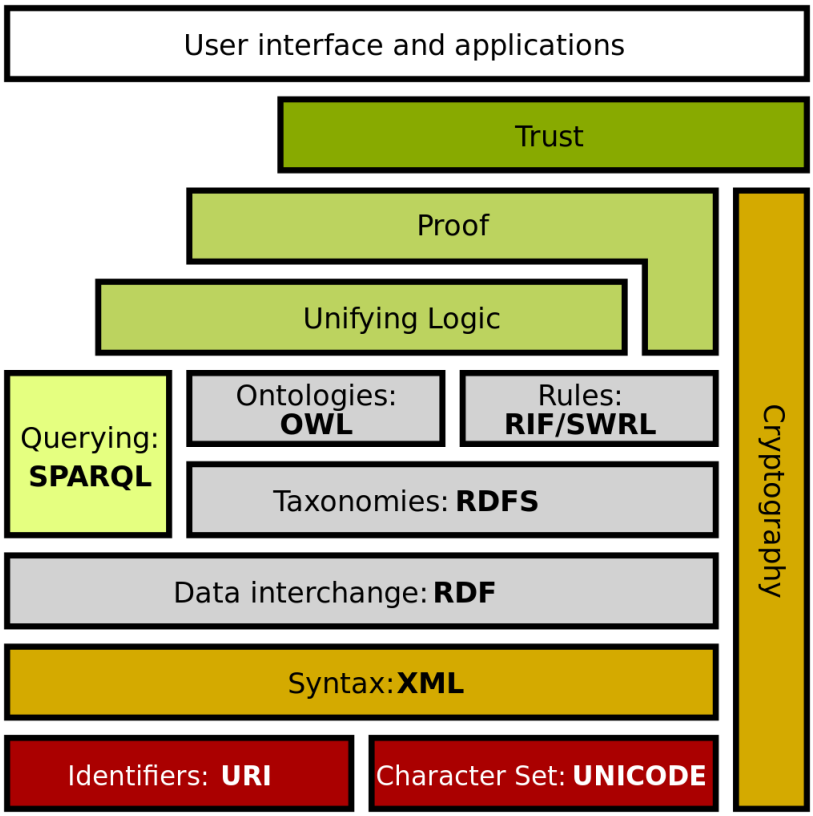


FIGURE 2.1 – Semantic web stack - Architecture en couches du web sémantique [SHADBOLT, BERNERS-LEE et HALL, 2006]

2.3.1 XML : Langage de balisage extensible

XML (Extensible Markup Language) [BRAY et al., 1998; BRAY et al., 2008] est un méta-langage de balisage générique. Il sert donc à définir des vocabulaires à même de structurer l'information d'un domaine cible pour

en faciliter l'échange et le traitement. XML dispose de différents atouts. En particulier, la structuration des données en XML garde une certaine forme de lisibilité pour un humain et surtout est rendue exploitable par les ordinateurs. De plus, tout un chacun est à même de construire son propre vocabulaire avec XML en créant à cet effet ses propres balises. Enfin, les vocabulaires XML sont référencés au travers d'espaces de noms qui les rend interconnectables. Ces différents atouts ont fait du langage XML, un acteur incontournable du Web et un élément de base à l'empilement de couches dans le cadre du Web sémantique (voir figure 2.1). Cette structuration facilitent aussi la définition et la structuration de la communication entre différentes applications. Nous revenons sur la notion d'espace de noms. Afin de lever l'ambiguïté entre différents éléments ou attributs pouvant avoir le même nom mais désigner différentes choses dans deux vocabulaires différents, XML préconise l'utilisation d'URI (Uniform Resource Identifier) [BERNERS-LEE, FIELDING et MASINTER, 2005]. Cette méthode simple permet d'attribuer un nom unique aux ressources, l'URI, qui est défini par la concaténation de l'espace de noms [BRAY, HOLLANDER et LAYMAN, 1999 ; CONSORTIUM, 2006] et du nom local de la ressource. De cette manière, un URI permet d'identifier de manière unique une ressource, deux ressources différentes ne peuvent avoir le même URI. Par contre, deux URIs différentes peuvent référencer la même ressource. Le plus souvent, des préfixes sont utilisés pour remplacer les espaces de noms. Un préfixe est un alias lisible de l'espace des noms qui facilite sa manipulation.

2.3.2 RDF

Le langage RDF (Resource Description Framework) [MANOLA et MILLER, 2004] offre un cadre général pour la description et l'enrichissement de ressources. RDF s'appuie sur la notion de triplet {Sujet, Propriété, Objet} pour décrire toute ressource (sujet) au travers d'un couple propriété/valeur (prédicat/pbjet). Une ressource correspond à toute entité (abstraite ou réelle) référencée par une URI. Le sujet d'un triplet est la ressource à décrire, il peut être une référence URI ou un nœud anonyme [6]. L'objet est l'information donnée sur le sujet, il peut être une ressource, un littéral ou un nœud anonyme. Finalement, le prédicat (propriété) indique la relation, qui existe entre le sujet et l'objet du triplet. Un modèle RDF est une collection de triplets, il peut conceptuellement être vu comme un graphe orienté et étiqueté. Les nœuds sont constitués de l'ensemble des sujets et des objets, les arcs de l'ensemble des prédicats. Différentes syntaxes concrètes sont disponibles pour manipuler un graphe RDF : (RDF/XML) [BECKETT et MCBRIDE, 2004], N3, Turtle, ou encore JSON.

Reprenons l'exemple que nous avons introduit précédemment dans la section des logiques de description (section 2.2). Cette fois, nous allons partir d'une déclaration simple, que nous allons enrichir au fur et à mesure afin

6. Un nœud anonyme est un nœud dépourvu d'URI souvent introduit pour exprimer une notion d'agrégat difficile à modéliser avec la seule notion de propriété binaire.

d'illustrer les primitives mises à disposition par les différents langages de description.

La figure 2.2 représente le triplet décrivant une image, vue comme une ressource identifiée par l'URI `ex:image_1` et décrite par le prédicat `dc:date` et l'objet `13/10/2011`. La figure représente la déclaration RDF sous sa

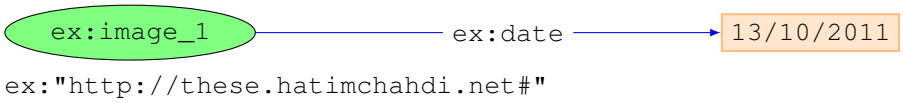


FIGURE 2.2 – Un exemple de triplet RDF

forme graphique (abstraite), elle peut être sérialisée en utilisant l'une des syntaxes mises à disposition à cet effet par le W3C. Nous donnons dans le listing 2.1, sa sérialisation en XML/RDF.

LISTING 2.1 – Exemple d'un document XML

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf=
        "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:ex="http://these.hatimchahdi.net#">
<rdf:Description rdf:about="ex:image_1">
        <ex:date>13/10/2011</ex:date>
</rdf:Description>
</rdf:RDF>
```

2.3.3 RDFS

Le langage RDFS (Resource Description Framework Schema) [BRICKLEY et GUHA, 2004] est un vocabulaire RDF, qui enrichit RDF au travers de différents éléments de modélisation. RDFS introduit notamment la notion de concept `rdfs:Class` (figure 2.3), et permet aussi d'organiser les concepts (classes) et les rôles (propriétés) au travers de hiérarchies. La spécification des domaines de définition et des ensemble d'arrivée des propriétés est aussi prise en charge par le langage. Reprenons la description RDF de notre exemple, nous pouvons déclarer les concepts `ex:Image` et `ex:Satellite` en utilisant `rdfs:Class`. Nous pouvons aussi introduire une nouvelle propriété `ex:date_production` et définir sa relation avec la propriété `dc:date` en utilisant `rdfs:subPropertyOf`. Il est donc rendu possible de modéliser, de manière générale, des ontologies simples avec le langage RDFS.

2.3.4 OWL

Le langage OWL (Web Ontology Language) est un langage conçu dans le but de représenter des connaissances riches et complexes. Il vient s'adosser aux langages RDF et RDFS (figure 2.1) et introduit des constructeurs complexes de concepts et de propriétés. Il emprunte à cet effet les procédés

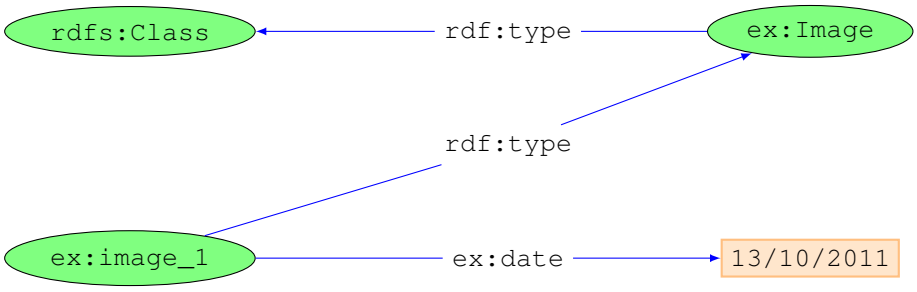


FIGURE 2.3 – Un exemple de triplet RDF enrichi avec le concept Image

de construction des logiques de description (section 2.2), qui représentent les fondements formelles de la sémantique du langage. OWL donne la possibilité d'exprimer l'équivalence, l'union et la disjonction entre concepts et entre rôles. Il introduit également les restrictions, les cardinalités et les concepts énumérés. Le standard actuellement adopté est OWL 2 [GROUP et OTHERS, 2009], il correspond à la logique $\mathcal{SROIQ}_{(\mathcal{D})}$, avec les types de données XML schéma [BIRON et MALHOTRA, 2004], qui viennent en support des domaines concrets. Dans leur présentation de la logique $\mathcal{SROIQ}$, KRÖTZSCH, SIMANCIK et HORROCKS, 2012 précisent la relation qu'elle entretient avec le langage OWL 2, ils montrent que OWL 2 peut être vu comme une variante syntaxique du langage $\mathcal{SROIQ}$, avec quelques différences notables, notamment dans l'adoption d'OWL des types de données XML schéma pour exprimer le domaine concret et l'utilisation d'axiomes de types clés [7] pour faciliter l'intégration de données.

La norme OWL spécifie aussi trois profils [MOTIK et al., 2009] à savoir OWL EL, OWL QL et OWL RL. Ces profils sont moins expressifs que le standard OWL 2 et sont conçus pour répondre à des usages pratiques. Ils offrent différents compromis entre l'expressivité exprimée par l'ontologie d'un coté, et la complexité des mécanismes de raisonnement de l'autre.

2.4 Raisonnement et interprétation automatisée

Les logiques de description permettent de formaliser les connaissances en utilisant les constructeurs introduits par les différents fragments. Chaque constructeur est doté d'une sémantique précise et bien définie. L'objectif du raisonnement est d'inférer des connaissances implicites en calculant les conséquences logiques des connaissances modélisées, et donc explicites. Les raisonneurs représentent les programmes qui implémentent les procédures de calcul des conséquences logiques des connaissances explicitées et

7. Clés OWL : https ://www.w3.org/TR/owl2-syntax/#Keys

permettre ainsi une interprétation automatisée. La figure 2.4 donne une vision "boîte noire" des raisonneurs. Son but est de montrer le fonctionnement des services d'inférence et non d'expliquer son fonctionnement interne.

Un raisonneur prend en entrée les connaissances formalisées (en OWL pour un raisonneur qui exploite ce standard) et calcule les conséquences logiques relatives aux tâches demandées au travers des requêtes. Les mécanismes de raisonnement offrent globalement des fonctionnalités s'articulant sur quatre mécanismes d'inférence :

1. La vérification de la consistance : Il s'agit de vérifier l'absence de faits (déclarations) contradictoires dans l'ensemble de l'ontologie. Elle implique une vérification de la consistance des déclarations au niveau de la TBox et une vérification du respect des instances de l'ABox des règles définis par les concepts de la TBox ;
2. La satisfiabilité d'un concept : Il s'agit de vérifier que pour un concept donné de la TBox, il est possible d'avoir des instances de l'ABox qui vont peupler ce concept ;
3. La classification de l'ontologie : Il s'agit de calculer les relations de « sous-classe » entre les concepts nommées de la TBox pour créer la hiérarchie complète entre ces concepts ;
4. La réalisation : Ce niveau de raisonnement permet de trouver le concept le plus spécifique de la TBox auquel appartient une instance de l'ABox.

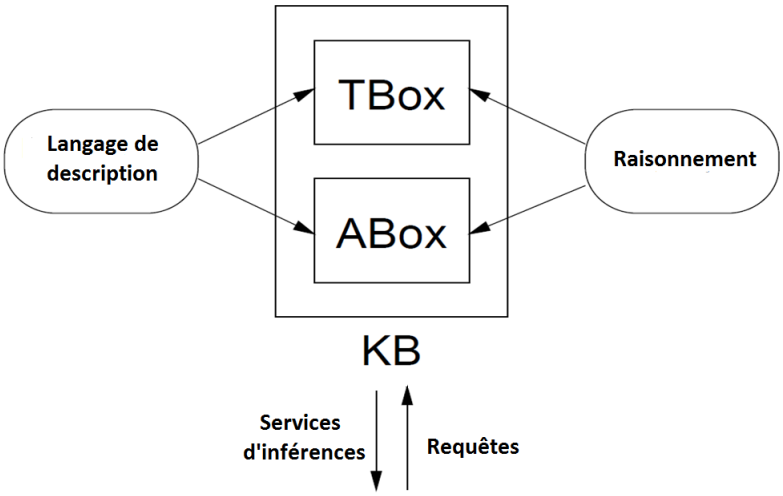


FIGURE 2.4 – Schéma général d'un système à base des logiques de description [BAADER, 2003]

2.5 Résumé

Dans ce chapitre, nous avons présenté les ontologies en tant que moyen de représentation des connaissances d'un domaine. Nous avons également

montré comment les logiques de description et les langages de représentation standardisés RDF, RDFS et OWL permettent une formalisation de ces connaissances. Finalement, nous avons illustré les avantages qu'offre cette formalisation pour inférer de nouvelles connaissances implicites en exploitant le raisonnement déductif.

Ce chapitre a permis de poser les concepts de base permettant la bonne compréhension du premier sous-domaine de l'intelligence artificielle concerné par nos travaux. Dans le prochain chapitre, nous allons introduire l'autre sous-domaine de recherche exploré, à savoir l'apprentissage statistique.

Chapitre 3

Classification à base d'apprentissage

Nos travaux explorent les nouvelles possibilités offertes par une utilisation concertée de la modélisation des connaissances et de l'apprentissage artificiel. Ce chapitre est consacré aux aspects relevant de l'apprentissage et s'attarde, dans un premier temps, sur une introduction générale de la classification des données, pour ensuite explorer plus en détail les méthodes non-supervisées.

Sommaire

"L'aveugle se détourne de la fosse où le clairvoyant se laisse tomber", **Averroès**

3.1 Introduction

À la différence du raisonnement à base d'ontologie, l'apprentissage artificiel s'appuie essentiellement sur un processus inductif qui s'applique à des données souvent en masse et décrites au travers de plusieurs dimensions (variables descriptives).

La spécificité de l'apprentissage réside dans sa capacité à extraire de nouvelles informations en utilisant des modèles statistiques. L'objectif du processus est la généralisation des relations trouvées dans l'échantillon de données analysé à tout l'espace de données du domaine.

L'apprentissage peut être exploité dans des problématiques très variées à titre d'exemple pour la recherche de gènes en biologie moléculaire ou la détection de tendances sur le Web. La catégorisation des données telle qu'elle est permise par l'apprentissage statistique, consiste à attribuer une catégorie (appelée classe ou cluster) à une donnée en raison d'une ressemblance particulière entre les données d'une même catégorie. Pour y parvenir, trois approches peuvent être utilisées : La classification supervisée (*classification* en anglais), la classification non-supervisée (*clustering*) et l'apprentissage semi-supervisé.

3.1.1 Classification supervisée

La *classification*, repose sur l'utilisation d'un échantillon de données étiquetées pour apprendre un modèle qui sera par la suite, utilisé pour déduire la classe d'appartenance (envisagée comme une étiquette) de nouvelles données non étiquetées.

La *classification* s'effectue généralement en deux étapes. La première étape est une étape d'apprentissage, elle consiste d'abord à choisir le type du modèle à apprendre (domaine des hypothèses), puis ensuite à apprendre, à partir de l'ensemble des données étiquetées, les paramètres de ce modèle. Cette étape d'apprentissage peut être itérative et nécessiter plusieurs "reparamétrages" du modèle appris. La deuxième étape est l'étiquetage des données dont la classe n'est pas connue, il s'agit d'évaluer le degré d'appartenance de chaque nouvelle donnée aux différentes classes prédéfinies en utilisant le modèle obtenu par apprentissage. Plus formellement, si on considère un ensemble de N paires d'apprentissage $\mathcal{X}_a = \{(x_i, y_l)\}_{i=1}^{N}$, où $x_i \in \mathbb{R}^d$ est un vecteur qui décrit l'instance x_i au travers de d attributs et $y_l \in Y = \{y_l\}_{l=1}^{k}$ est l'étiquette (label) de cette instance. L'objectif de la *classification* est d'apprendre un modèle $\mathcal{F}$ qui définit la dépendance entre x_i et y_l, afin de pouvoir prédire l'étiquette de nouvelles données $\mathcal{X} = \{x_j\}$ dont l'étiquette n'est pas connue. Le livre de CORNUÉJOLS et MICLET (2011) peut être consulté pour avoir plus de détails sur ce type d'apprentissage.

3.1.2 Classification non-supervisée

Le clustering englobe les techniques d'analyse exploratoire de données [JAIN, MURTY et FLYNN, 1999]. Il désigne l'ensemble de méthodes utilisées

quand il n'existe pas d'informations disponibles a priori sur les données. Le clustering vise à partitionner de gros volumes de données non étiquetées en un ensemble de sous-groupes homogènes au regard de leurs similarités. Si on considère un ensemble de données non-étiquetées $\mathcal{X} = \{x_i\}$, l'objectif est de trouver automatiquement des groupes homogènes $\mathbb{C}_l$, en se basant uniquement sur les attributs des instances et sans connaissance a priori des classes cibles. La plupart du temps, chaque groupe (*cluster*) résultant peut être représenté au travers d'un centroïde (ou prototype), résumant ainsi les caractéristiques communes des données du même groupe. En fonction des algorithmes de clustering et du type des données à analyser, plusieurs critères peuvent être utilisés pour opérer la séparation des données en clusters. D'une manière générale, le but est d'assigner chaque instance $x_i \in \mathcal{X}$ à un ou plusieurs clusters $Cl_i \in \mathbb{C}_l$ avec $i = \{1, .., k\}$, où k est le nombre de clusters. Cela revient à trouver une distribution des données qui maximise la similarité intra-classe et minimise la similarité inter-classe, toujours au regard du critère de similarité choisi.

3.1.3 Apprentissage semi-supervisé

La *classification semi-supervisée* [CHAPELLE, SCHÖLKOPF et ZIEN, 2006] est la plus récente des trois approches. Elle désigne les techniques se situant à mi-chemin entre la classification supervisée et la classification non-supervisée. Elle est généralement utilisée dans deux cas de figure. Le premier cas de figure se présente quand la quantité de données à analyser est très importante par rapport à l'échantillon des données étiquetées disponible. Le deuxième cas de figure se produit quand aucun échantillon étiqueté n'est disponible, mais que d'autres informations a priori (contraintes, taille des clusters, distribution, ...) le sont. Les techniques semi-supervisées peuvent être divisées en deux sous catégories. La première catégorie, appelée *Classification* semi-supervisée, représente l'ensemble des techniques utilisant des données non-étiquetées pour renforcer la phase d'apprentissage. Ces techniques ont un cadre similaire à la classification supervisée, avec une phase d'apprentissage alimentée cette fois par des données étiquetées et des données non étiquetées, et une deuxième phase de prédiction de la classes des données. La deuxième catégorie, appelée *clustering semi-supervisé*, regroupe l'ensemble de techniques ayant un cadre similaire au clustering, avec l'introduction des informations disponibles *a priori* afin d'améliorer le clustering et d'y ajouter de la supervision, sans une phase d'apprentissage préalable.

La classification supervisée est la première approche développée en apprentissage, et son efficacité n'est plus à prouver. Elle possède, cependant, l'inconvénient majeur de rendre obligatoire l'obtention préalable d'un large échantillon de données étiquetées pour l'apprentissage. Avec la multiplication des capteurs, des systèmes de collecte de données et des données publiées sur le Web, l'acquisition d'un tel échantillon de données étiquetées

devient de plus en plus difficile, voire même impossible dans certains cas. Dans le contexte de la télédétection par exemple, l'étiquetage des images satellites demande une intervention lourde d'un expert humain ayant des connaissances approfondies du domaine. Au regard de la taille et de la complexité des images satellites, ce travail peut prendre plusieurs mois à temps complet. C'est pourquoi, les approches non-supervisées et semi-supervisées deviennent aujourd'hui de plus en plus populaires. Elles représentent des alternatives intéressantes pour le traitement automatique de grandes masses de données. Dans la suite du chapitre, nous détaillons en conséquence les techniques de classification non-supervisées.

3.2 Clustering

La classification non-supervisée, *clustering*, regroupe l'ensemble des techniques de catégorisation non-supervisées. Opérant sur des données non étiquetées, le clustering est particulièrement efficace quand il s'agit d'analyser des données sans connaissances a priori sur les classes. À la différence de la *classification*, les catégories cibles ne sont pas connues à l'avance. Les méthodes de clustering se sont révélées particulièrement utiles dans plusieurs domaines d'application. Elles ont notamment été utilisées dans plusieurs pour la compression de données, la réduction de leurs dimensions, la segmentation d'images [JAIN et FLYNN, 1996], la reconnaissance de voix [GÓRRIZ et al., 2006 ; EYBEN, BUCHHOLZ et BRAUNSCHWEILER, 2012] ou encore la catégorisation de clients [WEDEL et KAMAKURA, 2012].

3.2.1 Les approches de clustering

Les techniques de clustering sont diverses et variées, elles peuvent être distinguées au regard de différents critères portant sur le choix de la représentation de données, des techniques de séparation, des modalités de regroupement ou encore du critère de similarité adopté. Ces caractéristiques ont permis la caractérisation des méthodes de clustering en sous-groupes de techniques plus ou moins similaires.

Avant de présenter les différents types de clustering, nous discutons d'abord la notion de similarité, qui est un élément fondamental et déterminant pour tout algorithme de clustering. La mesure de similarité définit comment un algorithme de clustering évalue la proximité entre les instances, et donc la manière avec laquelle il va former les clusters. Elle peut être vue comme une fonction qui quantifie la ressemblance/dissemblance entre deux instances. Suivant la nature du clustering et le type des données à traiter (numériques, catégorielles, binaires ...), les mesures de similarité varient. Quand il s'agit de calculer la similarité entre des données numériques, plusieurs distances existent (voir tableau des distances en Annexe). La distance Euclidienne figure parmi les mesures les plus utilisées, c'est une similarité spécifique aux espaces vectoriels, qui respecte des conditions spécifiques, comme la symétrie et l'inégalité triangulaire. Pour deux instances

x et y, appartenant à $\mathbb{R}^d$, la distance euclidienne peut être définie par la formule suivante :

$$d_E(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{3.1}$$

où x_i, y_i, avec $i \in \{1, ..., d\}$, représentent les valeurs de x et y à la i^{eme} dimension.

Plusieurs catégorisations ont été proposées dans la littérature [JAIN, MURTY et FLYNN, 1999 ; XU et WUNSCH, 2005 ; BERKHIN, 2006]. Nous avons choisi de les présenter suivant ces catégories de méthodes :

— **Méthodes hiérarchiques** [JOHNSON, 1967 ; MURTAGH et CONTRERAS, 2012]
— **Méthodes à base de densité** [ESTER, 2009]
— **Méthodes probabilistes** [MCLACHLAN et BASFORD, 1988]
— **Méthodes à base de graphes** [CRISTIANINI, SHAWE-TAYLOR et KANDOLA, 2001 ; FILIPPONE et al., 2008]
— **Méthodes à base de distance** [XU et WUNSCH, 2005]

que nous allons décrire ci-après.

Méthodes hiérarchiques

Le type de regroupement permet de différencier les méthodes hiérarchiques d'autres méthodes dites plates. Les méthodes hiérarchiques regroupent les données en un ensemble de clusters imbriqués dans d'autres clusters sous une forme arborescente. Deux grandes familles de clustering hiérarchique co-existent : les approches dites ascendantes (ou agglomératives) et les approches dites descendantes (ou divisives).

Pour ce qui concerne les approches ascendantes [GUHA, RASTOGI et SHIM, 1998 ; WARD JR, 1963], l'intuition est de commencer par former n clusters correspondant à l'ensemble des instances, puis de regrouper au fur et à mesure ces clusters, par agglomération progressive des instances en s'aidant d'un critère de similarité, pour en arriver à des clusters plus généraux et de niveau supérieur.

La plupart des algorithmes de classification ascendante hiérarchique se base sur les étapes suivantes :

1. Chaque instance est assignée à un cluster ;
2. La similarité entre chaque cluster est calculé suivant la mesure préalablement choisie ;
3. Les clusters les plus similaires sont fusionnés ;
4. Les étapes précédentes sont répétées jusqu'au regroupement de toutes les instances en un seul cluster.

La différence notable entre les différentes proposition réside dans la mesure de similarité adoptée.

Dans le cadre des approches divisives, le principe est inversé et il est d'usage de considérer que toutes les instances appartiennent à un seul et même cluster qui est ensuite divisé en deux clusters, qui sont eux mêmes divisés à leur tour, toujours sur la base d'une fonction de similarité. La figure 3.1 illustre le principe général d'un algorithme hiérarchique.

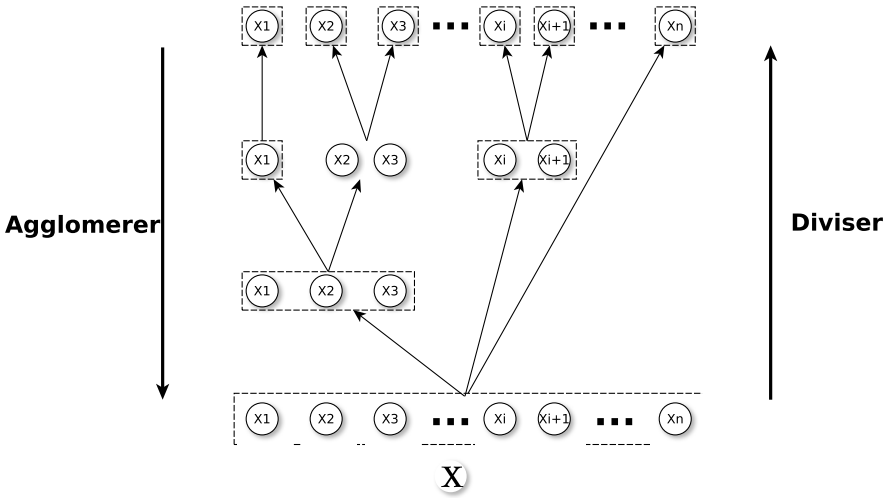


FIGURE 3.1 – Principe général du clustering hiérarchique

Les méthodes de clustering hiérarchique ont l'avantage de proposer une classification des données suivant plusieurs niveaux. Cependant, ce type d'algorithme se heurte à deux principales limitations.

La première est la nécessité de l'intervention de l'utilisateur pour le choix du niveau de coupe dans la hiérarchie des clusters obtenus (figure 3.1). Ce choix reste un problème difficile malgré les différentes propositions existantes.

Le deuxième inconvénient de ces approches est leur grande complexité. Tous les algorithmes de clustering hiérarchique ont une complexité au minimum proportionnelle au carré du nombre des instances ($\mathcal{O}(n^2)$). Cela empêche l'application de ce type de clustering à de grandes masses de données à l'exemple des images satellites.

Méthodes à base de densité

Les méthodes à base de densité [ESTER, 2009] sont des algorithmes qui séparent les données en se basant sur la concentration (ou densité) des instances. Les clusters sont constitués de régions à forte densité et sont séparés entre eux par les régions à faible densité. L'idée sous-jacente à ces méthodes consiste à définir un espace de voisinage (N_{Eps}) qui va servir à calculer la densité de chaque instance des données, puis d'un seuil de voisinage (Eps) minimal à considérer. Plus le voisinage d'une instance est riche en instances, plus cette instance sera considérée comme appartenant au même

cluster. Le voisinage N_{Eps} d'une instance x_i dans un ensemble $\mathcal{X}$ est défini de la manière suivante [ESTER et al., 1996] :

$$N_{Eps}(x_i) = \{x_j \in X \mid dist(x_i, x_j) \leq Eps\} \tag{3.2}$$

Une approche naïve consisterait à séparer les instances en établissant un nombre d'instances minimum $MinPts$ dans le voisinage de chaque instance pour former un cluster. Cependant, cette approche naïve aurait des problèmes à détecter les instances sur les bordures des clusters [ESTER et al., 1996]. L'approche généralement adoptée par les algorithmes repose plutôt sur trois notions importantes pour créer les clusters. Ces notions sont désignées par : densité directement-accessible (*directly density-reachable*), densité-accessible (*density-reachable*) et densité-connectée (*density-connected*).

- Une instance x_i est à une *densité directement accessible* d'une instance x_j si et seulement si, $x_i \in N_{Eps}(x_j)$ et $\mid N_{Eps}(x_j) \mid \geq MinPts$;
- Une instance x_i est à une densité accessible d'une instance x_j si et seulement si une séquence d'instances $\{x_i = x_1, x_2..., x_l = x_j\}$ existe tel que : x_k est à une *densité directement accessible* de x_{k+1} pour $k = 1, ..., l-1$;
- Deux instances x_i et x_j sont à une *densité connectée* par rapport à un seuil de voisinage (Eps) et un voisinage N_{Eps} s'il existe une instance $x_l \in X$ tel que x_i et x_j sont toutes les deux à une *densité accessible* de x_l.

À partir de ces concepts de base, différents algorithmes ont été proposés dans la littérature. DBSCAN [ESTER et al., 1996] est l'une des méthodes les plus connues, et prend comme paramètres d'entrée le seuil de voisinage Eps et le nombre d'instances minimum du voisinage $MinPts$. Appliqué à un ensemble de données non-étiquetées $\mathcal{X}$, l'algorithme forme les clusters $\mathbb{C}_l = \{Cl_1, ..., Cl_k\}$ en maximisant la densité des instances à densité-connectée en satisfaisant les deux conditions suivantes :

- $\forall x_i, x_j$: Si $x_i \in Cl_l$ est à densité-accessible de x_j en respectant Eps et que $MinPts$ est non-vide, alors $x_j \in Cl_l$ (maximisation de la densité) ;
- $\forall (x_i, x_j) \in Cl_l$: x_i est à densité-connectée de x_j en respectant Eps et $MinPts$ (connectivité) ;

L'algorithme est décrit en détail dans l'article de [ESTER et al., 1996].

L'avantage des approches à base de densité réside dans leurs capacités à découvrir par elles-mêmes le nombre de clusters et à détecter des clusters ayant des formes concaves. Cependant, ces algorithmes ont du mal à séparer les données qui n'ont pas de régions à faible densité [HAN, PEI et KAMBER, 2011].

Méthodes probabilistes

Les méthodes probabilistes sont des approches qui supposent que les données aient été générées suivant un modèle de mélange de plusieurs lois de probabilité [MCLACHLAN et BASFORD, 1988] et qu'à l'intérieur de chaque cluster, les instances suivent (globalement) la même loi. Chaque

cluster Cl_i peut ainsi être associé à une loi de probabilité $P(x_i, \theta_i)$, où θ_i représente les paramètres qui vont estimer la probabilité d'appartenance de x_i à Cl_i. Si on suppose que les données sont générées par k lois de probabilités et π_i la proportion de la i^{eme} loi dans le modèle de mélange, alors la fonction des distributions peut être exprimée sous cette forme :

$$P(x_i, \Phi) = \sum_{i=1}^{k} \pi_i P(x_i, \theta_i) \tag{3.3}$$

avec $\Phi = (\pi_1, ..., \pi_k, \theta_1, ..., \theta_k)$, l'ensemble des paramètres à calculer pour le modèle du mélange. L'algorithme EM (Espérance Maximisation) [DEMPSTER, LAIRD et RUBIN, 1977] est parmi les plus utilisés pour estimer les paramètres de ce type de modèle de mélange.

Méthodes à base de graphes

Les méthodes à base de graphes [CRISTIANINI, SHAWE-TAYLOR et KANDOLA, 2001 ; FILIPPONE et al., 2008] sont une famille d'algorithmes qui ont une forte relation avec la théorie des graphes.

L'idée générale de ces méthodes est de construire un graphe à partir des données et d'appliquer par la suite le clustering pour le séparer en sous-groupes. Cette représentation en graphe est indépendante du type des données.

Sur un ensemble $\mathcal{X} = \{x_i\}_{i=1}^{n}$ de données non-étiquetées, on crée un graphe $G = (V, A)$ complet, pondéré et non orienté, où V est l'ensemble des noeuds représentant les instances x_i et A est l'ensemble des arcs du graphe. Les arcs ont chacun un poids qui représente la similarité entre les deux instances connectées par l'arc. L'objectif des algorithmes revient à former les clusters à partir des noeuds fortement connectés. La séparation des noeuds dépend donc fortement de la méthode de calcul des poids des arcs. Plusieurs algorithmes ont été proposés, comme le clustering spectral [NG, JORDAN et WEISS, 2002 ; VON LUXBURG, 2007], ROCK [GUHA, RASTOGI et SHIM, 1999], ou encore CHAMELEON [KARYPIS, HAN et KUMAR, 1999].

Méthodes à base de distance

Une autre famille d'algorithmes, certainement la plus riche en alternatives et la plus populaire, est la famille des approches à base de distance. Ces méthodes se basent sur une mesure de distance pour évaluer la proximité entre les instances. Ce type d'algorithmes est très efficace pour analyser des jeux de données à très grande dimension.

Nous présentons ci-dessous deux algorithmes se basant principalement sur les distances pour séparer les données. Il s'agit des algorithmes K-Means et SOM.

3.2.2 K-Means

K-Means est de manière certaine le plus connu des algorithmes de clustering et a été introduit par MACQUEEN (1967), l'objectif est de séparer un ensemble de données $\mathcal{X} = \{x_i\}_{i=1}^{n}$, en k clusters $\mathbb{C}_l = \{Cl_1, ..., Cl_k\}$ disjoints, où chaque cluster est représenté par un centroïde (prototype) μ, calculé à partir de la moyenne des instances du cluster.

$$\mu_j = \frac{1}{\mid Cl_j \mid} \sum_{x_i \in Cl_j} x_i. \tag{3.4}$$

Pour atteindre son objectif, K-Means s'appuie essentiellement sur deux étapes :

- **Première étape : Affectation des instances**
  - La distance entre chaque instance et les centroïdes est calculée ;
  - Chaque instance est affectée au cluster du plus proche centroïde.

- **Deuxième étape : Mise à jour des centroïdes**
  - La moyenne des instances de chaque cluster est calculée ;
  - Les centroïdes de chaque cluster sont mis à jour.

Une fois le nombre de cluster k fixé et les centroïdes initialisés [PENA, LOZANO et LARRANAGA, 1999], les deux étapes de l'algorithme sont répétées jusqu'à stabilisation de l'algorithme. L'enchaînement de ces deux étapes revient à optimiser la fonction objective suivante (somme des carrés) :

$$R_{KMeans} = \sum_{j=1}^{k} \sum_{x_i \in Cl_j} \|x_i - \mu_{Cl_j}\|^2 \tag{3.5}$$

avec, $\mathcal{X} = \{x_i\}_{i=1}^{n}$ les instances à analyser, k le nombre des clusters, Cl_j les instances affectées au cluster $j \in \{1, ..., k\}$, avec $\mathcal{X} = \{Cl_1 \cup Cl_2 \cup ... \cup Cl_k\}$ et μ_{Cl_j} le centroïde du cluster Cl_j.

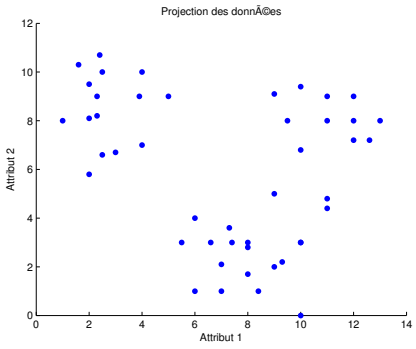


FIGURE 3.2 – Exemple d'un ensemble de données décrit par deux attributs

L'algorithme K-Means converge vers un minimum local [SELIM et ISMAIL, 1984]. Afin de mieux comprendre son fonctionnement, nous illustrons par la figure 3.3 les différentes itérations de l'algorithme appliqué à un jeu de données artificiel (figure 3.2) comportant 46 instances de deux dimensions. On peut voir, dans la figure 3.3, l'enchaînement des itérations de l'algorithme pour trouver une bonne séparation des données. Nous avons fixé le nombre de clusters à $k = 3$ et initialisé les centroïdes aléatoirement. Les couleurs des instances montrent leur affectation à chaque itération. Les déplacements des trois centroïdes est, quant à lui, indiqué par la trajectoire du $\times$ sur les illustrations.

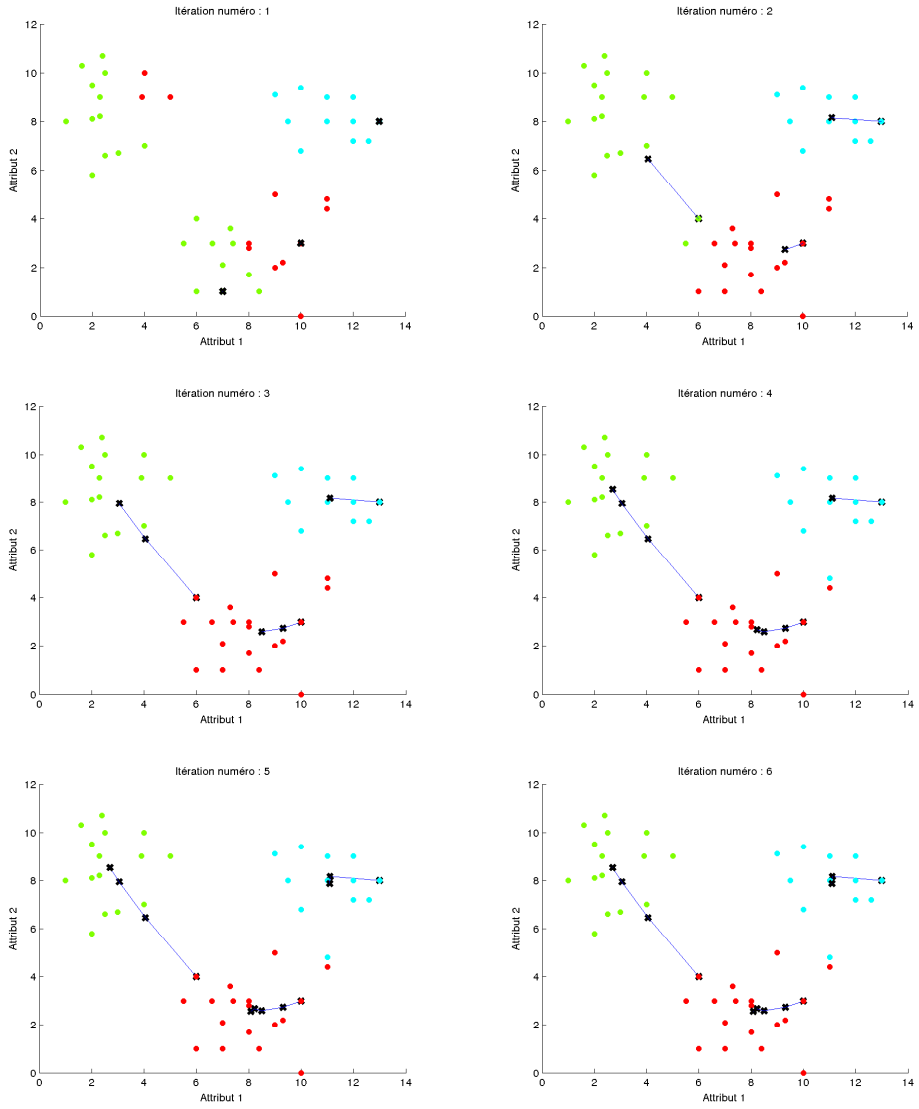


FIGURE 3.3 – Évolution des centroïdes et de l'affectation des instances aux clusters à chaque étape de K-Means

3.2.3 Cartes auto-organisatrices

Les cartes auto-organisatrices ,*Self Organizing Maps* (SOM), sont un algorithme non-supervisé à base de réseau de neurones. Introduit initialement par KOHONEN (1990) et KOHONEN (2013), les cartes SOM ont vite montré leurs utilités dans le cadre de plusieurs applications. Elles sont notamment utilisées pour la visualisation [VESANTO, 1999], la compression de données et la réduction de leurs dimensions.

L'idée est de représenter les données par une carte spatialement organisée qui peut être vue comme un graphe de neurones non orienté et connecté. Durant l'apprentissage, les instances vont être présentées une à une à l'algorithme et un neurone de la carte va être activé à chaque fois. Ce neurone correspondra au neurone gagnant sélectionné par compétition (minimisant la distance avec l'instance présentée). Ainsi, les neurones vont capter des instances similaires. De plus, les neurones connectés s'influencent mutuellement afin d'organiser spatialement la carte, cela permet que les instances proches dans l'espace d'entrée soient associées à des neurones proches sur la carte. Cette influence mutuelle est pondérée par une fonction de voisinage $\mathcal{K}$:

$$\mathcal{K}_{i,j} = \frac{1}{\lambda(t)} \exp^{-\frac{d_1^2(i,j)}{\lambda^2(t)}} \tag{3.6}$$

avec i,j, deux neurones, $\lambda(t) = \lambda_i(\frac{\lambda_f}{\lambda_i})^{\frac{t}{t_{max}}}$ une fonction de température qui contrôle l'étendue du voisinage et $d_1^2(i,j)$ la distance de Manhattan.

Chaque neurone a un prototype (vecteur de référence) w_j qui lui est associé, représentant les instances captées par celui-ci.

L'objectif de l'apprentissage des cartes SOM peut être formulé comme un problème de minimisation de la fonction objective :

$$R_{SOM} = \sum_{i=1}^{n} \sum_{j=1}^{k} \mathcal{K}_{j,\chi(x_i)} \|x_i - w_j\|^2 \tag{3.7}$$

où n est le nombre des instances du jeu de données, k le nombre des neurones de la carte, $\chi(x_i)$ assigne x_i au neurone le plus proche.

3.2.4 Qualité des résultats du clustering

Les méthodes de clustering analysent des données non étiquetées, sans supervision, et produisent des clusters à la fin de leur exécution.

Dans ce cadre, la validation des résultats et la mesure de leur qualité deviennent des éléments essentiels. L'évaluation des résultats peut faciliter et guider le choix de l'utilisateur pour l'algorithme à appliquer aux données analysées ainsi que les paramètres à utiliser afin d'obtenir les meilleures performances.

Trois types de mesures de qualité existent, des mesures externes, des mesures internes et des mesures relatives [HALKIDI, BATISTAKIS et VAZIRGIANNIS, 2001].

Les mesures externes évaluent la qualité des résultats en se basant sur des informations externes comme un étiquetage établi par l'expert. Les mesures internes s'appuient , quant à elles, sur des critères propres aux données. La qualité du partitionnement peut être évaluée par exemple en se basant sur les distances entre les objets du même cluster. Finalement, les mesures relatives comparent, suivant une mesure de qualité prédéfinie (externe ou interne), les résultats produits sur un jeu de données par le même algorithme en utilisant plusieurs paramètres.

Plusieurs mesures d'évaluation de la qualité du clustering ont été proposées dans la littérature [RENDÓN et al., 2011a]. Dans la suite de cette section, nous commençons par la présentation de quelques mesures externes, puis nous définissons par la suite un ensemble de mesures internes.

Mesures externes

Nous supposons dans le cadre des mesures externes l'existence d'une classification de référence (vérité terrain). Cette classification est admise correcte, elle est la plupart du temps le résultat d'une catégorisation manuelle d'un expert ayant une connaissance des données et du problème étudié.

Le **rappel** et la **précision** ont été introduits par KENT et al. (1955). En apprentissage, le **rappel** mesure la complétude de la classification obtenue, tandis que la précision renseigne sur l'exactitude des résultats. Afin de faciliter la compréhension des définitions, nous commençons par définir les quatre cas possibles lorsqu'il s'agit de comparer la classification d'un d'instances avec un algorithme d'apprentissage par rapport à une vérité terrain :

— Vrais positifs (**VP**) : Le cas des instances correctement attribuées à une classe ;
— Faux positifs (**FP**) : Le cas des instances attribuées à une classe à qui elles n'appartiennent pas ;
— Vrais négatifs (**VN**) : Le cas des instances correctement non attribuées à une classe ;
— Faux négatifs (**FN**) : Le cas des instances non attribuées à une classe alors qu'elles appartiennent à cette classe ;

À partir de ces définitions, nous pouvons définir le rappel et la précision par les rapports suivants :

$$Rappel = \frac{VP}{VP + FN}, \tag{3.8}$$

$$Précision = \frac{VP}{VP + FP}. \tag{3.9}$$

L'utilisation du rappel sans la précision ou de la précision sans le rappel peut fausser l'évaluation des résultats. En effet, si l'on met toutes les instances dans un seul cluster, cela donne un rappel maximal alors que la précision de la classification reste très basse. Afin d'éviter ce problème, la **F1-Mesure** combine ces deux métriques et introduit une nouvelle mesure

d'évaluation supervisée [RIJSBERGEN, 1979].

$$F1 - Mesure = \frac{2 * Précision * Rappel}{Précision + Rappel} \tag{3.10}$$

Trois autres mesures sont aussi très utilisées pour évaluer les résultats d'une classification ou d'un clustering. Il s'agit de l'**exactitude**, de la **spécificité** et de l'**entropie**.

L'**exactitude** représente la fraction correctement classifiée. Elle est définie par l'équation suivante :

$$Exactitude = \frac{TP + RN}{VP + FN + FP + VN} \tag{3.11}$$

La **spécificité** mesure quant à elle la capacité d'une classification à bien détecter les résultats négatifs. Elle est définie par l'expression suivante :

$$Specificit = \frac{VN}{VN + FP} \tag{3.12}$$

L'**Entropie** est un indice de qualité calculé en plusieurs étapes. La probabilité qu'une instance appartenant à un cluster j appartient à la classe i est d'abord estimée par la formule :

$$p_{ij} = \frac{m_{ij}}{m_j} \tag{3.13}$$

où m_j est le nombre d'instances du cluster j et m_{ij} le nombre d'instances appartenant à la classe i et attribuées au cluster j.

À partir de cette probabilité, l'entropie de chaque cluster j est calculée par la formule :

$$e_j = \sum_{i=1}^{L} p_{ij} \log_2(p_{ij}) \tag{3.14}$$

avec L le nombre de classes.

L'entropie totale est ensuite obtenue par la somme des entropies de chaque cluster pondérée par la taille des clusters (eq. 3.15).

$$e = \sum_{i=1}^{k} \frac{m_j}{m} e_j \tag{3.15}$$

où k est le nombre des clusters.

Mesures internes

L'obtention d'une vérité terrain est parfois très coûteuse voire même impossible, l'utilisation des mesures internes peut remédier à ce problème et donner une idée sur les performances des algorithmes. RENDÓN et al. (2011b) ont montré que les indices internes apportent également des informations pertinentes sur les résultats.

Le **coefficient silhouette** [ROUSSEEUW, 1987] est une méthode d'évaluation de la séparabilité et la compacité des clusters obtenus $\mathbb{C}_l = \{Cl_1, ..., Cl_k\}$. Le coefficient $\mathcal{S}$ est calculé pour chaque instance x_i appartenant à un cluster Cl_i suivant la formule :

$$\begin{aligned} \mathcal{S}(x_i) = 1 - \frac{a(x_i)}{b(x_i)} sia(x_i) < b(x_i) \\ sinon \mathcal{S}(x_i) = \frac{b(x_i)}{a(x_i)} - 1 \end{aligned} \tag{3.16}$$

où $a(x_i)$ est la distance moyenne entre x_i et les instances du cluster Cl_i et $b(x_i)$ est le minimum des moyennes de distance entre x_i et les instances appartenant aux autres clusters $\mathbb{C}_l = \{Cl_1, ..., Cl_k; k \neq i\}$.

Le coefficient prend des valeurs entre -1 et 1. Si la valeur $\mathcal{S}(x_i)$ est proche de 1, cela montre que les instances du même cluster sont plus proches de cette instance x_i que des instances des autres clusters. Le coefficient d'un cluster est la moyenne des coefficients des instances de ce cluster. Elle se situe aussi entre -1 et 1, et peut être interprétée de la même manière.

Une autre mesure populaire est l'indice **Davies et Bouldin** [DAVIES et BOULDIN, 1979], il permet aussi d'évaluer la compacité et la séparabilité des clusters. Il est défini par l'équation :

$$BD = \frac{1}{k} \sum_{i=1}^{k} Max_{i \neq j} \left\{ \frac{S(Cl_i) + S(Cl_j)}{d(\mu_i, \mu_j)} \right\} \tag{3.17}$$

Où k représente le nombre des clusters, $d(\mu_i, \mu_j)$ la distance entre les centroïdees de Cl_j et Cl_j, et $S(Cl_i)$ la distance moyenne entre chaque instance du cluster Cl_i et son centroïde μ_i, défini par :

$$Cl_i = \frac{1}{\mid Cl_i \mid} \sum_{x_i \in Cl_i} d(x_i, \mu_i) \tag{3.18}$$

Plus la valeur de l'indice est faible, plus les résultats du clustering sont jugés de meilleure qualité.

La **cohésion** (SSE) et la **séparation** (BSS) sont deux autres indices d'évaluation internes. LLe premier mesure la proximité entre les instances d'un même cluster. Le deuxième évalue quant à lui le degré de séparation entre les clusters.

$$SSE = \sum_{i} \sum_{x \in Cl_i} (x - \mu_i)^2 \tag{3.19}$$

$$BSS = \sum_{i} \mid Cl_i \mid (\mu - \mu_i)^2 \tag{3.20}$$

3.2.5 Complexité des algorithmes

La complexité algorithmique est un élément important dans le choix des techniques de clustering à appliquer. Elle permet d'estimer les temps de

calcul ainsi que la capacité du passage à l'échelle des approches. Le tableau 3.1 résume la complexité de quelques algorithmes évoqués précédemment, avec n le nombre d'instances, d le nombre des variables (espace de définition $\mathbb{R}^d$), k le nombre de clusters, m_n la moyenne des voisins pour chaque instance et m_m le nombre de voisins maximum de la même instance [XU et WUNSCH, 2005].

Algorithme	Complexité
Hiérarchique	$\mathcal{O}(n^2)$ au minimum
ROCK	$\mathcal{O}(n^2 + n \times m_m \times m_n + n^2 \log n)$
DBSCAN	$\mathcal{O}(n \log n)$
K-Means	$\mathcal{O}(k \times n \times d)$
SOM	$\mathcal{O}(n \times k)$

TABLE 3.1 – Complexité en temps des algorithmes de clustering

3.3 Résumé

Dans ce chapitre, nous avons introduit les notions fondamentales de l'apprentissage artificiel et présenté plusieurs algorithmes de clustering. Ce panorama, loin d'être exhaustif, montre la diversité des algorithmes existants mais aussi la difficulté du problème du clustering. Nous avons tenté d'organiser au mieux les approches disponibles pour en proposer une synthèse et présenter pour chacune d'elle un exemple d'algorithme de clustering. Cette catégorisation reste subjective et peut varier selon le point de vue et les critères choisis pour catégoriser les méthodes. Il existe par ailleurs d'autres familles d'algorithmes qui ne sont pas abordées ici comme les approches de clustering flou [BARALDI et BLONDA, 1999 ; YANG, 1993] ou encore celles à base de grilles [BERKHIN, 2006].

En dépit du nombre important d'algorithmes proposés, aucun de ces algorithmes ne peut être considéré comme le "meilleur". En évaluant différents types d'algorithmes sur plusieurs jeux de données, plusieurs études [JAIN et al., 2004 ; MAULIK et BANDYOPADHYAY, 2002 ; XU et WUNSCH, 2005] ont montré que les performances des algorithmes variaient suivant le contexte, la nature des données, les motifs recherchés... L'évaluation même de ce qu'est un bon résultat reste difficile et fortement dépendant du critère choisi.

La nature généraliste du clustering qui lui permet d'aborder une multitude de problèmes limite aussi ses performances. Contrairement à la classification supervisée, qui utilise les données étiquetées pour se spécialiser et adapter son modèle au problème étudié, le clustering attaque tous les problèmes de la manière et poursuit généralement l'objectif de maximiser la similarité intra-classe (entre les instances du même cluster) et minimiser la similarité inter-classe (maximiser la séparation des clusters). Cela rend le clustering difficilement adaptable à des problèmes spécifiques.

La nature totalement non-supervisée du clustering limite parfois son application. L'image satellite par exemple peut être partitionnée de plusieurs manières et l'évaluation des résultats peut fortement varier suivant le point de vue de l'utilisateur. La non prise en compte des connaissances externes dans le clustering est un verrou important qui limite ses performances.

Nous allons examiner dans le chapitre 5 les différentes techniques de la littérature permettant l'intégration de ces informations dans le cadre des techniques semi-supervisées.

Chapitre 4

Les images satellites

Ce dernier chapitre introductif sera consacré à la présentation des images satellites et de leurs spécificités. Les images satellites sont le domaine d'application privilégié des contributions informatiques développées dans la suite du manuscrit.

Dans cette perspective, il est essentiel de considérer la nature complexe des images et d'aborder les principes liés à la forte volumétrie des données ou bien au fort pouvoir d'expressivité rendu nécessaire pour la représentation du contenu de l'image. Ces principes sous-tendent les implémentations et expérimentations effectuées pour la validation de nos propositions.

Sommaire

"The new electronic independence re-creates the world in the image of a global village", **Marshall McLuhan**

4.1 Introduction

Les images satellites sont devenues une source d'information incontournable. Elles peuvent apporter, après traitements appropriés, des éléments de réponse à diverses problématiques, qu'elles soient de nature environnementale ou sociétale. Elles constituent donc de véritables atouts, en particulier pour les sciences de l'environnement et de la santé, en offrant de nouveaux regards sur l'espace et au travers du temps. Cependant, les images satellites sont des données particulièrement complexes à manipuler et à analyser. En effet, l'obtention des connaissances contenues dans les milliers d'images acquises par les multiples satellites observant la Terre est particulièrement coûteuse et difficile. Afin de mieux cerner cette complexité, comprendre à la fois la nature de ces images et les procédés qui ont mené à leur acquisition est certainement l'une des premières choses à entreprendre.

Le monde de l'imagerie satellitaire est un monde pluridisciplinaire par excellence. Cette pluridisciplinarité favorise l'ambiguïté dans les termes utilisés par les différentes communautés scientifiques. Elle rend aussi le consensus sur les significations des termes du domaine, plus difficile. Nous nous efforçons dans cette section de définir les concepts et les termes du domaine de l'imagerie satellitaire en nous basant sur des travaux reconnus dans la communauté ; ces définitions sont celles que nous adopterons par la suite pour expliquer nos propositions.

4.2 Les images numériques

Le mot **image** est polysémique et la sémantique du mot change selon le contexte de son utilisation. Le dictionnaire Larousse [1] propose ainsi une dizaine de définitions, parmi lesquelles on retrouve :

- **Ensemble de points ou d'éléments représentatifs de l'apparence d'un objet, formés à partir du rayonnement émis, réfléchi, diffusé ou transmis par l'objet** ;
- Symbole ou représentation matérielle d'une réalité invisible ou abstraite ;
- Représentation mentale élaborée à partir d'une perception antérieure.

La première définition donnée semble être la plus en adéquation avec l'image en tant que donnée numérique, il faut cependant garder à l'esprit que même cette définition de l'image en tant que donnée numérique, doit être considérée avec précaution. En effet, la différence entre l'objet représenté par l'image et sa représentation numérique par les pixels de l'image peut parfois être sujet à confusion. Il faut donc prêter attention à la différence subtile entre les deux. L'image dans notre cas désigne la représentation numérique d'un objet (réel), formée par les rayonnements émis, réfléchis et diffusés par l'objet et capturés par les capteurs de l'appareil émettant l'image.

1. Définition "Image" : www.larousse.fr/dictionnaires/francais/image/41604

Deux types d'images numériques sont généralement distinguées : les images dites matricielles et les images dites vectorielles. Les premières sont composées d'un ensemble (matrice) de points (pixels), où chaque point (pixel) possède un ensemble de valeurs. L'espace spatial que représente la matrice des pixels dépend de la distance qui sépare chaque pixel de la partie de l'objet auxquelles les valeurs font référence. On parle de la **définition** de l'image. Les différentes valeurs admises par chaque pixel, correspondent à sa capacité de représentation des couleurs et de leur intensité. On parle de la **profondeur** de l'image. La qualité de l'image dépend donc de sa définition et sa profondeur. Les images vectorielles sont quant à elles composées d'objets géométriques individuels (segments de droite, polygones, arcs de cercle, ...), qui sont définis mathématiquement à l'aide de différents attributs à l'exemple de la forme, de la position ou de la couleur.

4.3 Les images d'observation de la Terre

Les images satellites sont des images matricielles qui possèdent des caractéristiques propres. Elles font ainsi partie de la famille des images d'observation de la Terre au même titre que des photos aériennes. La télédétection est la discipline concernée par l'acquisition, la mise à disposition, l'étude et l'analyse de ce type de données. Elle est définie par le centre canadien de cartographie et d'observation de la Terre[2] comme *la technique qui, par l'acquisition d'images, permet d'obtenir de l'information sur la surface de la Terre sans contact direct avec celle-ci. La télédétection englobe tout le processus qui consiste à capter et à enregistrer l'énergie d'un rayonnement électromagnétique émis ou réfléchi, à traiter et à analyser l'information, pour ensuite mettre en application cette information* [TERRE CCRS, 2013]. Nous allons exposer les différents processus liés aux images satellites pour mieux comprendre leurs spécificités et la complexité qui en découle.

4.3.1 L'acquisition des images satellites

L'observation de la Terre en télédétection se fait donc au moyen de satellites. Ces satellites embarquent des capteurs qui vont enregistrer le rayonnement de la surface cible. Deux types de capteurs existent, des capteurs dits passifs et des capteurs dits actifs [TERRE CCRS, 2013]. Les capteurs passifs perçoivent l'énergie du soleil réfléchie et l'énergie émise naturellement par la cible, tandis que les capteurs actifs produisent leur propre énergie (rayonnement électromagnétique) et perçoivent, par la suite, l'énergie réfléchie par la cible. La figure 4.1 présente le processus général de l'acquisition des images par utilisation des satellites à capteurs passifs. Nous avons choisi de décomposer ce processus en six étapes :

2. CCCOT (Centre canadien de cartographie et d'observation de la Terre) : http://www.rncan.gc.ca/accueil

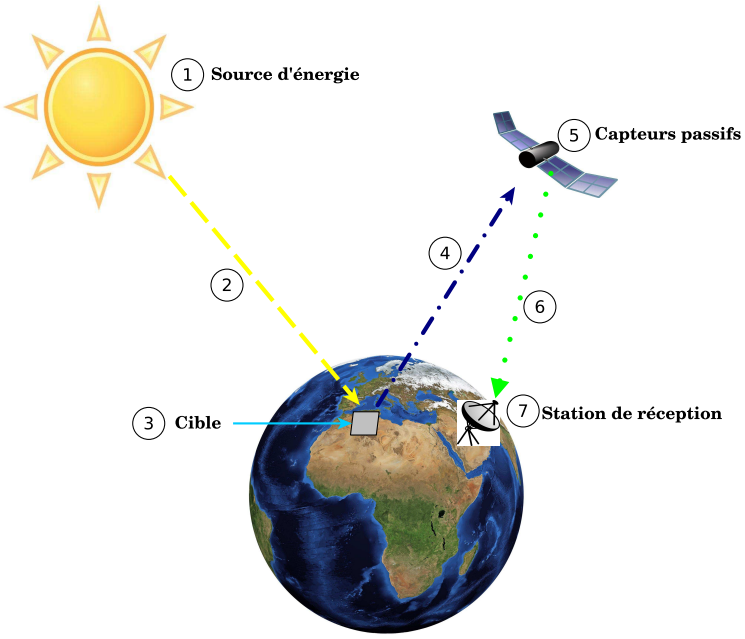


FIGURE 4.1 – Processus d'acquisition d'images d'observation de la Terre avec un satellite doté de capteurs passifs

1. La source d'énergie : cette source est le soleil pour les capteurs passifs ;
2. Les rayonnements : l'énergie traverse l'atmosphère pour atteindre la cible, les rayonnements interagissent avec l'atmosphère durant son parcours ;
3. La cible : c'est la surface de la terre observée. L'énergie interagit, suivant les propriétés de la surface, avec la cible ;
4. L'énergie émise ou reflétée par la surface observée traverse l'atmosphère une deuxième fois ;
5. Satellite d'observation : les capteurs du satellite enregistrent l'énergie captée à distance ;
6. Le satellite transmet l'énergie captée en utilisant des moyens électroniques à une station de réception ;
7. La station de réception reçoit l'information du satellite et la transforme en images.

Ces étapes retracent le processus d'obtention des images satellites d'observation de la Terre. Les satellites passifs ne diffusent pas de signaux. Ils enregistrent les rayonnements (réfléchis ou émis) des différents objets de la surface suivant les longueurs d'onde qu'ils peuvent capter. Chaque satellite peut donc être caractérisé par le nombre de bandes spectrales qu'il est capable d'observer et les intervalles sur chacune de ses longueurs d'onde (fenêtre spectrale). On parle alors de la **résolution spectrale** du satellite.

Une autre caractéristique importante du satellite est sa **résolution spatiale**. Similaire à la définition de l'image, elle désigne la taille de ce que représente un pixel par rapport à la surface de la Terre. Un pixel peut par exemple représenter une surface carrée de 5m de coté, on parle alors d'une résolution de 200 pixels par kilomètre, ou par abus de langage d'une résolution de 5 mètres.

Deux autres caractéristiques, moins importantes dans notre cadre, peuvent être utilisées pour la catégorisation des satellites. La première est la **résolution radiométrique**, qui désigne la capacité des capteurs du satellite à détecter des petites variations dans l'énergie électromagnétique. La deuxième est la **résolution temporelle** du capteur, qui correspond à son temps de revisite, c'est à dire le temps qu'il faut au satellite pour observer la même cible (scène) à partir de la même position.

Ces caractéristiques sont très importantes pour le choix des images par les experts. Suivant l'objectif des études entreprises, les types des images à analyser et à interpréter changent. La résolution spatiale est par exemple une caractéristique qui va déterminer les types d'objets de la surface possiblement identifiables par les images. Plus la résolution est petite, plus la précision des images est grande et donc le nombre d'objets réels pouvant être caractérisés est plus grand. D'après KERGOMARD (1990), les images satellites du domaine de la télédétection civile sont organisées en quatre classes de résolution :

— Les images de basse résolution : 1000 mètres ou plus ;
— Les images de moyenne résolution : 80 mètres ;
— Les images de haute résolution : entre 30 et 10 mètres ;
— Les images de très haute résolution : 5 mètres ou moins.

4.3.2 Les images satellites en tant que données complexes

Les images produites par les satellites d'observation de la Terre ont différentes caractéristiques, en fonction des caractéristiques des capteurs embarqués. Nous donnons ici quelques types d'images les plus répandues en télédétection :

— L'**image monospectrale** : C'est une image dont les pixels ont une seule valeur numérique (souvent comprise entre 0 et 255) exprimant la moyenne du rayonnement émis par la surface et capté par le satellite, représentée sur un intervalle donné de longueur d'onde (fenêtre spectrale). Quand cette fenêtre spectrale est étroite et assimilable à une longueur d'onde, on parle d'**image/bande spectrale**. Quand la fenêtre spectrale est large, on utilise le terme **image/bande panchromatique** ;
— L'**image multispectrale** : C'est une collection d'images monospectrales prises au même instant d'une même surface terrestre, les pixels sont représentés par un ensemble de valeurs, où chaque valeur représente une fenêtre spectrale ou une longueur d'onde précise. Quand l'image contient plusieurs dizaines de bandes ou plus, on parle d'une

image hyperspectrale.

La figure 4.2 donne une représentation graphique d'une image multispectrale. Elle peut être vue comme une matrice, où chaque pixel est représenté par un élément de la matrice, la profondeur de cette matrice correspond aux bandes spectrales de l'image. À droite de l'image, on peut voir une illustration du rayonnement réfléchi par la surface de la Terre représentée par le pixel x_i sur toute le spectre électromagnétique. Les capteurs du satellite vont enregistrer le rayonnement de cette surface uniquement sur les fenêtres spectrales indiquées sur notre figure. L'intensité du rayonnement sur ces fenêtres spectrales vont représenter les valeurs numériques du pixel x_i. La surface que représente chaque pixel, les fenêtres spectrales enregistrées et leur précision sont liées aux capacités du satellite. Plus les satellites sont performants, plus la résolution est fine, les bandes nombreuses et l'intervalle des fenêtres spectrales réduit.

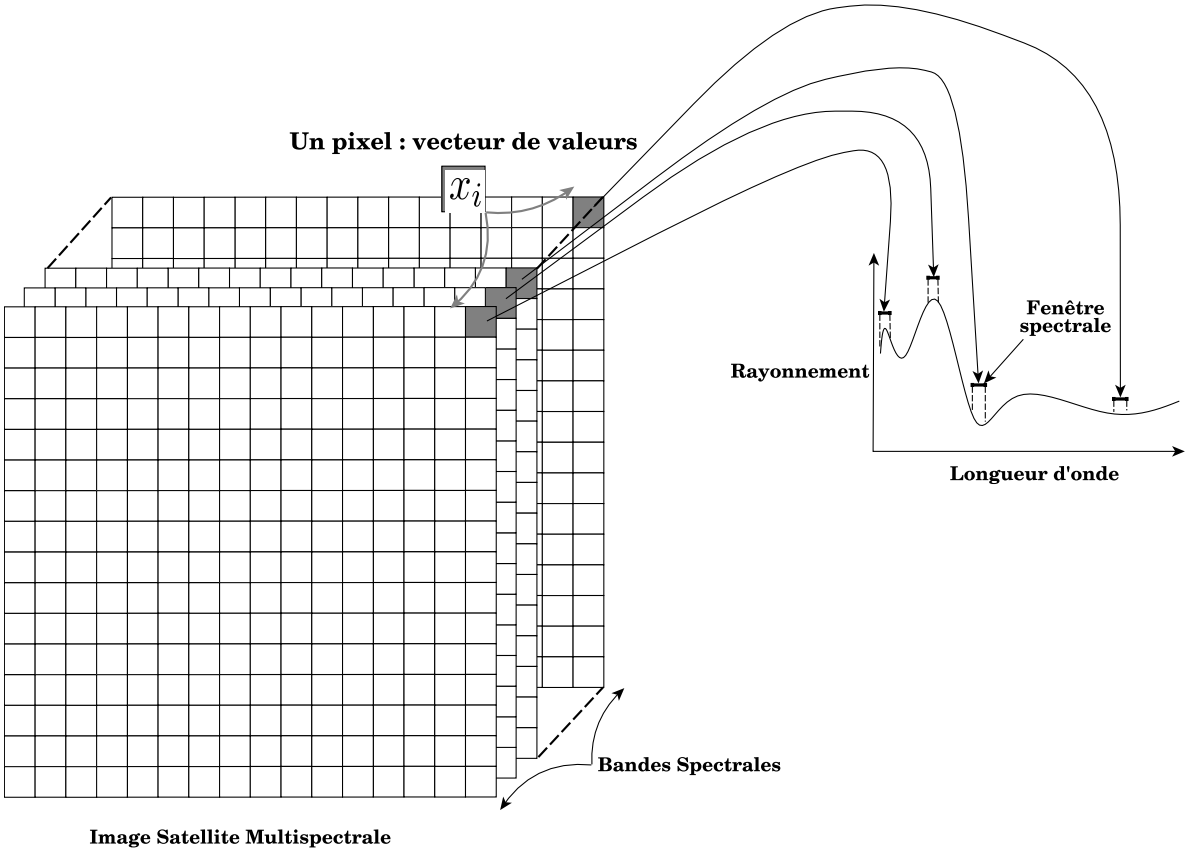


FIGURE 4.2 – Représentation simplifiée d'une image satellite multispectrale

Métadonnées

La figure 4.2 représente la nature de l'image en tant que donnée numérique. Les images satellites sont accompagnées d'informations supplémentaires représentées au travers de métadonnées, en complément de la matrice des vecteurs numériques.

```
GROUP = L1_METADATA_FILE
  GROUP = METADATA_FILE_INFO
    ORIGIN = "Image courtesy of the U.S. Geological Survey"
    REQUEST_ID = "0101406055082_00003"
    LANDSAT_SCENE_ID = "LT52280622009296CUB01"
    FILE_DATE = 2014-06-06T06:57:43Z
    STATION_ID = "CUB"
    PROCESSING_SOFTWARE_VERSION = "LPGS_12.4.1"
    DATA_CATEGORY = "NOMINAL"
  END_GROUP = METADATA_FILE_INFO
  GROUP = PRODUCT_METADATA
    DATA_TYPE = "L1T"
  .
    SPACECRAFT_ID = "LANDSAT_5"
    SENSOR_ID = "TM"
    SENSOR_MODE = "BUMPER"
    WRS_PATH = 228
    WRS_ROW = 062
    DATE_ACQUIRED = 2009-10-23
    SCENE_CENTER_TIME = 13:44:34.7070690Z
    CORNER_UL_LAT_PRODUCT = -1.95421
    CORNER_UL_LON_PRODUCT = -57.01798
  .
    FILE_NAME_BAND_1 = "LT52280622009296CUB01_B1.TIF"
  .
  END_GROUP = PRODUCT_METADATA
  GROUP = IMAGE_ATTRIBUTES
    CLOUD_COVER = 0.00
    IMAGE_QUALITY = 9
    SUN_AZIMUTH = 110.01688635
    SUN_ELEVATION = 62.92185986
  .
  END_GROUP = IMAGE_ATTRIBUTES
  GROUP = MIN_MAX_RADIANCE
    RADIANCE_MAXIMUM_BAND_1 = 193.000
    RADIANCE_MINIMUM_BAND_1 = -1.520
  .
    RADIANCE_MINIMUM_BAND_7 = -0.150
  END_GROUP = MIN_MAX_RADIANCE
  GROUP = MIN_MAX_PIXEL_VALUE
    QUANTIZE_CAL_MAX_BAND_1 = 255
    QUANTIZE_CAL_MIN_BAND_1 = 1
  .
  .
  END_GROUP = MIN_MAX_PIXEL_VALUE
  GROUP = PROJECTION_PARAMETERS
    MAP_PROJECTION = "UTM"
    DATUM = "WGS84"
    ELLIPSOID = "WGS84"
    UTM_ZONE = 21
    GRID_CELL_SIZE_REFLECTIVE = 30.00
    GRID_CELL_SIZE_THERMAL = 30.00
    ORIENTATION = "NORTH_UP"
    RESAMPLING_OPTION = "CUBIC_CONVOLUTION"
    MAP_PROJECTION_L0RA = "NA"
  END_GROUP = PROJECTION_PARAMETERS
END_GROUP = L1_METADATA_FILE
END
```

FIGURE 4.3 – Extrait d'un fichier de métadonnées complétant une image LANDSAT 5 TM

Une métadonnée est définie au sens large comme une donnée portant sur une autre donnée. Il s'agit souvent d'une information factuelle qui va permettre de contextualiser les données décrites, à l'exemple de la date d'acquisition ou bien de l'emprise spatiale.

Le W3C (World Wide Web Consortium) les envisage comme des données structurées venant compléter toute ressource et pouvant être exploitées par des agents logiciels.

PRIÉ et GARLATTI, 2004 précisent les définitions précédentes et considèrent les métadonnées comme des informations susceptibles d'organiser les ressources disponibles afin d'en faciliter l'indexation et donc la recherche et d'en optimiser l'exploitation par des agents logiciels.

Pour ce qui concerne les images satellites, il s'agit le plus souvent de métadonnées structurées et standardisées qui sont essentielles pour la compréhension du contenu de l'image. Pour donner un exemple, dans la figure 4.2), les métadonnées permettent de préciser la correspondance entre la dimension de la matrice et les longueurs d'onde traitées.

Différents standards de métadonnées ont actuellement cours dans le périmètre de l'observation de la Terre. ISO 19115 [*ISO 19115 - Geographic Information Metadata* 2003] est une référence dans le domaine. Proposé par l'ISO (Organisation Internationale de Normalisation), le standard fournit une synthèse exhaustive d'éléments de métadonnées venant structurer l'information géographique. Les fournisseurs d'images satellites établissent aussi leurs propres standards qui sont plus orientés vers la caractérisation des satellites et de leurs capteurs.

Au regard de la finalité du standard, les éléments de métadonnées changent et/ou se présentent différemment. Les images LANDSAT sont par exemple fournies au moyen d'un fichier texte accompagnant les images. Ce fichier décrit, entre autres, avec des métadonnées standardisées, les caractéristiques de l'image, des capteurs et des conditions d'acquisition. Il fournit également des informations sur la date d'acquisition et l'emprise spatiale de l'image, ainsi que les pré-traitements appliqués. La figure 4.3 montre un extrait d'un fichier de métadonnées d'un image LANDSAT 5 TM.

4.3.3 Les différents satellites d'observation de la Terre

Avec le développement des technologies et l'intensification de l'utilisation des images satellites pour l'étude de diverses problématiques en particulier liées au territoire, différents satellites d'observation de la Terre ont été conçus et mis en orbite par différents états et organisations. Il existe actuellement un nombre important de satellites opérationnels. Les États-Unis d'Amérique ont réussi, en 1972, à mettre en orbite l'un des premiers satellites d'observation de la Terre, baptisé Landsat-1. **LANDSAT** est devenu par la suite un programme dédié à des fins civiles dont les images sont publiques et libres d'utilisation. D'autres types satellites ont vite été rendus fonctionnels. La France dispose dans ce cadre de programmes de pointe

dans le domaine, comme les programmes **SPOT** et **Pléiades** [3], conçus principalement par le **CNES** [4]. Nous nous limiterons dans la suite de cette section à présenter les trois programmes SPOT, Pléiades et LANDSAT.

SPOT

SPOT [5] est un **S**ystème **P**our l'**O**bservation de la **T**erre [IMAGE, 1988] conçu à l'origine par le CNES en collaboration avec le la Belgique et le Suède. SPOT a été le premier programme Européen dédié à l'observation de la Terre, et se compose d'un ensemble de satellites lancés successivement de 1986 à 2015. Les premiers satellites (SPOT 1 à SPOT 3) étaient identiques et ont embarqué chacun deux capteurs HRV (haute résolution visible) qui ont permis de capturer des images dans le spectre du visible et de livrer ensuite des images à trois bandes spectrales à 20 mètres de résolution et d'une bande panchromatique à 10 mètres de résolution. SPOT 4, lancé en 1998, a embarqué deux capteurs identiques HR VIR (haute résolution visible et infra-rouge) qui ont permis de produire des images à quatre bandes spectrales à 20 mètres de résolution et d'une bande monospectrale à 10 mètres de résolution. La troisième génération de SPOT, SPOT 5, a embarqué deux nouveaux instruments HRG (haute résolution géométrique), à même de capturer des images à très haute résolution constituées d'une bande panchromatique de 2,5 mètres et 10 mètres en mode multispectral. SPOT 5 avait aussi embarqué un nouvel instrument de prise de vue HRS (Haute Résolution Stéréoscopique) fonctionnant en mode panchromatique pointant à la fois vers l'avant et vers l'arrière du satellite ce qui permettait de restituer le relief.

Le programme SPOT a été clôturé par le CNES avec la désorbitation de SPOT 5 en 2015. Il s'agit désormais d'une entreprise commerciale Airbus Defence & Space [6] qui produit les images SPOT 6 et SPOT 7 mettant à profit les applications opérationnelles de leurs prédécesseurs.

Pléiades

À la suite du programme SPOT, le CNES a lancé en 2011 et en 2012 deux satellites de pointe, Pléiades 1A et Pléiades 2B. Ce nouveau programme pour l'observation de la Terre et l'acquisition d'images à très haute résolution a été développé dans le cadre du programme Franco-Italien **Orféo** [7], en collaboration avec les industriels d'Airbus Defence & Space et de Thales Alenia Space. La constellation Pléiades en est la composante optique. Munis d'un capteur opérant dans le spectre visible et proche infrarouge, les satellites Pléiades capturent des images avec une résolution de 0,7 mètres pour la bande panchromatique, et une résolution de 2,8 mètres pour les quatre 4 bandes spectrales bleue, verte, rouge et proche infra rouge.

3. CNES Pléiades : https ://pleiades.cnes.fr/
4. CNES (Centre National d'Études Spatiales) : https ://cnes.fr/fr
5. SPOT (Système Pour l'Observation de la Terre) : https ://spot.cnes.fr/
6. Airbus Defence & Space https ://airbusdefenceandspace.com/
7. Orféo : Optical and Radar Federated Earth Observation

LANDSAT

Le programme **LANDSAT** regroupe un ensemble de satellites mis en orbite par la NASA. Huit satellites ont été lancés depuis le début du programme. Les trois premiers satellites possèdent les mêmes caractéristiques et sont équipés d'un capteur multispectral MSS capable de fournir des images à quatre bandes spectrales avec une résolution spatiale de 80 mètres. Les deux satellites LANDSAT-4 et LANDSAT-5 ont embarqué un scanner MSS dotés de détecteurs plus performants et d'un scanner TM (*Thematic Maper*). Cela a permis à ces deux satellites de fournir des images à sept bandes spectrales avec une résolution spatiale de 30 mètres pour six des sept bandes et de 120 mètres pour la bande thermique. Lancé après la perte de LANDSAT-6 pendant sa mise en orbite. LANDSAT-7 a embarqué un capteur amélioré ETM+, permettant de capturer des images à 8 bandes, avec 6 bandes à 30 mètres, une bande panchromatique à 15 mètres et une bande dotée d'une résolution de 60 mètres dans l'infrarouge thermique. LANDSAT-8 est opérationnel depuis Le 11 février 2013, et produit des images satellites à 11 bandes spectrales, dont 8 sont similaires à LANDSAT-7.

LANDSAT-5 peut être considéré comme le satellite le plus abouti du programme LANDSAT. Initialement prévu pour 3 ans de service, il a permis l'acquisition de millions d'images pendant 29 ans, établissant ainsi un record de longévité inégalé, et qui surtout permet de pouvoir observer les changements de la couverture des sols avec les mêmes résolutions spectrales et spatiales. Le tableau 4.1 détaille les caractéristiques des images multispectrales produites par ce satellite.

Bandes spectrales	Résolution spatiale	Fenêtre spectrale
Bande 1	30 mètres	$0.45 - 0.52\,\mu m$
Bande 2	30 mètres	$0.52 - 0.60\,\mu m$
Bande 3	30 mètres	$0.63 - 0.69\,\mu m$
Bande 4	30 mètres	$0.76 - 0.90\,\mu m$
Bande 5	30 mètres	$1.55 - 1.75\,\mu m$
Bande 6	120 mètres	$10.40 - 12.50\,\mu m$
Bande 7	30 mètres	$2.08 - 2.35\,\mu m$

TABLE 4.1 – Les bandes spectrales LANDSAT 5 avec leurs résolution et fenêtre spectrale

Pour donner un ordre d'idée, une image satellite LANDSAT-5 (Scène) brute représente une surface terrestre de 170 km x 185 km (figure 4.4).

FIGURE 4.4 – Exemple d'une scène LANDSAT 5 TM sur le Sud de la France acquise le 13/10/2011

4.3.4 Accès aux images satellites

Une fois les images capturées et numérisées dans les stations de réception (figure 4.1), il reste à les mettre à disposition des différents acteurs et experts. Face à la multiplication des satellites mis en orbite et l'hétérogénéité des capteurs disponibles, un effort conséquent doit alors porter sur la publication et le partage des images dès lors que l'on souhaite leur valorisation. En effet, une image satellite qui n'est pas analysée et interprétée n'est d'aucune utilité. Le moyen le plus habituel pour la publication et le partage des images est la création de portails d'accès, de recherche et de téléchargement des différentes données disponibles. Les portails déploient des infrastructures de données spatiales intégrant des données hétérogènes et distribuées [GRANELL et al., 2009]. Conscient de l'importance de la mise à disposition et de la facilitation de la découverte des images, plusieurs initiatives gouvernementales et industrielles ont vu le jour, parmi lesquelles on retrouve des initiatives à l'échelle mondiale (GEOSS [8], Earth Explorer USGS [9]), européenne (INSPIRE [DIRECTIVE, 2007], COPERNICUS [10]) ou

8. GEOSS : http ://www.geoportal.org
9. Earth Explorer : http ://earthexplorer.usgs.gov/
10. http ://www.copernicus.eu

nationale (GEOSUD [KAZMIERSKI et al., 2014], THEIA [BAGHDADI et al., 2015]).

Ces catalogues d'images satellites ont pour principales missions le stockage et l'indexation des images, afin de faciliter leur découverte et téléchargement. **GEOSUD** et **Earth Eplorer** sont deux exemples de catalogues riches en images et faciles d'usage.

GEOSUD [11] est un projet EQUIPEX, bénéficiant du Programme Investissements d'Avenir. Il vise à développer une infrastructure nationale de données satellitaires accessible gratuitement par les communautés scientifiques et autres acteurs publics. L'initiative a été lancée à partir du constat de sous-utilisation des données spatiales par les équipes de recherche et les acteurs publics travaillant sur la gestion des milieux et des ressources. Malgré la disponibilité des images, elles sont souvent délaissées par ces communautés scientifiques au profit de méthodes basées sur les données terrain. Parmi les explications données à ce phénomène figure la difficulté d'accès aux images. Le projet GEOSUD vient faciliter l'accès aux images de plusieurs satellites d'observation de la Terre et se positionne dans la perspective de l'initiative européenne COPERNICUS, qui vient répondre aux besoins des utilisateurs dans les domaines de l'agriculture, de l'environnement, de l'aménagement du territoire, de la société, de l'enseignement et de la recherche.

Earth Explorer [12] est un portail d'accès aux images satellites des différents programmes LANDSAT. Les images LANDSAT y sont distribuées gratuitement par l'USGS. Elles constituent une source d'information très riche, notamment grâce à la longévité du programme et ses couvertures répétées de toute la surface de la Terre. Les produits sont fournis au niveau *L1T*, correspondant à des images ortho-rectifiées exprimées en réflectances *TOA* [CALOZ et COLLET, 2001]. Pour la France par exemple, des produits de Niveau 2A, exprimées en réflectance de surface après correction atmosphériques et accompagnées d'un masque de nuages et d'ombres de nuages sont également disponibles gratuitement. L'interface Earth Explorer est simple et intuitive, elle permet la localisation et le téléchargement rapide des images. Le portail offre par exemple à l'utilisateur la possibilité de localiser les images suivant plusieurs critères, comme leur emprise spatiale (directement sur la carte ou en utilisant des coordonnées), les dates d'acquisition ou bien le type de satellite LANDSAT (Earth Explorer référence uniquement des produits LANDSAT). Le projet GEOSUD donne accès aux images provenant de différents programmes à l'exemple de LANDSAT, SPOT ou Pléiades. L'infrastructure de données spatiales exploite à cet effet un profil d'application qui permet de tirer parti de manière unifiée des différents standards de métadonnées venant décrire les images [DESCONNETS, CHAHDI et MOUGENOT, 2014 ; MOUGENOT, DESCONNETS et CHAHDI, 2015].

11. IDS GEOSUD : http ://ids.equipex-geosud.fr/
12. Earth Explorer : http ://earthexplorer.usgs.gov/

4.4 Résumé

Il s'agissait dans ce chapitre de présenter, dans les grandes lignes, les particularités des images satellites et de rendre compte de la richesse en matière de savoir que sont susceptibles de véhiculer ces milliers d'images accumulées depuis plus de cinquante ans. Il se dégage d'ores et déjà un certain nombre d'enseignements. Les images satellites sont des données volumineuses qui proviennent de différents programmes qui ont chacun fait des choix technologiques différents avec des améliorations régulières des satellites et des capteurs imageurs. Les images sont donc aussi des données hétérogènes. Des différences se font jour en particulier dans la résolution spatiale, spectrale et dans la répétitivité temporelle. Ces différences ont une conséquence directe sur la structure de l'image elle même et sur l'organisation des métadonnées qui la complètent. Des solutions génériques permettant l'analyse de toute catégorie d'image sont dès lors complexes à concevoir.

Chapitre 5

Connaissances et Apprentissage

Les deux premiers chapitres ont introduit les concepts de bases et les notions fondamentales des deux axes concernés par nos travaux de recherche, à savoir l'interprétation et la modélisation des connaissances ainsi que l'analyse de données exploratoires à base d'apprentissage. Le troisième chapitre a par la suite présenté les images satellites et leurs spécificités.

Ces chapitres ont posé le cadre général de la thèse et ont permis de cerner les domaines étudiés. Dans le même esprit, ce chapitre vient compléter cette première partie du mémoire par une présentation de travaux connexes se révélant des sources d'inspiration pour nos propres travaux. Nous fournissons ainsi un état de l'art des propositions existantes tout en expliquant les limites possibles. Par extension, nous discutons également les différences avec nos propositions.

Sommaire

"Knowledge that is not being used for winning of further knowledge does not even remain- it decays and disappears",
J.D. Bernal

5.1 Introduction

Des travaux de différentes natures sont menés pour bénéficier des connaissances du domaine parfois en présence de gros volumes de données à interpréter. Deux grandes familles d'approches peuvent dès lors être envisagées. Il s'agit des systèmes à base de connaissances et des systèmes de classification semi-supervisée.

Ces systèmes ne cherchent pas à faire le même usage des connaissances et en conséquence en proposent des cadres d'exploitation radicalement différents.

D'une part, les systèmes à base de connaissances s'appuient essentiellement sur un processus déductif. Les connaissances sont au cœur du système et les déductions permettent d'expliciter de nouveaux faits à partir de règles ou axiomes formulées par les experts et des données à analyser. D'autre part, les systèmes d'apprentissage semi-supervisés se basent essentiellement sur un processus inductif et utilisent les connaissances plutôt comme un moyen d'amélioration et de renforcement des inductions trouvées. Les connaissances interviennent ainsi pour guider le clustering et ne représentent donc pas la composante centrale de l'approche.

Dans la suite du chapitre, nous commençons par présenter le fossé sémantique qui est lié à l'analyse et l'interprétation des images. Nous expliquons brièvement les deux grandes approches habituellement exploitées pour analyser une image satellite. Nous abordons ensuite successivement les travaux qui font usage des systèmes à base de connaissances pour s'affranchir de ce fossé et les contributions en apprentissage semi-supervisé.

5.2 Fossé sémantique : De l'image au concept

L'interprétation de gros volumes de données quantitatives, à l'exemple des images satellites, est confrontée à un obstacle de taille appelé fossé sémantique [HARE et al., 2006] ou encore ancrage des symboles en intelligence artificielle [HARNAD, 1990].

Le fossé sémantique est lié à la difficulté d'attacher des concepts de haut niveau à des descripteurs de bas niveau qui sont extraits par calcul à partir des données de l'image. Il est ainsi défini par SMEULDERS et al. (2000) comme le manque de concordance entre ce qu'un humain est capable de percevoir d'une image et l'information qui peut être extraite de l'image à l'aide de processus de traitement du signal.

Il est à noter qu'une image n'a de sens que pour les objectifs qui lui sont fixés. Une image peut en conséquence être interprétée de diverses manières puisque la sémantique n'est pas dans l'image mais dépend d'un système de connaissances extérieur susceptible de varier en fonction du contexte et des besoins de l'analyse.

Nous avons vu dans le chapitre 4, qu'une image satellite est représentée numériquement comme une matrice de pixels, où chaque pixel est envisagé comme un vecteur de valeurs de densité énérgétique spectrale (figure 4.2). A partir de ces agrégats de données, une manière de réduire le fossé

sémantique consiste en premier lieu à trouver des éléments de réponse aux questions suivantes :

— Que représentent les valeurs numériques attachées aux pixels pour une longueur d'onde et un capteur donné ?
— Quel sens donner à ces pixels ou groupement de pixels dans un contexte d'étude précis ?

Il demeure donc très difficile d'espérer s'affranchir du problème du fossé sémantique sans aucune intervention humaine.

Avec son expérience et ses connaissances, un expert en photo-interprétation acquiert la capacité à contextualiser l'information et reconnaître visuellement les concepts thématiques. Le fossé est donc réduit par une forte implication de l'expert qui guide l'interprétation.

D'une manière similaire, dans le cadre de la classification supervisée, les classes thématiques sont fournies par l'expert, qui passe un temps considérable à élaborer l'échantillon d'apprentissage, et l'algorithme trouve par apprentissage le modèle qui relie les variables descriptives de bas niveau (les entrées) à ces classes thématiques.

Pour ce qui concerne le clustering, les algorithmes produisent des groupements de pixels (cluster) qui n'ont aucune sémantique et il revient ensuite à l'expert de trouver manuellement les correspondances entre les clusters et les classes thématiques cibles.

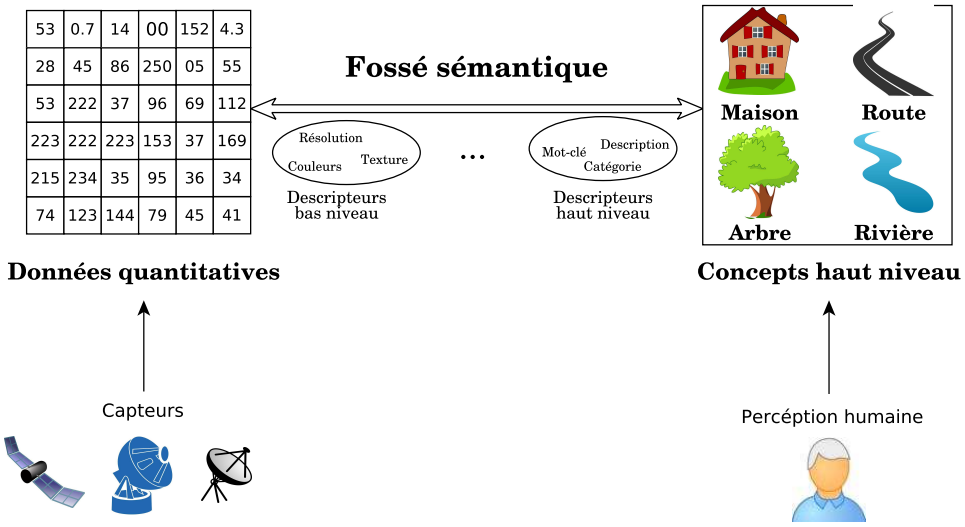


FIGURE 5.1 – Illustration du problème du fossé sémantique entre les données quantitatives et les concepts de haut niveau

La figure 5.1 illustre quelques éléments de la problématique du fossé sémantique. Elle montre d'un coté, des capteurs (satellites, appareils photos, enregistreur vidéo, ...) qui fournissent des données numériques qui sont dénuées de sens. Et de l'autre coté, la perception que peut en avoir l'expert

et qui lui permet d'associer des concepts de haut niveau (Arbre, Maison, Rivière, Route . . .) à ces données.

Le passage du numérique vers le symbolique est le cœur de la problématique du fossé sémantique. Généralement, il n'existe pas de liens directs entre les données numériques et les concepts de haut niveau [LIU et al., 2007].

Il est à remarquer dans la figure la présence de quelques termes (Couleurs, Texture, Résolution, ...) à coté des données quantitatives, qui représentent des **descripteurs de bas niveau**. Les techniques de traitement d'images disponibles aujourd'hui peuvent extraire des informations sur les couleurs, sur la résolution des images ou encore des informations par rapport à la texture des pixels. Ces traitements s'appuient la plupart du temps sur les métadonnées pour produire ces descriptions, qui restent de bas niveau, mais qui introduisent une première couche sémantique qui vient contextualiser les données. Ces descripteurs ne sont pas suffisants pour réduire le fossé sémantique mais constituent une première étape dans sa résolution. Ils peuvent être vus comme un premier niveau d'abstraction. C'est pour cette raison que nous les avons placés à l'extrémité de la flèche représentant le fossé sémantique.

À l'autre extrémité de la flèche, à coté des concepts de haut niveau, nous pouvons voir les termes "Mot-clé", "Description" et "Catégorie". Ce sont des **descripteurs de haut niveau**, proches de la perception subjective de l'expert, qui sont souvent obtenus par annotation manuelle et permettent de lier les contenus aux concepts. Nous les avons placé à l'autre extrémité en raison de leur proximité avec les concepts de haut niveau. Ils représentent un niveau d'abstraction supérieur.

L'obtention de ces deux formes de description est une étape primordiale dans la réduction du fossé sémantique, mais elle ne représente qu'une partie de la solution. Une part importante du travail consiste à formaliser la relation entre les deux niveaux d'abstraction (descripteurs de bas niveau et de haut niveau) et à automatiser le passage d'un niveau à l'autre.

Les ontologies figurent parmi les solutions exploitées par la communauté scientifique pour remédier à ce problème. Leurs apports en matière de modélisation, de contextualisation et de raisonnement sur de la connaissance en ont fait des outils privilégiés pour la réduction du fossé sémantique. Nous présentons dans la section (5.4) différentes propositions qui ont fait usage des ontologies à cette fin.

5.3 Approches d'analyse d'images satellites

La vision par ordinateur est devenu au cours des années un champ de recherche très actif ; et regroupe les techniques permettant d'analyser des images à l'aide de méthodes informatiques. Dans le contexte de la télédétection, deux grandes familles d'approches sont distinguées, à savoir les approches basées pixel et les approches basées région.

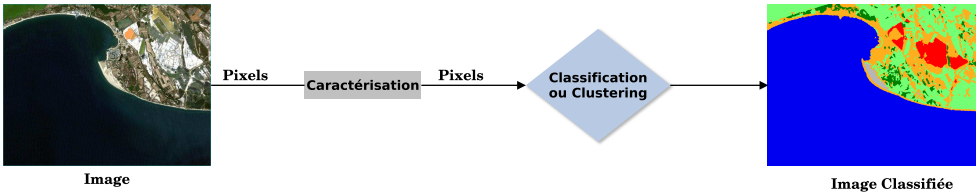


FIGURE 5.2 – Schéma général des approches d'analyse basées pixels

5.3.1 Approches basées pixel

Les approches basées pixel analysent directement les pixels et les considèrent comme les seuls porteurs d'information. À ce titre les pixels sont les entités traitées par les algorithmes de classification ou de clustering [JENSEN, 1986].

Avant analyse, les pixels peuvent être enrichis par les valeurs provenant d'indices spectraux et/ou de texture. Ces indices sont calculés à partir des bandes spectrales de l'image. L'objectif de cet enrichissement est d'améliorer la caractérisation des pixels par l'ajout de nouveaux attributs que les télédétecteurs jugent plus pertinents pour détecter de l'eau ou de la végétation par exemple. À l'issue de cette étape d'enrichissement, les entités à analyser restent cependant les pixels.

L'image est donc vue comme un ensemble de données $\mathcal{X}$, composé de n pixels $\{x_i\}_{i=1}^{n} \in \mathbb{R}^d$, où chaque pixel est décrit par d variables, qui correspondent aux valeurs des pixels sur les différentes bandes spectrales, possiblement augmentés par les valeurs des indices calculés.

La figure 5.2 schématise le principe général du déroulement de ces approches. À l'issue du processus d'analyse, chaque pixel va être attaché à un cluster ou une classe.

Dans le cadre de la classification supervisée, les pixels sont associés aux étiquettes des classes fournies par l'expert. À l'opposé, quand il s'agit du clustering, l'algorithme associe les pixels à des clusters non étiquetés. L'expert a par la suite la possibilité d'attribuer des étiquettes à ces regroupements dès lors qu'ils représentent une réalité thématique.

5.3.2 Approches basées régions

Les approches basées région consistent à segmenter l'image avant sa classification ou son clustering [HAY et CASTILLA, 2008]. L'idée principale derrière l'utilisation de la segmentation réside dans la construction d'agrégats de pixels homogènes et connexes, appelées segments, à partir de l'image afin d'obtenir des entités qui correspondent à des objets géographiques avec une réalité thématique à l'exemple des parcelles agricoles, des bâtiments ou encore des routes.

Après la segmentation (manuelle ou automatique) de l'image, les segments sont caractérisés par de multiple attributs à l'exemple de la taille, de la forme ou de données contextuelles. Ces segments sont les entités qui vont

être traitées par l'algorithme de classification ou de clustering. Il faut noter que le nombre d'instances à classifier correspond au nombre de segments et non au nombre des pixels. $X = \{x_j\}_{j=1}^{m}$, où m représente le nombre des segments.

À l'issue de la classification, chaque segment de l'image se voit assigné une classe ou un cluster. La figure 5.3 illustre d'une manière générale les différentes étapes de ce type d'approche.

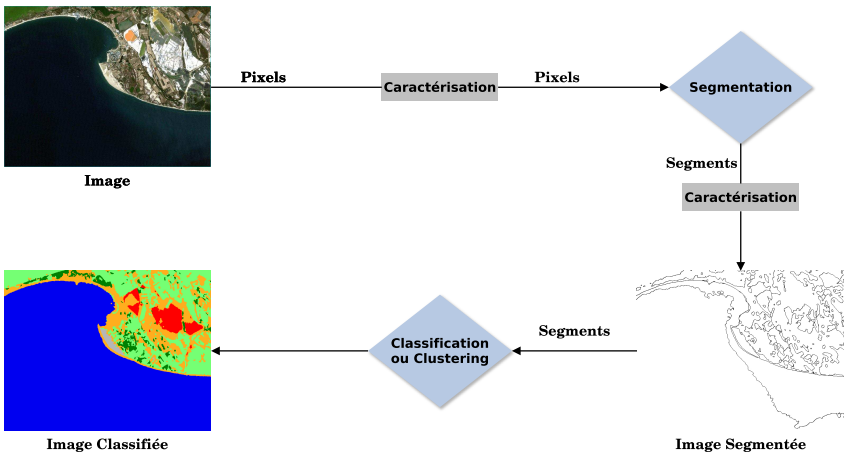


FIGURE 5.3 – Schéma général des approches d'analyse basées régions

5.3.3 Avantages et limites des deux approches d'analyse

Les approches basées pixels ont été les premières à être étudiées dans la littérature [RICHARDS et RICHARDS, 1999]. Elles restent aujourd'hui largement utilisées par les experts dans plusieurs contextes. Avec l'évolution de la résolution des images et l'apparition des images satellites à très haute résolution. Les approches basées région ont gagné en popularité et sont quelques fois vues comme des solutions plus appropriées.

L'avantage des approches basées région réside dans l'extraction de nouvelles entités qui peuvent être plus proches des concepts recherchés. Il faut cependant noter que la segmentation est une étape très complexe. Il n'existe pas actuellement d'algorithme de segmentation idéal. Les approches basées pixels sont quant à elles plus précises dans leur classification, puise qu'elles agissent au niveau du pixel.

Pour résumer, aucun des deux paradigmes ne fournit une solution parfaite pour l'analyse d'images satellites. Les deux types d'approches ont leurs avantages et inconvénients. Il n'existe pas d'approche miracle adaptée à tous les problèmes d'analyse des images satellites. L'adoption d'une des approches va grandement dépendre des besoins de l'expert. Plusieurs paramètres peuvent cependant l'aiguiller vers son choix. Notamment le type des classes thématiques recherchées et la résolution de l'image analysée.

Si le but recherché est d'identifier des classes d'occupation du sol très générales comme l'eau, la végétation ou les surfaces minérales ; effectuer

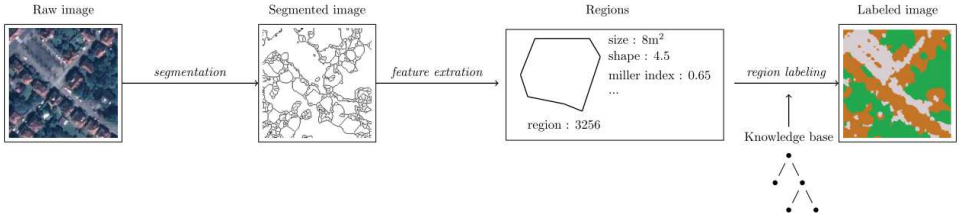


FIGURE 5.4 – Processus d'étiquetage par ontologie des segments d'une image satellite, par FORESTIER et al. (2012)

une segmentation avant de classifier l'image peut parfois détériorer la qualité des résultats à cause de la propagation des erreurs. Une classification directe des pixels est plus judicieuse dans ce cas de figure.

Par contre, l'identification de classes d'occupation du sol comme des parcs ou des lacs nécessite une segmentation préalable de l'image. La segmentation permet d'extraire des segments qui peuvent être caractérisés par leurs formes et compositions. Cette caractérisation supplémentaire peut aider à faire correspondre les segments à ce type de classes, qui ne peuvent pas être détectées au niveau pixel.

5.4 Connaissances pour l'interprétation et la classification d'images

L'un des principaux verrous scientifiques du processus d'extraction des connaissances reste donc l'automatisation de l'interprétation du contenu de l'image. L'une des pistes de recherche les plus explorées est de privilégier la modélisation des connaissances expertes afin de réduire le fossé sémantique sans toutefois faire appel à un expert à chaque interprétation d'image. Ces travaux [DURAND et al., 2007 ; NEUMANN et MÖLLER, 2008 ; GÓMEZ-ROMERO et al., 2011 ; FALOMIR et al., 2011 ; ANDRES, ARVOR et PIERKOT, 2012 ; FORESTIER et al., 2012] ont montré la capacité des systèmes à base de connaissances à fournir une interprétation de haut niveau en s'appuyant sur les connaissances expertes. Il est à noter une grande variété dans les rôles joués par les ontologies.

Ainsi, les approches des travaux cités font appel aux ontologies à différents moments de l'analyse, soit pour qualifier les objets extraits des phases de clustering ou de segmentation, soit pour directement décrire les pixels de l'image [ANDRÉS, 2013]. De même les ontologies sont parfois peuplées par les éléments des images qui en deviennent les individus.

A contrario, les ontologies ne sont pas matérialisées à partir des images mais sont plutôt exploitées pour annoter sémantiquement les éléments extraits des images. Les services de raisonnement ne sont alors pas mis à contribution.

DURAND et al. (2007) et FORESTIER et al. (2012) ont ainsi proposé une approche pour l'étiquetage des segments d'une image à partir d'une ontologie formalisant des concepts relatifs à des objets géographiques.

La figure 5.4 (extraite de FORESTIER et al. (2012)) illustre le processus général de la démarche. Deux étapes de traitement de données précèdent l'étiquetage par ontologie. Dans un premier temps, l'image est segmentée au moyen de l'algorithme de segmentation de DERIVAUX et al. (2006), paramétré manuellement [DURAND et al., 2007]. Dans un deuxième temps, les segments obtenus sont caractérisés par de nouveaux attributs, spectraux et spatiaux, à l'exemple d'indices de végétation et de surface. Les segments enrichis et les concepts des ontologies sont ensuite comparés à l'aide d'une procédure spécifique développée par les auteurs [DURAND et al., 2007 ; FORESTIER et al., 2012] qui en évalue la similarité et qui attribue un score à chaque paire (segment, concept de l'ontologie). Le segment est alors étiqueté par le concept avec lequel il forme la paire ayant le score le plus élevé.

DURAND et al. (2007) ont mené des expérimentations sur une image QuickBird [1] de la région de Strasbourg et en exploitant quatre concepts thématiques (végétation, eau, bâtiment au toit orange et route). FORESTIER et al. (2012) ont poursuivi les expérimentations en élargissant les jeux de données à quatre images QuickBird dont une sur la zone de Marseille.

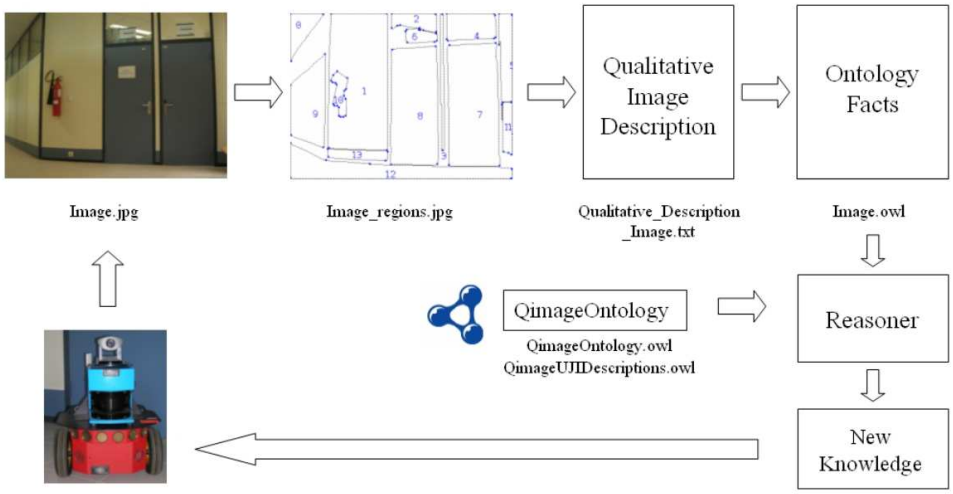


FIGURE 5.5 – Approche à base des logiques de description pour la description qualitative d'images [FALOMIR et al., 2011]

FALOMIR et al. (2011) ont proposé une approche à base d'ontologie pour automatiser l'interprétation d'images numériques. Cette interprétation s'est appuyée sur la capacité de leur approche à décrire qualitativement les objets de l'image afin de les relier aux concepts de l'ontologie. Ils ont construit à cet effet une ontologie d'image, *QImageOntology*, en adoptant une conceptualisation à trois niveaux d'abstraction et une formalisation des connaissances en OWL.

1. https ://en.wikipedia.org/wiki/QuickBird

Le niveau d'abstraction le plus élevé représente la conceptualisation de référence, qui contient les connaissances valides quel que soit le domaine d'interprétation (définitions des formes géométriques ou couleur ...).

Le niveau intermédiaire contient quant à lui les connaissances contextuelles formalisant les concepts spécifiques au domaine d'application (ici les concepts porte, sol, ...)

Finalement, le niveau le plus bas concerne les assertions sur les images. C'est à ce niveau que les descripteurs de bas niveau de l'image sont raccrochés à des propriétés qualitatives. Les deux premiers niveaux d'abstraction sont fixés indépendamment de l'image traitée. Le niveau assertionnel, qui concerne les faits de l'image est quant à lui obtenu par l'algorithme de segmentation [FELZENSZWALB et HUTTENLOCHER, 2004] paramétré manuellement, comme le montre la vue générale de l'approche (figure 5.5).

Plus récemment, d'autres travaux ont été publiés par les mêmes auteurs en proposant de nouvelles expérimentations tout en améliorant le temps de calcul nécessaire à l'interprétation des images de scènes [FALOMIR et OLTEȚEANU, 2015]. D'autres travaux ont aussi exploré l'utilisation des logiques de description pour l'interprétation des scènes d'images multimédia, à l'exemple de NEUMANN et MÖLLER (2008). Cependant, ces travaux n'ont pas précisé les techniques exploitées pour obtenir les descriptions sémantiques de l'image.

Les travaux d'ANDRES, ARVOR et PIERKOT (2012) ont proposé l'utilisation du raisonnement à base des logiques de description pour étiqueter sémantiquement les objets, pixels ou segments, d'une image satellite. Les auteurs se sont inspirés de la conceptualisation à trois niveaux d'abstraction proposée par FALOMIR et al. (2011) pour modéliser les connaissances et en particulier distinguer l'ontologie de l'image et l'ontologie de la connaissance experte en télédétection. Une représentation modulaire de la connaissance en OWL 2 a ainsi été définie pour le domaine de la télédétection.

Selon les entités individuelles considérées, segment ou pixel, deux approches sont mises en place. Dans la première approche, des segments sont préalablement extraits d'images LANDSAT à l'aide d'une préclassification par pixel proposée par BARALDI et al. (2006). L'importance est donnée à l'élicitation de régions qui sont le reflet d'objets cohérents de l'image sans trop s'attarder sur les problèmes liés à la recherche d'une bonne segmentation [BLASCHKE, LANG et HAY, 2008]. Les segments obtenus deviennent les assertions de l'ontologie et un raisonneur est alors exploité pour mettre en œuvre des mécanismes de déduction et en particulier assigner les segments à une classe (à rapprocher de classes d'occupation du sol) au travers d'un service dédié dit de reconnaissance d'instances. Dans la seconde approche [ANDRÉS, 2013], Andres se démarque des analyses basées région et exploite directement les pixels en tant qu'instances de l'ontologie sans segmentation préalable de l'image. Le raisonneur est alors appliqué aux pixels pour en déduire les classes d'appartenance. L'ontologie mise en jeu est construite à partir des règles et des 46 catégories spectrales définies dans [BARALDI et al., 2006]. La complexité des descriptions mises en jeu et le nombre considérable de pixels à évaluer ont conduit à une reconnaissance des pixels pris

séparément. L'approche a permis ainsi de contourner la complexité des logiques de description et effectuer un étiquetage des pixels dans un temps linéairement proportionnel au nombre des pixels.

Dans les propositions où une phase de segmentation[2] précède l'interprétation par connaissances des images. Le recours à cette étape est justifiée par le besoin d'extraire de nouvelles entités construites à partir des pixels avant d'entamer l'interprétation. Dans les images satellites à très haute résolution, quelques concepts thématiques cibles (parking, toit de maison...) n'apparaissent pas au niveau du pixel mais sont formés par des groupes de pixels (segments). Cependant, l'obtention d'une bonne segmentation est une tâche ardue. Elle demande une forte implication de l'expert et l'amélioration des techniques de segmentation reste une problématique de recherche ouverte.

5.4.1 Les connaissances pour l'enrichissement des descriptions

La section précédente s'est intéressée à l'exposition des travaux exploitant les systèmes à base de connaissances pour automatiser l'interprétation de données quantitatives (essentiellement des images) tout en réduisant le fossé sémantique. D'autres travaux dans la littérature, tous aussi intéressants ont exploité différemment les ontologies, notamment en les utilisant pour enrichir les descriptions textuelles des images par des conceptualisations de références.

Les travaux de Bannour et Hudelot [BANNOUR et HUDELOT, 2011 ; BANNOUR et HUDELOT, 2012 ; BANNOUR et HUDELOT, 2014] ont utilisé les ontologies pour annoter des images multimédias et enrichir leurs descriptions textuelles. Pour cela, les auteurs ont proposé une approche qui s'appuie sur les annotations des images et leurs métadonnées[3] pour construire une ontologie multimédia floue formalisée en OWL. Cette ontologie est enrichie par des annotations obtenues en exploitant les hiérarchies, les relations et les contextes spécifiés pour mesurer la similarité entre les images et les concepts de l'ontologie initiale [BANNOUR et HUDELOT, 2012]. L'approche [BANNOUR et HUDELOT, 2014 est résumée dans la figure 5.6, où l'on peut distinguer le processus général de cette construction. L'objectif premier de ces travaux est d'améliorer la pertinence des résultats des recherches d'un catalogue d'images, en améliorant les annotations attribuées aux images.

Dans le domaine de l'imagerie médicale, les travaux de HUDELOT, ATIF et BLOCH (2008) ont exploré l'utilisation des connaissances pour l'interprétation d'images radiologiques du cerveau. HUDELOT, ATIF et BLOCH (2008) ont construit une ontologie de relations spatiales, qu'ils ont intégré à une ontologie du domaine existante FMA [ROSSE et MEJINO JR, 2003], conceptualisant les modèles anatomiques de base. Dans ce travail, les domaines concrets [LUTZ, 2003] ont permis de relier les concepts de l'ontologie médicale, à leurs représentations qui admettent du flou [BLOCH, 2005 ; ATIF

2. Pour rappel, la segmentation est le processus de décomposition de l'image (au sens numérique 4.2) en segments constitués de pixels homogènes et spatialement connexes.

3. Métadonnée : Dans sa plus simple définition, une donnée sur la donnée.

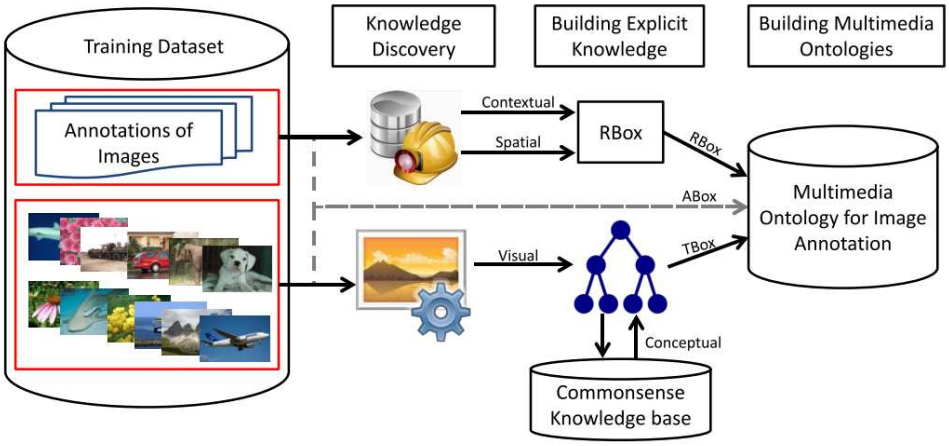


FIGURE 5.6 – Approche proposée pour la construction d'ontologies multimédia dédiées aux annotations d'images [BANNOUR et HUDELOT, 2014]

et al., 2007]. Ces travaux ont démontré les possibilités offertes par le raisonnement, dans le cadre opérationnel des images radiologiques, pour déduire les relations spatiales et les positions relatives des segments dans l'image. D'autres travaux des mêmes auteurs ont proposé une autre approche ayant le même objectif en se basant cette fois sur les graphes pour la représentation des connaissances et la modélisation des relations.

5.4.2 Autres propositions à base d'ontologie et apprentissage

D'autres travaux [SCHOBER, HERMES et HERZOG, 2004 ; SHEEREN et al., 2006 ; PUISSANT et al., 2006], plus éloignés de notre problématique de recherche, ont investi les possibilités de l'utilisation mutuelle de l'apprentissage et des connaissances formalisées. Ces travaux se sont principalement intéressés à l'utilisation de l'apprentissage pour aider à la construction d'ontologies. La formalisation des connaissances reste un verrou important en ingénierie des connaissances [CULLEN et BRYMAN, 1988 ; BOICU et al., 2001], ces approches cherchent à faciliter la phase de formalisation en offrant à l'utilisateur une vision analytique des données traitées. Cette vision s'appuie sur les résultats de l'apprentissage supervisé pour alimenter les connaissances ou permettre à l'expert de trouver des correspondances entre les données et le concepts.

SHEEREN et al. (2006) et PUISSANT et al. (2006) ont exploré l'utilisation de l'apprentissage pour découvrir des règles logiques de classification d'images satellite urbaines. L'acquisition des règles se base sur un algorithme génétique introduit par KOZA (1992). Après la segmentation de l'image, plusieurs segments sont étiquetés par l'expert et fournis à l'algorithme génétique. Les règles de classification sont apprises successivement pour chaque classe. A chaque étape, le but est d'apprendre à discriminer une classe par rapport à toutes les autres (*one vs all*). Les règles apprises ont

ensuite été formalisées et utilisées pour classifier une image Quickbird de la région de Strasbourg.

Dans la même perspective, les travaux de MAILLOT et THONNAT (2008) et MAILLOT, THONNAT et BOUCHER (2004) ont proposé un cadre pour l'enrichissement d'ontologie par apprentissage afin de créer un système de reconnaissance d'objets complexes pour des images. L'ontologie contient, dans un premier temps, les concepts du domaine ainsi que la définition des relations spatiales. La première étape consiste à annoter manuellement les segments par les concepts. La phase d'apprentissage intervient par la suite pour trouver automatiquement les relations entre les descripteurs quantitatives de bas niveau et les concepts symboliques de l'ontologie. Les connaissances enrichies sont utilisées par la suite pour étiqueter des nouveaux objets.

SCHOBER, HERMES et HERZOG (2004) ont développé un système pour la recherche d'image basée sur le contenu. Le système proposé est basé principalement sur deux composantes. La première est l'ontologie du domaine décrite par le formalisme DAML+OIL [HORROCKS, 2002] (ancêtre du standard OWL). Cette ontologie contient les concepts ciblés par les recherches (montagne, nuage, arbre, ...). La deuxième composante repose sur des classifieurs supervisés entraînés à reconnaître les segments des images multimédia du système. Une fois la classification effectuée, les résultats sont transformés en axiomes et viennent enrichir la base de connaissances. Un raisonneur est par la suite utilisé pour corriger les incohérences de la classification, en se basant sur les spécifications de l'ontologie. Cela permet d'éviter d'avoir des étiquetages illogiques de segments de la même image. Le raisonneur est aussi utilisé dans ce système pour répondre aux requêtes des utilisateurs. L'un des points clés du système est la bonne segmentation des images. Comme soulevé par les auteurs eux-mêmes [SCHOBER, HERMES et HERZOG, 2004], les images doivent être parfaitement segmentées avant d'être présentées aux classifieurs. Cela peut fortement pénaliser l'apprentissage et donc introduire des incohérences qui peuvent rendre le raisonnement impossible.

5.4.3 Discussion des travaux

Les travaux cités utilisent de différentes manières la connaissance experte, mais aucune des approches trouvées dans la littérature n'intervient pendant une phase de clustering. La principale utilisation des ontologies reste l'interprétation sémantique des objets extraits, ou des instances des données. DURAND et al. (2007) et FORESTIER et al. (2012) 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 al. (2011) et ANDRES, ARVOR et PIERKOT (2012) 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 et enrichir l'étiquetage obtenu par les connaissances.

L'amélioration constante des performances des services d'inférence fournis par les raisonneurs a grandement favorisé l'usage des logiques de description et par extension de OWL. Ces services, allant par exemple de l'inclusion de concepts au test d'appartenance d'une instance à un concept, ont tout d'abord été rendus efficaces pour répondre aux besoins grandissants du Web sémantique [BERNERS-LEE, HENDLER et LASSILA, 2001 ; SHADBOLT, BERNERS-LEE et HALL, 2006]. Cette dynamique aidant, les raisonneurs ont pris en charge des fragments de logiques de description de plus en plus expressifs à l'exemple de SROIQ(D), comme le montrent différentes études comparatives [GARDINER, HORROCKS et TSARKOV, 2006 ; ABBURU, 2012 ; EITER et al., 2014 ; MATENTZOGLU et al., 2015]. Même si les premiers développements des raisonneurs concernaient en premier lieu les applications liées au domaine du Web sémantique, et donc essentiellement le traitement d'informations symboliques ; les propositions présentées dans ce chapitre démontrent que les raisonneurs peuvent être exploités pour jouer un rôle essentiel dans l'interprétation automatique de tout type de connaissance formalisée.

Dans le contexte de la télédétection, le fait de disposer de services de raisonnement efficaces couplés des logiques de description expressives intégrant en particulier la représentation de domaines concrets, participe à la réduction du fossé sémantique. Un point important à souligner porte sur la grande souplesse de l'usage qui peut être fait de ces services. Si l'on considère qu'une image satellite va être analysée à l'aide de différents traitements qui vont lui être appliqués en séquentiel ou en parallèle [4], il est à noter que les services de raisonnement peuvent intervenir à tout moment apportant ainsi de la modularité et de la généricité dans l'ensemble du processus d'analyse et de caractérisation des objects de l'image.

Il ressort de notre étude des différents travaux usant des connaissances dans l'analyse et l'interprétation des images (et d'autres données numériques) que les ontologies peuvent être considérées comme une solution pérenne pour réduire le fossé sémantique et automatiser l'interprétation des données sans recours permanent à l'expert. La capacité des ontologies à formaliser la composante des connaissances indépendamment de l'implémentation algorithmique apporte de la modularité et de la généricité aux différentes chaînes de traitements. Cependant un verrou important réside dans la difficulté à acquérir les connaissances [BOICU et al., 2001 ; CULLEN et BRYMAN, 1988]. En effet, l'accès à l'interprétation automatisée par raisonnement demande un effort de modélisation et de formalisation préalables. Comme souligné par SESTER (2000), les experts, accompagnés par les modélisateurs, doivent pouvoir décrire précisément les concepts d'intérêt. Combiné avec la nature déductive du processus de raisonnement, ce

4. L'analyse d'images implique la mise en œuvre de chaînes de traitement

verrou impose deux limitations aux systèmes à base de connaissances :

— Une incomplétude des spécifications qui implique que les instances ne vont pas toutes appartenir à un concept cible ;
— Une incapacité à découvrir de nouveaux concepts.

5.5 Intégration des connaissances en apprentissage

Les premiers travaux de recherche portant sur la classification des données ont envisagé principalement deux cadres d'études. Un cadre supervisé, où les données étiquetées sont suffisantes pour apprendre le modèle de prédiction des classes, et un cadre non-supervisé, où l'apprentissage se fait uniquement sur des données non-étiquetées, et ne tient pas compte des données étiquetées possiblement disponibles.

Ces dernières années, l'apprentissage semi-supervisé est devenu un champ de recherche particulièrement actif [CHAPELLE, SCHÖLKOPF et ZIEN, 2006 ; ZHU, 2005 ; GRIRA, CRUCIANU et BOUJEMAA, 2004]. Les méthodes de classification (au sens large) à base d'apprentissage semi-supervisé peuvent être séparées en deux catégories :

— *Classification* semi-supervisée : Utilisation de données non-étiquetées dans la phase d'apprentissage pour améliorer la *classification* [CHAPELLE, SCHÖLKOPF et ZIEN, 2006]
— *Clustering* semi-supervisé : Utilisation d'information *a priori* (instances étiquetées ou contraintes) pour améliorer le *clustering* [DAVIDSON et BASU, 2007]

***Classification* semi-supervisée**

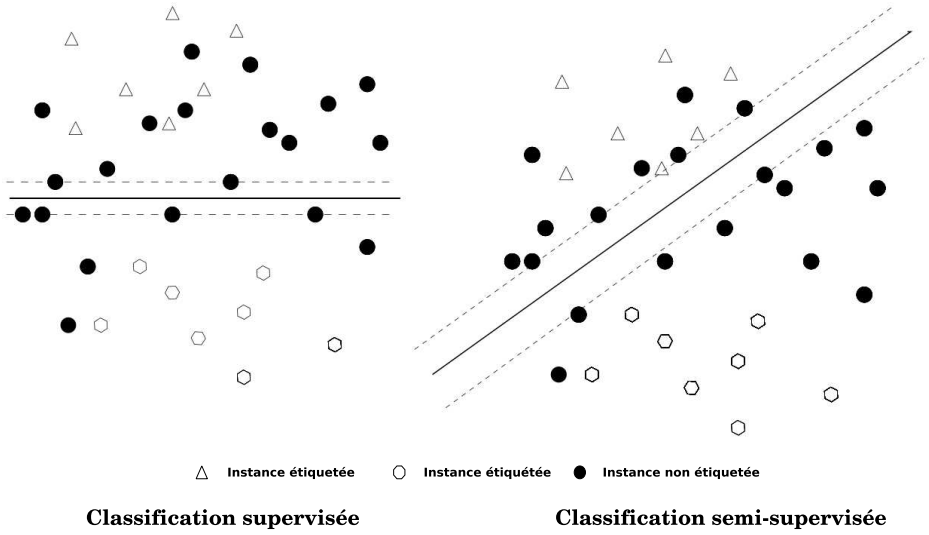


FIGURE 5.7 – Illustration de l'apport des données étiquetées pour la classification, figure inspirée de BENNETT et DEMIRIZ (1999)

La classification semi-supervisée regroupe les méthodes qui essayent d'exploiter les données non-étiquetées afin d'améliorer la classification [CHAPELLE, SCHÖLKOPF et ZIEN, 2006]. Les algorithmes supervisés utilisent uniquement les données étiquetées pour apprendre leurs modèles. Cependant, comme nous l'avons mentionné à plusieurs reprises, l'obtention d'exemples étiquetés consomment énormément de temps et peut être très difficile. Les algorithmes de classification semi-supervisée utilisent mutuellement des données étiquetées et non-étiquetées dans la phase d'apprentissage. Le problème de la *classification* semi-supervisée revient donc à prédire la classe de données non-étiquetées en utilisant un jeu de données $\mathcal{X} = \mathcal{X}_l \cup \mathcal{X}_u$, avec $\mathcal{X}_l = \{(x_i, y_i)\}$ l'ensemble des instances étiquetées par $y_i \in Y = \{y\}_{i=1}^{k}$, et $\mathcal{X}_u = \{(x_j)\}$ l'ensemble des instances non-étiquetées.

Les premiers travaux pointant l'intérêt de l'utilisation de données non étiquetées sont apparus bien avant la formalisation du problème de la classification semi-supervisée. Vapnik [VAPNIK et CHERVONENKIS, 1974 ; VAPNIK et STERIN, 1977] avait déjà relevé l'intérêt de l'utilisation des données non étiquetées dans le cadre de ce qu'il avait appelé l'*apprentissage transductif*. À la différence de l'apprentissage inductif, l'objectif ici n'est pas de produire un modèle qui va être généralisé à tout l'espace de données et donc chercher à minimiser l'erreur de généralisation [VAPNIK, 1995], mais plutôt de classifier uniquement les instances de la base de test.

Dans leurs travaux sur la classification de données linéairement séparables, JOACHIMS (1999) et BENNETT et DEMIRIZ (1999) ont illustré l'apport de l'intégration des données non-étiquetées pour bien apprendre l'hyperplan des séparateurs à vaste marge (SVM) [BOSER, GUYON et VAPNIK,

1992; CORTES et VAPNIK, 1995]. La figure 5.7 montre qu'en prenant en compte les données non étiquetées, le choix de l'hyperplan séparateur peut être fortement amélioré.

La plupart des approches de classification semi-supervisée s'appuient sur des hypothèses *a priori* concernant la nature des données [CHAPELLE, SCHÖLKOPF et ZIEN, 2006]. L'hypothèse de la régularité est parmi les hypothèses les plus importantes, et est aussi partagée par la classification supervisée. Elle stipule que si deux instances x_i et x_j sont proches (appartiennent à la même zone de forte densité), alors elles partagent probablement la même étiquette y_l. La deuxième hypothèse concerne la distribution des données. On suppose que la distribution des données d'apprentissage (étiquetées et non-étiquetées) est assez proche du partitionnement souhaité par la classification. En d'autres termes, si x_i et x_j sont dans le même cluster, alors ils appartiennent probablement à la même classe. La troisième hypothèse porte sur la séparation des données en se basant sur la densité. Les contours des classes se trouvent dans les zones à faible densité. Cela permet, entre autres, de s'assurer que les instances non-étiquetées proches apparaissant dans des zones à faible densité peuvent se retrouver dans différentes classes. La quatrième et dernière hypothèse est celle de la dimensionnalité. Les données en grande dimension peuvent être exprimées dans des espaces à plus petite dimension.

En se basant sur une partie ou l'intégralité de ces hypothèses. Plusieurs méthodes ont été proposées dans la littérature, notamment les premières méthodes semi-supervisées à base des SVM [JOACHIMS, 1999; BENNETT et DEMIRIZ, 1999], des méthodes génératives [SEEGER, 2000], des méthodes de propagation à base de graphes [ZHU, GHAHRAMANI et LAFFERTY, 2003] ou encore des méthodes semi-supervisées se basant sur la minimisation de l'entropie [GRANDVALET et BENGIO, 2004].

Depuis les années 90, la classification semi-supervisée est devenue un champ de recherche très actif. Cependant, similairement à la *classification* supervisée, la classification semi-supervisée ne s'applique que si une connaissance *a priori* existe sur **toutes** les classes cibles [CHAPELLE, SCHÖLKOPF et ZIEN, 2006]. En d'autres termes, des exemples étiquetés représentant toutes les classes doivent être présents dans le jeu de données d'apprentissage. Cette condition n'est pas toujours satisfaite, dans le cas des images satellite par exemple, l'expert n'a pas toujours connaissance de toutes les classes thématiques présentes et ne peut pas toujours fournir des exemples sur toutes les classes. Dans ce cadre, une alternative est alors le clustering semi-supervisé [DAVIDSON et BASU, 2007].

5.5.1 Clustering semi-supervisé

Comme nous l'avons expliqué à la fin du chapitre 3, le clustering fait émerger un partitionnement des données qui n'est pas toujours celui attendu par l'utilisateur. L'activité de partitionnement peut dans ce contexte être guidée pour aboutir à un résultat qui est plus fidèle aux attentes de l'utilisateur. L'idée est donc de prendre en compte la vision exprimée par

l'utilisateur tout en permettant au clustering de découvrir de nouvelles structures. Le clustering semi-supervisé est défini par BILENKO, BASU et MOONEY (2004) comme l'utilisation d'exemples étiquetés ou de contraintes pour améliorer le clustering.

L'un des plus gros avantages du clustering semi-supervisé par rapport à la classification semi-supervisée réside dans sa capacité à gérer l'incomplétude des échantillons disponibles par rapport aux clusters à découvrir. A la différence de la classification semi-supervisée, dans le cadre du clustering semi-supervisé, les exemples étiquetés ou les contraintes fournies peuvent ne pas concerner toutes les classes d'intérêt. Par exemple, dans une classification à quatre clusters, les contraintes peuvent être spécifiées uniquement sur deux clusters. Ceci donne un avantage notable au clustering semi-supervisé, notamment pour des applications où les données étiquetées sont difficiles à obtenir ou/et les connaissances *a priori* sont incomplètes.

Deux approches de clustering semi-supervisé existent dans la littérature, des approches à base d'apprentissage de métriques [XING et al., 2003 ; YANG et JIN, 2006 ; KULIS, 2012 ; BELLET, HABRARD et SEBBAN, 2013] et des approches à base de contraintes [DAVIDSON et BASU, 2007 ; BASU, DAVIDSON et WAGSTAFF, 2008 ; COVOES, HRUSCHKA et GHOSH, 2013].

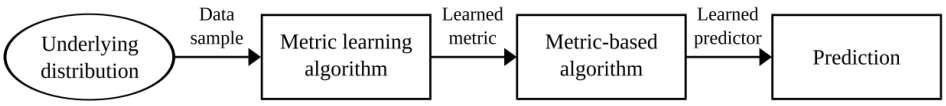
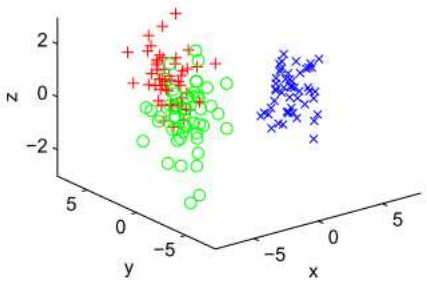
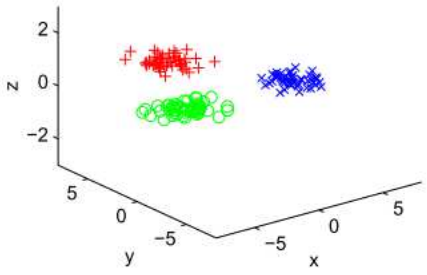


FIGURE 5.8 – Processus général d'apprentissage à base de métriques [BELLET, HABRARD et SEBBAN, 2013]

Les approches à base de métriques utilisent les exemples étiquetés ou les contraintes *a priori* pour apprendre une métrique. Cette métrique va par la suite être utilisée dans l'algorithme de clustering pour partitionner les données, comme le montre la figure 5.8 extraite de BELLET, HABRARD et SEBBAN (2013). L'apprentissage de la distance demande le calcul d'une matrice sur tout l'espace des données et de l'utiliser pour pondérer les distances dans la phase du clustering. L'objectif est d'apprendre une nouvelle distance qui va minimiser la distance entre les données de la même classe/-cluster, et l'exploiter ensuite pour séparer les données. La figure 5.9, extraite de XING et al. (2003), illustre cette pondération par un exemple de projection des données avant et après l'apprentissage de la métrique. Le calcul de la matrice de pondération sur tout l'espace de données peut causer des problèmes de passage à l'échelle qui peut limiter l'application de ces méthodes à des grandes masses de données comme les images satellite.



Projection des données suivant la distance euclidienne (Les couleurs indiquent les classes)



Projection des données suivant la distance pondérée après apprentissage de la métrique

FIGURE 5.9 – Illustration de l'apprentissage d'une métrique par la projection des données suivant la métrique apprise [XING et al., 2003]

5.5.2 Clustering par contraintes

Dans le cadre clustering par contraintes, les connaissances *a priori* sont exprimées sous-forme de contraintes opérées sur les instances pendant le clustering. Les travaux sur le clustering par contraintes sont relativement récents, les travaux de WAGSTAFF et CARDIE (2000) ont été les premiers à formellement définir les contraintes au niveau des instances. Plusieurs approches ont été proposées par la suite dans la littérature [WAGSTAFF et al. (2001), BILENKO, BASU et MOONEY (2004), DAVIDSON et RAVI (2005a), DAVIDSON et RAVI (2005b), KULIS et al. (2009), ALLAB et BENABDESLEM (2011), WU et al. (2013), KHOREVA et al. (2014), ANAND et al. (2014) et WANG, QIAN et DAVIDSON (2014)].

Il existe principalement quatre types de contraintes pour le clustering [DAVIDSON et BASU, 2007] :

— la contrainte *must-link* : $ml(x_i, x_j)$ spécifie que deux instances, notées x_i et x_j, doivent se retrouver dans le même cluster final ;

— la contrainte *cannot-link* : $cl(x_i, x_j)$ spécifie que les deux instances ne doivent pas appartenir au même cluster ;

— la contrainte δ : Cette contrainte spécifie que pour toute instance x_i, lier avec des *must-link* toutes les instances $x_j \in \mathcal{X}$
Tel que $d(x_i, x_j) < \delta$;

— la contrainte ϵ : Cette contrainte spécifie que pour toute instance x_i, lier avec au minimum un *must-link* à une instance x_j
Tel que $d(x_i, x_j) \leq \epsilon$.

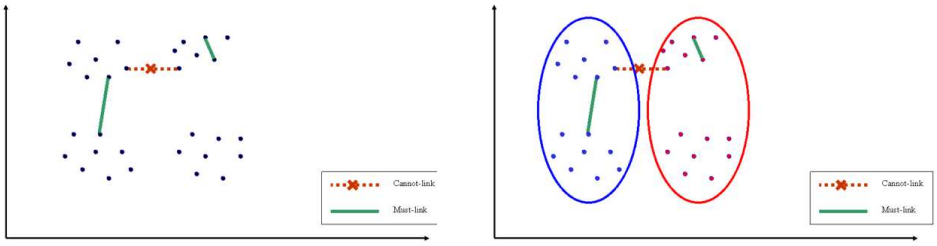
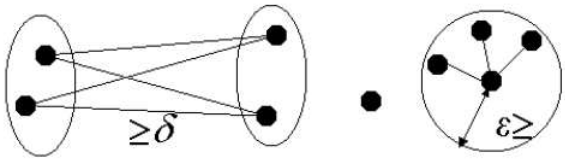


FIGURE 5.10 – Illustration des contraintes *must-link* et *cannot-link* [DAVIDSON et BASU, 2007]

La figure 5.10 illustre l'utilité des contraintes pour superviser le clustering. Dans l'exemple donné, il est à remarquer que le respect des contraintes permet d'obtenir un partitionnement fidèle à la vision de l'utilisateur. Les deux autres contraintes (δ et ϵ) représentent un regroupement de contraintes, où la contrainte δ peut être vue comme une contrainte de séparation minimale qui va permettre d'imposer que les instances de deux clusters différents doivent être séparées par une distance supérieure à δ (conjonction de contraintes ml). Tandis que la contrainte ϵ permet de spécifier que les instances séparées d'une distance inférieure ou égale à ϵ sont automatiquement liées par des contraintes *must-link* (disjonction de contraintes ml) (figure 5.11).



(a) Contraintes delta **(b) Contraintes epsilon**

FIGURE 5.11 – Illustration des contraintes δ et ϵ [DAVIDSON et BASU, 2007]

Les contraintes *must-link* sont transitives, réflexives et symétriques :

$$ml(x_i, x_j) \cap ml(x_j, x_k) \Rightarrow ml(x_i, x_k) \tag{5.1}$$

La transitivité des contraintes permet la formation de groupes connectés de *must-link* $GC = \{ml(x_i, x_j)|\forall i, j\}$. Ces groupes de contraintes sont eux-mêmes transitifs, cela implique que si deux instances x_i et x_j, appartenant respectivement à deux groupes connectés différents GC_i et GC_j, sont connectées par une contrainte *must-link*, alors toutes les instances de ces deux groupes sont aussi liées par des contraintes *must-link* :

$$x_i \in GC_i, x_j \in GC_j, ml(x_i, x_j) \Rightarrow \forall x_k \in GC_i, x_l \in GC_j; ml(x_k, x_l) \tag{5.2}$$

De la même manière, cette transitivité permet de déduire des contraintes *cannot-link* sur des groupes de contraintes sur la base de l'implication suivante :

$$x_i \in GC_i, x_j \in GC_j, cl(x_i, x_j), \Rightarrow \forall x_k \in GC_i, x_l \in GC_j; cl(x_k, x_l) \quad (5.3)$$

Différentes techniques ont été proposées dans la littérature pour intégrer les contraintes dans les algorithmes de clustering. La grande majorité utilise des contraintes *must-link* et *cannot-link*. Ces techniques reposent sur les modifications de différentes composantes des algorithmes de clustering, qui peuvent être regroupées en trois catégories [DAVIDSON et BASU, 2007 ; BASU, DAVIDSON et WAGSTAFF, 2008] :

— Modification de la phase de mise à jour de l'affectation des instances aux clusters ;
— Modification de la phase d'initialisation des clusters ;
— Modification de la fonction objective du clustering.

Clustering par contraintes sur l'affectation des instances

La vérification de l'affectation des instances est l'une des premières techniques proposées pour intégrer les contraintes [WAGSTAFF et al., 2001 ; SHENTAL et al., 2004]. 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 d'une instance à un cluster.

Ainsi, lors de l'affectation d'une instance x_i au cluster le plus proche Cl_i. Cette affectation 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 x_j liée par une contrainte *must-link* à x_i, et qui est déjà assignée à un autre cluster Cl_j, et qu'il n'existe aucune instance liée par une contrainte *cannot-link* avec x_i dans le cluster Cl_i. Si ces deux conditions ne sont pas satisfaites, l'algorithme passe au cluster suivant, jusqu'à la vérification de toutes les contraintes.

L'algorithme COP-KMEANS [WAGSTAFF et al., 2001] est une variante de K-Means (section 3.2.2) qui adopte cette technique.

La seule différence entre COP-KMEANS et K-Means se trouve dans l'étape d'affectation (algorithme 1, (3)) où l'affectation des instances à un cluster est conditionnée par la non-violation des contraintes spécifiées.

Avec cette modification, WAGSTAFF et al. (2001) ont montré l'efficacité de l'algorithme à analyser les traces GPS et à former des clusters avec des formes allongées pouvant détecter les trajets des voitures, ce que K-Means ne réussissait pas à faire.

L'une des limitations de cette technique de prise en compte des contraintes est sa sensibilité aux bruits dans les contraintes. En effet, l'applicabilité de l'algorithme suppose qu'il n'existe pas de contradiction dans les contraintes proposées.

Paramètres : Données $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$
Nombre des clusters : k
Contraintes *must-link* :
$\mathcal{ML} = \{ml(x_i, x_j)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$
Contraintes *cannot-link* : $\mathcal{CL} = \{cl(x_i, x_j)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$

Résultat : Partitionnement : $\mathbb{C}_l = \{Cl_1, ..., Cl_k\}$
Ou : $\emptyset$ clustering infaisable

(1). *Initialisation des clusters* $\mathbb{C}_l = Cl_1, ..., Cl_k$;
(2). *Affectation des instances*;
pour chaque $x_i \in \mathcal{X}$ **faire**
    Affecter x_i *à* $Cl_j \in \mathbb{C}_l$ ***Tel que*** ;

$$d(x_i, \mu_{Cl_j}) = argmin \sum_{j=1}^{k} \|x_i - \mu_{Cl_j}\|^2 \tag{5.4}$$

$$V_{contraintes}(x_i, Cl_j, \mathcal{ML}, \mathcal{CL}) = \textbf{Faux} \tag{5.5}$$

fin
(3). *Mise à jour des centroïdes*;
pour chaque $\mu_{Cl_j} \in \{\mu_{Cl_j}\}_{j=1}^{k}$ **faire**
    Calculer les centroïdes des clusters $\mathbb{C}_l$;

$$\mu_{Cl_j} = \frac{1}{\mid Cl_j \mid} \sum_{x_i \in Cl_j} x_i \tag{5.6}$$

fin
(4). *Répéter* (2) *et* (3) *jusqu'à convergence* ;
Fonction $V_{contraintes}(x_i, Cl_j, \mathcal{ML}, \mathcal{CL})$
    pour chaque $(x_i, x_j) \in \mathcal{ML}$ **faire**
        si $x_j \notin Cl_j$ **alors**
            Retourner **Vrai**;
        fin
    fin
    pour chaque $(x_i, x_j) \in \mathcal{CL}$ **faire**
        si $x_j \in Cl_j$ **alors**
            Retourner **Vrai**;
        fin
    fin
    Retourner **Faux**;

Algorithme 1 : Algorithme COP-KMEANS

D'autres algorithmes ont adopté la même technique, à l'exemple de COP-COWEB [WAGSTAFF et CARDIE, 2000], une variante de COBWEB [FISHER, 1987], d'une version contrainte d'*Expectation Maximization* proposée par SHENTAL et al. (2004) ou encore de l'algorithme CrTM [BELLAL, BENABDESLEM et AUSSEM, 2008], une version modifiée de SOM.

Contraintes sur l'initialisation des clusters

La seconde technique consiste à apporter une modification dans la phase d'initialisation de l'algorithme. L'initialisation est une étape importante qui influence fortement la formation des clusters. L'idée est d'utiliser les contraintes pour initialiser les clusters par les instances concernées et éventuellement veiller par la suite au respect de ces contraintes pendant le clustering.

Une variante de K-Means, appelée *Constrained-KMeans*, a été proposée par BASU, BANERJEE et MOONEY (2002). Elle adopte cette technique pour introduire les contraintes. Cet algorithme utilise des instances étiquetées comme connaissances *a priori*, notées $\mathcal{S} = \cup_{l=1}^{k} \mathcal{S}_l$. Ces instances peuvent être vues comme un groupe de contraintes *must-link* $GC = \{ml(x_i, x_j) \forall i, j\}$.

L'algorithme affecte les instances étiquetées aux clusters et calcule les centroïdes lors de l'initialisation. L'affectation des instances n'est pas remise en cause par le calcul des distances minimales avec les centroïdes des autres clusters. Les autres instances et le calcul des centroïdes prend en compte à la fois les instances nouvellement assignées et celles imposées. Toutes les étapes de cette proposition sont décrites dans l'algorithme 2.

Dans le même esprit, DAVIDSON et RAVI (2005a) ont proposé une variante par contraintes du clustering ascendant hiérarchique. Des groupes d'instance sont calculés lors de l'initialisation à partir des contraintes ml et ϵ en utilisant la procédure introduite dans [DAVIDSON et RAVI, 2005b]. L'algorithme calcule par la suite la hiérarchie des clusters tout en veillant au respect des affectations prédéfinies lors de l'initialisation.

La modification de la phase d'affectation et celle de l'initialisation ont été les premières approches à intégrer les contraintes au niveau du clustering. Ces approches améliorent significativement les résultats du clustering. Ils permettent une prise en compte efficace du point de vue de l'utilisateur dans le partitionnement. Cependant, ces algorithmes appartiennent à une catégorie de clustering dite **hard constrained** [DAVIDSON et BASU, 2007]. La stratégie de ces approches est d'imposer à l'algorithme de clustering de **respecter toutes les contraintes**. Cela rend ces propositions très sensibles aux bruits et pose des problèmes de faisabilité en présence de contraintes incohérentes. Des expérimentations menées par DAVIDSON, WAGSTAFF et BASU (2006) montrent même une dégradation des résultats dans certains cas.

Paramètres : Données $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$
Nombre des clusters : k
Ensemble d'instances étiquetées : $\mathcal{S} = \cup_{l=1}^{k} \mathcal{S}_l$, $\mathcal{SSX}$

Résultat : Partitionnement : $\mathbb{C}_l = \{Cl_1, ..., Cl_k\}$

(1). *Initialisation des clusters* $\mathbb{C}_l = Cl_1, ..., Cl_k$;
pour $Cl_l \in \mathbb{C}_l$, **faire**

$$\mu_{Cl_j} = \frac{1}{\mid \mathcal{S}_l \mid} \sum_{x_i \in \mathcal{S}_l} x_i \tag{5.7}$$

fin
(2). *Affectation des instances*;
pour chaque $x_i \in \mathcal{X}$ **faire**
  si $x_i \in \mathcal{S}_j$ **alors**
    Affecter x_i à Cl_j;
  sinon
    Affecter x_i *à* $Cl_j \in \mathbb{C}_l$ ***Tel que*** ;

$$d(x_i, \mu_{Cl_j}) = argmin \sum_{j=1}^{k} \|x_i - \mu_{Cl_j}\|^2 \tag{5.8}$$

  fin
fin
(3). *Mise à jour des centroïdes*;
pour chaque $\mu_{Cl_j} \in \{\mu_{Cl_j}\}_{j=1}^{k}$ **faire**
  Calculer les centroïdes des clusters $\mathbb{C}_l$;

$$\mu_{Cl_j} = \frac{1}{\mid Cl_j \mid} \sum_{x_i \in Cl_j} x_i \tag{5.9}$$

fin
(4). *Répéter (2) et (3) jusqu'à convergence* ;

Algorithme 2 : Constrained-KMeans

Contraintes par modification de la fonction objective

D'autres travaux ont proposé des algorithmes tolérant la violation de quelques contraintes. Ces approches appartiennent à la famille de clustering dite **soft constrained**. L'idée est de chercher à trouver **le meilleur partitionnement tout en respectant le maximum de contraintes**. L'objectif est de faire en sorte que le clustering satisfasse le plus de contraintes possibles tout en permettant le non-respect de celles-ci quand elles semblent illogiques (trop coûteuses). Ces approches peuvent être envisagées quand il y a du bruit dans les contraintes ou des erreurs dans l'échantillon étiqueté.

La plupart de ces propositions reposent sur une modification de la fonction objective des algorithmes. Les contraintes sont incorporées en ajoutant un terme de pénalité à la fonction objective. *PCKMeans* [BASU, BANERJEE et MOONEY, 2004] est une variante de K-Means qui utilise cette technique.

Les auteurs ont proposé d'intégrer les contraintes via deux termes pondérés qui viennent influencer la distance calculée, ce qui donne la fonction objective suivante :

$$R_{pckm} = \frac{1}{2}\sum_{x_i \in \chi} \|x_i - \mu_{li}\|^2 + \sum_{(x_i,x_j)\in \mathcal{ML}} w_{ij} 1[l_i \neq l_j] + \sum_{(x_i,x_j)\in \mathcal{CL}} \overline{w}_{ij} 1[l_i = l_j] \tag{5.10}$$

Les deux termes introduits vont artificiellement modifier la distance, soit en la réduisant quand il s'agira d'une contrainte *must-link*, soit en l'agrandissant quand il s'agira d'une contrainte *cannot-link*. Dans PCKMeans, les contraintes sont aussi utilisées dans la phase d'initialisation pour former des composantes connectées, appelées ensembles de voisinage.

À la différence de la technique d'introduction de contraintes sur phase d'initialisation, l'affectation de ces ensembles de voisinage peut être modifiée par la suite. Les ensembles de voisinage sont inférés en calculant les fermetures transitives des contraintes *must-link*, comme cela a été proposé dans WAGSTAFF et al. (2001). λ ensembles $= \{N_p\}_{p=1}^{\lambda}$ sont ainsi produits à partir des contraintes. Avec la condition que chaque paire d'ensembles N_p et $N_p^{'}$ doivent contenir deux instances liées par un *cannot-link*. L'algorithme 3 décrit les étapes de PCKMeans.

D'autres algorithmes appartenant à la même famille ont été proposés dans la littérature, comme HMRF-Kmeans [BASU, BILENKO et MOONEY, 2004], LCVQE [PELLEG et BARAS, 2007], SS-KERNEL-KMEANS [KULIS et al., 2009], SKMS [ANAND et al., 2014] ou encore S3OM [ALLAB et BENABDESLEM, 2011]. D'autres approches hybrides, à base de contraintes et d'apprentissage de métriques ont été développées. MPCK-Means [BILENKO, BASU et MOONEY, 2004] est une variante de K-Means qui en fait partie de cette famille d'algorithmes.

5.5.3 Discussions des travaux

Le domaine du clustering par contraintes a reçu un fort intérêt ces dernières années. Cependant, peu de travaux se sont intéressés à la génération

Paramètres : Données $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$
Nombre des clusters : k
Contraintes *must-link* : $\mathcal{ML} = \{ml(x_i, x_j)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$
Contraintes *cannot-link* : $\mathcal{CL} = \{cl(x_i, x_j)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$

Résultat : Partitionnement : $\mathbb{C}_l = \{Cl_1, ..., Cl_k\}$

(1). Initialisation des clusters $\mathbb{C}_l = Cl_1, ..., Cl_k$*;*
Créer λ *ensembles de voisinage :* $\{N_p\}_{p=1}^{\lambda}$ *à partir de* $\mathcal{ML}$ *et* $\mathcal{CL}$*;*
Trier les ensemble N_p *suivant leurs tailles;*
si $\lambda \geq k$ **alors**
    Initialiser $\{\mu_{Cl_j}\}_{j=1}^{k}$ *par la moyenne des instances de* $\{N_p\}_{p=1}^{k}$*;*
sinon
    Initialiser $\{\mu_{Cl_j}\}_{j=1}^{k}$ *par la moyenne des instances de* $\{N_p\}_{p=1}^{\lambda}$ *;*
    si $\exists$ *instance* x_i *liée par cannot-link à tout* $\{N_p\}_{p=1}^{\lambda}$ **alors**
        Initialiser $\mu_{Cl_{\lambda+1}}$ *avec* x_i
    fin
    ;
    Initialiser aléatoirement les centroïdes restants ;
fin
(2). Affectation des instances;
pour chaque $x_i \in \mathcal{X}$ **faire**
    Affecter x_i *à* $Cl_j \in \mathbb{C}_l$ ***Tel que*** *;*

$$d(x_i, \mu_{Cl_j}) = argmin(\sum_{j=1}^{k} \|x_i - \mu_{Cl_j}\|^2 + \sum_{(x_i,x_j)\in\mathcal{ML}} w_{ij} 1[l_i \neq l_j] + \sum_{(x_i,x_j)\in\mathcal{CL}} \overline{w}_{ij} 1[l_i = l_j]) \tag{5.11}$$

fin
(3). Mise à jour des centroïdes;
pour chaque $\mu_{Cl_j} \in \{\mu_{Cl_j}\}_{j=1}^{k}$ **faire**
    Calculer les centroïdes des clusters $\mathbb{C}_l$*;*

$$\mu_{Cl_j} = \frac{1}{\mid Cl_j \mid} \sum_{x_i \in Cl_j} x_i \tag{5.12}$$

fin
(4). Répéter (2) et (3) jusqu'à convergence ;

Algorithme 3 : Algorithme PCKMEANS

automatique des contraintes. Les méthodes proposées utilisent toujours des contraintes obtenues manuellement, que ce soit directement sous formes de *must-links* et *cannot-links* de l'expert, ou à partir de données étiquetées disponibles. De plus, l'information sémantique apportée par la classe d'appartenance des instances liées ou préalablement étiquetées n'est pas exploitée.

5.6 Utilisation mutuelle des connaissances formalisées et du clustering

Les travaux qui proposent une utilisation mutuelle des connaissances formalisées en ontologie et du clustering sont relativement rares. En parcourant la littérature, nous avons constaté que la plupart des approches proposées concernent l'analyse et l'interprétation du texte [HOTHO, MAEDCHE et STAAB, 2002 ; HOTHO, STAAB et STUMME, 2003 ; SEDDING et KAZAKOV, 2004 ; BLOEHDORN et al., 2005 ; JING et al., 2006 ; HU et al., 2009].

JING et al. (2006) ont développé une approche qui prend en compte l'ontologie dans le calcul de la distance dans le cadre du clustering K-Means. Les corrélations entre les termes spécifiés dans WordNet [MILLER, 1995] sont calculées et ajoutées comme information supplémentaire aux vecteurs de données initiaux. Deux variantes de K-Means sont par la suite utilisées pour partitionner les données. Les auteurs ont montré au travers les expérimentations que la prise en compte des connaissances de l'ontologie WordNet améliore les résultats du clustering.

HOTHO, STAAB et STUMME (2003) et HOTHO, MAEDCHE et STAAB (2002) ont quant à eux proposé un cadre général qui tire partie de l'ontologie et du clustering à différentes étapes du processus d'extraction des connaissances. Dans un premier temps, l'ontologie est utilisée dans la phase de pré-traitement afin d'obtenir différentes représentations du texte. Plusieurs K-Means sont par la suite appliqués sur ces représentations et les concepts de l'ontologie les plus pertinents sont utilisés pour expliquer les différents résultats du clustering.

D'autres travaux [SEDDING et KAZAKOV, 2004], assez similaires, ont aussi utilisé WordNet pour enrichir le texte à analyser. Les synonymes et les homonymes de chaque mot du texte analysé sont extraits de WordNet et ajoutés à une représentation du corpus dite *bag-of-words* du document. Un clustering est appliqué par la suite aux documents afin de les catégoriser.

Les travaux de HOTHO, STAAB et STUMME (2003), SEDDING et KAZAKOV (2004) et JING et al. (2006), présentés plus haut, ont exploité WordNet comme support des connaissances. Cela rend les approches proposées sensibles à la richesse de WordNet. En effet, la qualité des résultats dépend fortement de la correspondance du texte à analyser avec les sujets couverts par WordNet.

Afin d'apporter une solution à ce verrou, HU et al. (2009) ont proposé d'intégrer Wikipedia comme connaissance externe dans leur approche. La première étape consiste à attacher les documents aux catégories et concepts

de Wikipedia. Une fois le mapping établi, HU et al. (2009) partitionnent les documents en se basant sur une métrique de similarité construite en combinant le contenu des documents, de celui des définitions des concepts et des informations concernant les catégories.

Les approches présentées font un usage très intéressant de l'ontologie et du clustering pour l'amélioration de l'extraction des connaissances. Cependant, ces approches sont spécifiques au traitement du texte et ne sont pas adaptées aux données quantitatives.

En effet, aucune des approches [HOTHO, MAEDCHE et STAAB, 2002; HOTHO, STAAB et STUMME, 2003; SEDDING et KAZAKOV, 2004; BLOEHDORN et al., 2005; JING et al., 2006; HU et al., 2009] ne permet de réduire le fossé sémantique.

De plus, l'utilisation principale de l'ontologie reste l'enrichissement des données par de nouvelles descriptions ou le remplacement des termes corrélés par les concepts qui leur correspondent. L'ontologie n'influence pas directement le partitionnement des données dans ces approches et un algorithme de clustering simple est appliqué dans la plupart des cas.

5.7 Résumé

Dans ce chapitre, nous avons parcouru les travaux de recherche portant à la fois sur l'utilisation des ontologies pour l'interprétation des données quantitatives et la réduction du fossé sémantique, et ceux qui ont exploré l'introduction des connaissances en apprentissage. Nous avons, à chaque fois, essayé d'expliquer les moyens mis en œuvre par les différentes approches, les apports, les limites et leurs différents objectifs.

Chapitre 6

Classification à base d'ontologie et de clustering semi-supervisé

Dans ce chapitre, nous détaillons la première de nos deux propositions ayant pour objectifs communs d'intégrer des modes de représentation quantitatifs et qualitatifs.

L'objectif général de la proposition est de disposer d'une approche hybride à même de tirer parti de connaissances expertes, par essence qualitatives, directement au sein d'activités de clustering.

Le clustering est connu pour donner de bons résultats en matière de définition de groupes homogènes, qui ne sont toutefois pas toujours le reflet des catégories qu'aurait aimé retrouver l'expert. L'idée est donc d'incorporer la vision de l'expert dans le clustering afin d'en guider le partitionnement et d'en améliorer si possible les résultats. De la même manière, les résultats du raisonnement portant sur les connaissances expertes mobilisées peuvent en retour être impactés par le clustering, contribuant ainsi à faire émerger de nouvelles connaissances.

Sommaire

"The true sign of intelligence is not knowledge but imagination", **Albert Einstein**

6.1 Introduction et motivations

Le chapitre 5 a présenté les principales approches retrouvées dans la littérature avec pour objectif d'exploiter des connaissances pour classifier et/ou interpréter les objets d'intérêt d'une image.

Nous avons ainsi vu que deux paradigmes se dégagent naturellement dès lors qu'il s'agit de classifier un ensemble de données non étiquetées en présence de connaissances du domaine.

Le premier paradigme s'appuie sur **une modélisation formelle des connaissances** pour pouvoir ensuite appliquer à ces connaissances des mécanismes de raisonnement, et en particulier de déduction.

Cette manière de voir les choses s'avère très efficace quand les connaissances expertes sont complètes et permettent d'attribuer une sémantique précise à chaque objet de l'image.

Pour ce faire, il est nécessaire non seulement de faire un choix sur la manière de modéliser le domaine en en décrivant dans un premier temps les concepts les plus généraux, mais aussi en faisant en sorte de définir les concepts les plus aboutis à partir de ces concepts généraux, et ce de manière exhaustive.

Il est cependant difficilement possible de disposer dans une ontologie, d'une couverture complète des concepts raffinés permettant de fournir les conditions nécessaires et suffisantes pour que tout individu (ici un pixel ou un segment de l'image) appartienne en extension à un concept en particulier.

En effet, l'acquisition et la formalisation des connaissances sont connues pour être des tâches particulièrement complexes dans le contexte d'applications réelles. En dépit des progrès réalisés dans la mise à disposition de méthodes et d'outils en ingénierie des connaissances, ces tâches nécessitent de lourds investissements en temps et en expertise.

Il arrive donc fréquemment qu'une ontologie ne fournisse des concepts dits définis [1] que pour une portion du domaine considéré, en fonction de l'état des connaissances sur le domaine ou bien des besoins des experts modélisateurs.

Ainsi, pour un domaine donné, les efforts de modélisation et de formalisation aboutiront la plupart du temps à une ontologie où, certains concepts dit primitifs [2] ne seront pas raffinés par des concepts plus précis. Et où, les concepts définis n'engloberont pas tous les cas de figure de l'apparition de leurs instances.

Prenons l'exemple du domaine de la télédétection, les experts auront plus de facilité à décrire des concepts, comme l'eau et la végétation, que

1. Un concept défini possède une à plusieurs conditions dites nécessaires et suffisantes
2. Un concept primitif possède une à plusieurs conditions dites nécessaires, mais qui ne suffisent cependant pas à faire appartenir les individus satisfaisant ces conditions au concept considéré

d'autres concepts plus complexes (et hétérogènes) à l'exemple des bâtiments. De plus, une image satellite est la transcription de valeurs numériques qui sont sujettes à variation en fonction des caractéristiques du capteur ayant servi à l'acquisition, des conditions d'acquisition elles mêmes ou bien des scènes étudiées. Des processus de calibration permettent de normaliser les images pour un capteur exploité et donc de s'affranchir d'une partie de ces variations. Cependant, la définition d'un concept comme l'eau reste dépendante de ces différentes considérations.

Ces difficultés n'enlèvent en rien l'intérêt de l'utilisation d'une ontologie en tant que support de connaissances formalisées. Les avantages sont en effet multiples (section 5.4).

Une ontologie apporte de la modularité en séparant les traitements des connaissances. Les langages de formalisation offrent des prédicats de modélisation qui permettent de raccrocher des concepts de haut niveau à des descripteurs de bas niveau, tout en contextualisant les connaissances, afin de réduire le fossé sémantique. Le raisonnement permet d'étiqueter les instances sans l'utilisation d'exemples, ce qui lui procure un avantage considérable sur les approches à base d'apprentissage.

Enfin, l'adoption de l'hypothèse du monde ouvert permet d'éviter les erreurs d'étiquetage. Le monde ouvert intègre parfaitement la notion d'incomplétude des connaissances. Les informations absentes ne sont non pas considérées comme fausses mais comme inconnues.

Le deuxième paradigme fait une toute autre projection sur les connaissances. L'objectif du **clustering semi-supervisé** est alors d'ajouter des connaissances *a priori*, aux données traitées soit sous forme d'instances étiquetées, soit sous forme de contraintes, pour améliorer le processus d'apprentissage et guider le clustering dans l'exploration des solutions. Ce modèle s'appuie essentiellement sur des mécanismes inductifs, et permet d'obtenir des résultats satisfaisants en particulier pour le traitement de données complexes et hétérogènes à l'exemple d'images.

Cependant, comme nous avons pu le constater, ce processus demande une intervention manuelle et préalable de l'expert pour définir les contraintes, et cela sur chaque jeu de données. Une autre limitation réside dans la perte de la sémantique des classes et la non capitalisation des résultats pour leur exploitation dans l'analyse d'autres jeux de données. Les contraintes sont définies manuellement pour chaque jeu de données et l'étiquette doit être donnée par l'expert en analysant les clusters obtenus.

Pour résumer, les approches semi-supervisées ne sont pas modulaires, les connaissances ne sont pas séparées des algorithmes de classification, ce qui impose le déploiement de nouvelles chaînes de traitement chaque fois que les concepts d'intérêt de l'expert et/ou les jeux de données à analyser changent.

Pour faire face à ces limitations, nous proposons l'utilisation mutuelle de l'ontologie comme référentiel des connaissances et du clustering semi-supervisé pour enrichir la classification, tout en introduisant une interprétation sémantique des clusters. L'idée est d'élaborer une approche hybride qui va s'appuyer sur le raisonnement à base des logiques de description afin

d'automatiser l'interprétation des connaissances du domaine, et utiliser les résultats du raisonnement pour guider le clustering par ces mêmes connaissances en s'appuyant sur des contraintes générées automatiquement.

Les avantages de notre approche hybride sont multiples. L'originalité est d'exploiter un processus déductif pour guider et renforcer un processus inductif, qui vient à son tour enrichir les résultats du processus déductif. Les inconvénients de chaque mécanisme pris séparément sont tour à tour contrebalancés par les apports de l'autre mécanisme.

Notre approche permet ainsi de pallier l'incomplétude et la relative incertitude des connaissances en renforçant l'étiquetage des instances avec le clustering par contraintes. L'utilisation de l'ontologie comme support des connaissances apporte aussi de la modularité à l'approche, puisque la classification s'adapte automatiquement aux concepts de l'ontologie. En accrochant des concepts de haut niveau aux données numériques, notre approche réduit aussi le fossé sémantique.

Un autre avantage de notre méthode réside dans sa capacité à générer automatiquement les contraintes à poser sur le clustering, ce qui permet d'adapter automatiquement, sans intervention manuelle, la classification des données à la vision de l'expert et aussi d'appliquer l'approche sur plusieurs jeux de données avec les mêmes connaissances du domaine.

6.2 Vue globale de l'approche

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.

Notre approche hybride exploite à la fois le raisonnement à base des logiques de description et le clustering afin de tirer le meilleur de ces deux processus et les améliorer mutuellement. 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 comprend une séquence d'étapes :

1. Conceptualisation et formalisation des connaissances expertes ;
2. Projection des données dans 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 par l'ontologie ;
5. Clustering guidé par les contraintes générées.
6. Capitalisation des résultats et propagation de l'étiquetage sémantique.

Comme le montre la figure 6.1, les étapes s'enchaînent pour garantir une collaboration efficace et une capitalisation des résultats obtenus à la fois par le raisonnement et par le clustering. Nous rappelons que nous n'utilisons pas de données étiquetées dans notre méthode. Nous détaillons dans ce qui suit chacune des étapes de la démarche, tout en illustrant les résultats intermédiaires obtenus à partir d'un exemple simple.

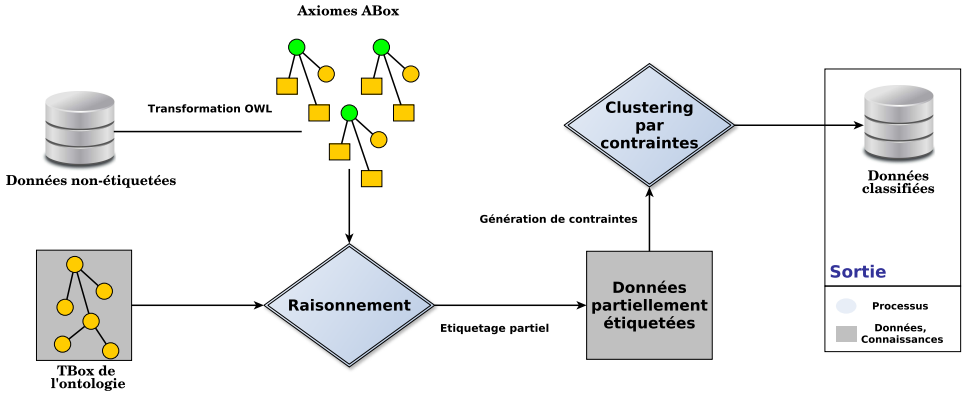


FIGURE 6.1 – Schéma global de l'approche hybride proposée

6.2.1 Conceptualisation et formalisation des connaissances expertes

Nous pouvons distinguer dans notre figure une double entrée. La première concerne les données non-étiquetées, la deuxième porte sur les connaissances expertes.

Notre proposition est une démarche méthodologique valable dans toute problématique qui dispose de données matricielles, et de connaissances formalisées (ou pouvant être formalisées) associées à ces données.

La connaissance que nous considérons est représentée au travers du langage OWL [GROUP et OTHERS, 2009]. Nous rappelons qu'OWL (chapitre 2, section 2.3.4) s'appuie sur la sémantique formelle et précise des logiques de description [KRÖTZSCH, SIMANCIK et HORROCKS, 2012]. Cette formalisation permet aux raisonneurs, en calculant les conséquences logiques, de déduire des nouvelles connaissances qui jusqu'alors étaient implicites. Les raisonneurs proposent différents services d'inférence relevant du raisonnement déductif, au regard des types de relations recherchées, à l'exemple du service d'inférence *réalisation* qui permet de retrouver les concepts les plus précis d'appartenance des instances.

Nous nous appuyons dans notre approche sur l'ontologie pour réduire le fossé sémantique (chapitre 5, section 5.2) et attacher les concepts de haut niveau aux propriétés de bas niveau que contiennent les données numériques. L'acquisition des connaissances n'est pas l'objectif de notre proposition. Nous supposons dans notre approche que des connaissances formalisées du domaine sont disponibles et qu'elles contiennent des concepts définis au travers des descripteurs de bas niveau. Cependant, afin de bien poser

le contexte et montrer la validité de notre proposition, nous détaillons dans cette section, ainsi que dans la section des expérimentations, des éléments sur la méthode de modélisation utilisée.

Nous exploitons dans notre approche les domaines concrets [LUTZ, 2003] pour réduire le fossé sémantique. Nous reprenons également à notre compte une conceptualisation à trois niveaux d'abstraction inspirée des travaux de FALOMIR et al. (2011) et ANDRÉS (2013).

L'ontologie va ainsi être décomposée en trois parties :

— Conceptualisation de référence
— Connaissances contextuelles
— Niveau assertionnel

Les deux premiers niveaux forment la TBox $\mathcal{T}$ de l'ontologie $\mathcal{O}$, ils sont le fruit d'une modélisation préalable des connaissances expertes et représentent la deuxième entrée fournie à notre approche (figure 6.1).

Le niveau le plus bas correspond à l'ABox $\mathcal{A}$ de l'ontologie $\mathcal{O}$, et contient les axiomes représentant les instances décrites à l'aide des descripteurs de bas niveau. Dans notre cas, cette description est obtenue par une transformation semi-automatique des données matricielles d'entrée, introduite dans la section 6.2.2.

Conceptualisation de référence

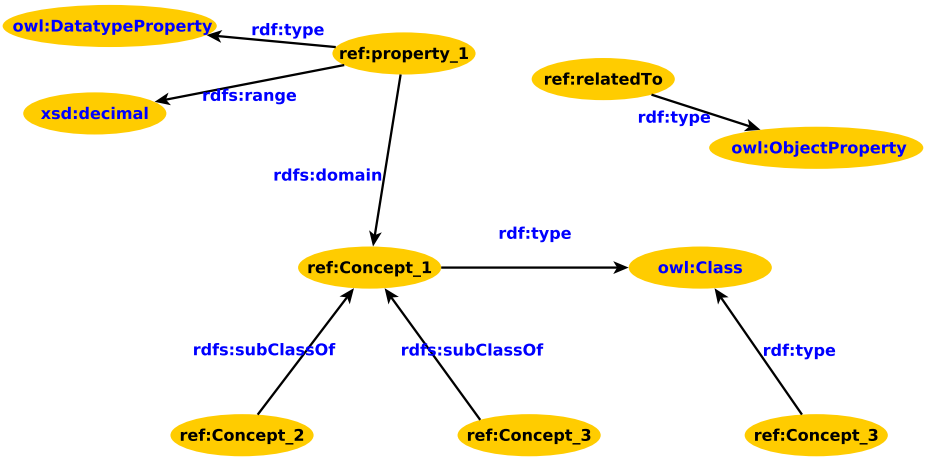
La conceptualisation de référence définit des concepts et des propriétés qui restent valides pour tout contexte d'utilisation. Les concepts de référence sont des concepts de haut niveau largement partagés au sein d'une communauté.

Cette conceptualisation est envisagée comme le cadre général d'un domaine. Elle définit les entités de base bénéficiant d'un large consensus, sans rentrer dans des détails opérationnels. En Télédétection par exemple, cette conceptualisation peut concerner des concepts comme *Image*, *Pixel* ou encore *Segment*, comme nous le verrons lors de la mise en œuvre de notre approche.

La figure 6.2 donne un exemple d'une possible conceptualisation de référence. Elle montre ici la déclaration de 4 concepts de référence et 2 propriétés. Les entités introduites sont formalisées à l'aide des langages OWL, RDFS et RDF (en bleu) et référencées à l'aide de l'espace de noms `ref : http://www.hchahdi.net/ConceptReference\#` spécifique.

Connaissances contextuelles

Les connaissances contextuelles représentent le niveau d'abstraction intermédiaire. Cette composante définit les concepts du domaine dans un cadre plus spécifique. La plupart du temps, les concepts des connaissances contextuelles spécialisent les concepts de la conceptualisation de référence. Ils représentent une vision thématique et sont définis d'une manière plus subjective et proche des besoins applicatifs.



Préfixes:

rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

owl : http://www.w3.org/2002/07/owl#

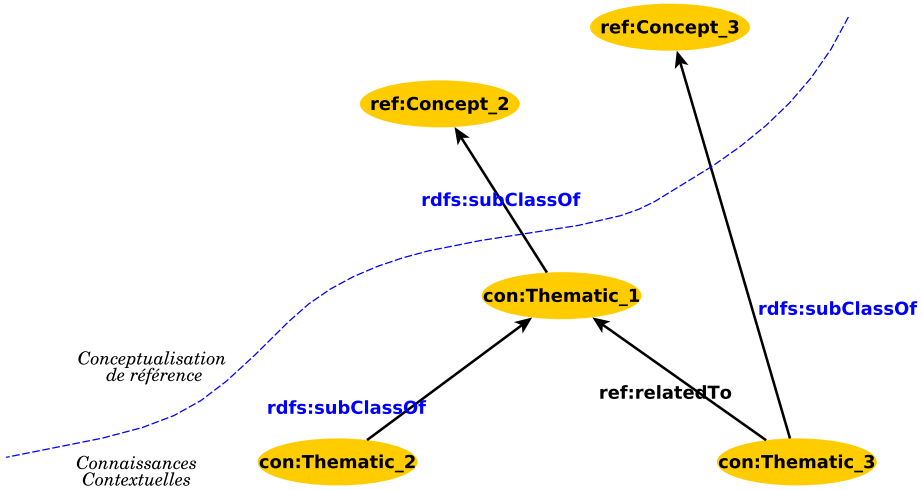
rdfs : http://www.w3.org/2000/01/rdf-schema#

H ref : http://www.hchahdi.net/ConceptReference#

FIGURE 6.2 – Exemple d'une conceptualisation de référence imaginaire

Nous nous intéressons dans notre approche aux domaines traitant de données quantitatives. L'interprétation de ces données nécessite la résolution du problème du fossé sémantique. C'est dans ce niveau d'abstraction que les descripteurs de bas niveau sont liés et combinés aux descripteurs de haut niveau. Les domaines concrets [LUTZ, 2003] jouent un rôle important dans cette liaison, ils permettent de spécifier dans les expressions logiques des concepts thématiques des valeurs et des conditions numériques.

OWL adopte les types de données XML comme supports aux domaines concrets. Ce qui offre beaucoup de possibilités pour la représentation de valeurs quantitatives de plusieurs types (entiers, réels, décimaux...).



Préfixes:

rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

owl : http://www.w3.org/2002/07/owl#

rdfs : http://www.w3.org/2000/01/rdf-schema#

ref : http://www.hchahdi.net/ConceptReference#

con : http://www.hchahdi.net/ConContextuelle#

FIGURE 6.3 – Exemple de connaissances contextuelles

La figure 6.3 représente un exemple de connaissances contextuelles. On y voit que les concepts thématiques sont introduits par une spécialisation des concepts de référence. La réduction du fossé sémantique se fait en s'appuyant sur des définitions similaires à l'expression suivante :

$$Thematic\_1 \equiv Concept\_2 \wedge \exists property\_1_{\{<2.3\}} \tag{6.1}$$

Avec cet exemple simple, nous illustrons la construction du concept *Thematic_1*. Nous le définissons par équivalence en exploitant le concept de référence *Concept_2* et la propriété *property_1*, ainsi qu'une condition d'inégalité $<$ posée sur cette propriété et à la valeur concrète 2.3. De cette manière, nous introduisons une définition thématique valide dans un cadre applicatif précis.

6.2.2 Projection des données dans l'ABox de l'ontologie

Une fois les connaissances expertes formalisées, la première étape est la projection des instances dans l'ABox de l'ontologie (niveau assertionnel). C'est une étape nécessaire et préalable à l'inférence sémantique des types de données par raisonnement. Cet étape consiste à produire une représentation OWL des données.

Afin de rendre notre approche générique, nous avons développé un processus de transformation semi-automatisé (algorithme 4) qui permet de représenter les données matricielles $x_i \in \mathcal{X}$ en instances OWL $a_i \in \mathcal{A}$.

Ce processus analyse les propriétés présentes dans la TBox de l'ontologie et les variables descriptives des données, et propose à l'utilisateur d'établir les correspondances. Une fois que l'utilisateur a indiqué les correspondances, le processus de transformation projette les données en instances OWL. Le processus décrit chaque instance avec les propriétés adéquates, alimente les valeurs des propriétés par celles présentes dans $\mathcal{X}$ et les typent en s'appuyant sur la TBox.

Paramètres :
Données $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$ décrites par : $V = \{v_j\}_{j=1}^{d}$
TBox : $\mathcal{T} =< \mathcal{N}_\mathcal{C}, \mathcal{N}_\mathcal{P} >$
Résultats :
ABox : $\mathcal{A} = \{a_i\}_{i=1}^{n}$
Méthode :
Pour tout p_k in $\mathcal{N}_\mathcal{P}$ **and** v_i in V **Faire**
Boolean Requête = Correspondance entre p_k et v_i
Si Requête.valide() **Alors**
$map(\mathcal{N}_\mathcal{P}, V)$.ajout($p_k$,$v_i$)
Fin Si
Fin Pour
Pour tout $x_i \in \mathcal{X}$ **Faire**
$a_i := createOWLInstance()$;
Pour tout $p_k \in map(\mathcal{N}_\mathcal{P}, V)$ **Faire**
$a_i.addProperty(p_k)$
$a_i.setPropertyType(p_k, \mathcal{T}.getPropertyType(p_k))$
$a_i.setPropertyValue(p_k, x_i.getValueOf(v_k))$
Fin Pour
return a_i : Axiomes OWL représentants x_i
$\mathcal{A}.add(a_i)$
Fin Pour

Algorithme 4 : Projection semi-automatique des données en instances OWL

Nous illustrons notre processus (algorithme 4) par un court exemple. Dans la figure 6.4, on représente la description produite de j^{eme} dimension de l'instance $x_i \in \mathcal{X}$. Le processus s'appuie sur les connaissances formalisées du domaine (voir extrait de la TBox dans la figure) pour représenter sémantiquement les données. La donnée x_i est représentée par une instance OWL $instance_i$. L'instance est décrite dans cet exemple par la propriété de la TBox $\#prop1$, qui est valuée par la valeur de x_i^j. Cette valeur est elle-même typée par $xsd : decimal$ en se basant sur les informations disponibles dans la TBox. Finalement, l'instance est associée au $\#concept_1$, qui peut correspondre à un concept de la TBox (Pixel ou Segment dans le cadre d'une image satellite par exemple).

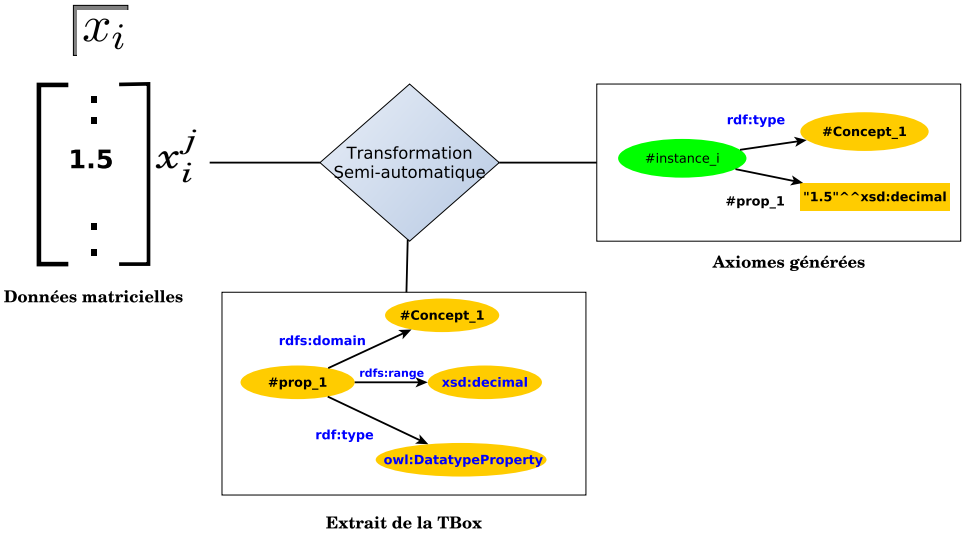


FIGURE 6.4 – Projection d'une donnée x_i en une instance OWL décrite par les propriétés de la TBox

Cette représentation des données en un ensemble d'axiomes sur les individus formant l'ABox de l'ontologie est une première étape dans l'interprétation sémantique. En effet, les données d'entrées non-étiquetées sont dénuées de sens et non contextualisées. Avec cette première étape de description, chaque donnée x_i est représentée par une instance a_i identifiable et dotée d'un ensemble de propriétés formellement définies.

On peut dire qu'avec cette étape, les données sont "habillées" sémantiquement et contextualisées afin de former l'ABox de l'ontologie.

6.2.3 Interprétation sémantique : Inférence du type des instances

L'ABox construite lors de l'étape précédente est alors associée à la TBox et aux connaissances expertes formalisées au préalable. La TBox et la ABox constituent alors la base de connaissances **KB**[3] de l'approche.

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 le service de réalisation encore appelé test à l'instanciation, qui consiste pour le raisonneur à retrouver pour une instance donnée de l'ABox, son concept d'appartenance le plus précis dans la TBox. Ainsi les instances répondant parfaitement aux critères des concepts, sont étiquetées sémantiquement par ces concepts. Comme déjà précisé, le langage OWL adopte l'hypothèse du monde ouvert, qui a comme conséquence le non étiquetage de certaines des instances de la base de connaissance. En effet, les instances qui ne répondent que partiellement à la description de concepts dits définis restent alors non-étiquetées.

3. KB : Knowledge Base

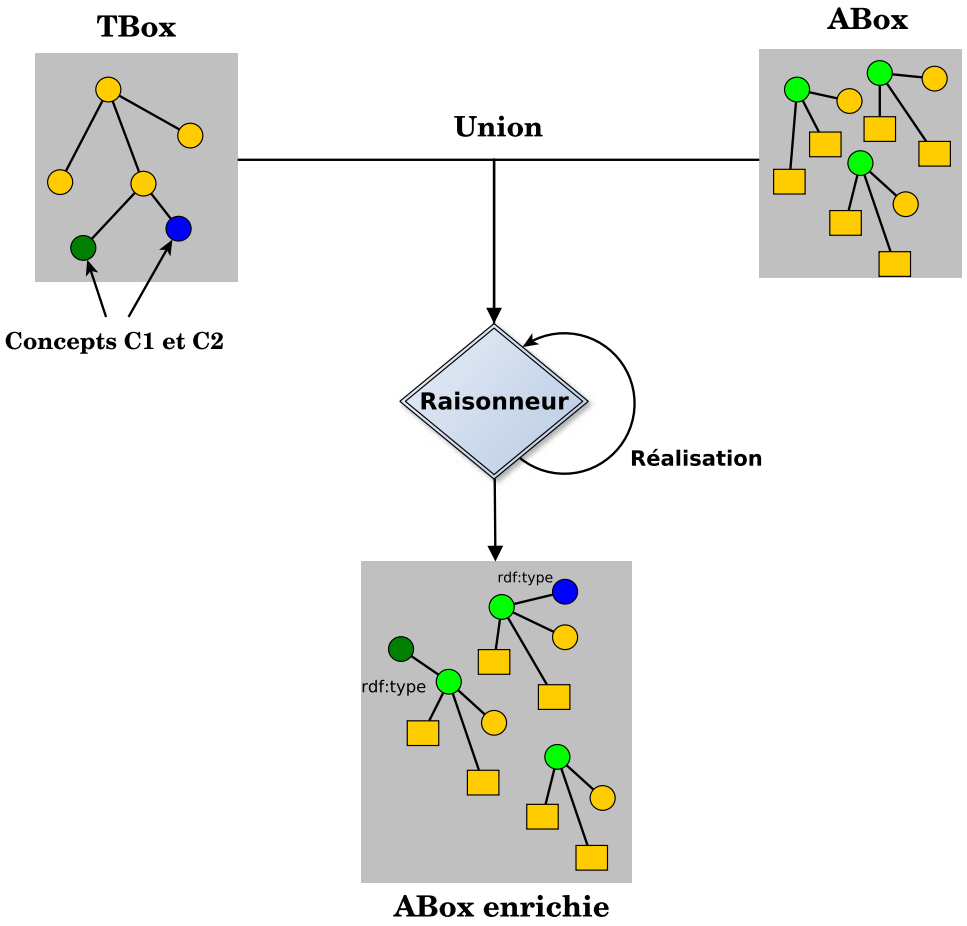


FIGURE 6.5 – Illustration de la Réalisation d'une ABox par raisonnement

La figure 6.5 illustre visuellement la réalisation d'une ABox associée à une TBox pour étiqueter les instances. Sur la figure, l'ABox contient trois instances et la TBox un nombre réduit de concepts dont deux concepts thématiques (C1 et C2). Dans cet exemple, la réalisation permet, après l'union de la TBox et l'ABox, d'étiqueter deux instances parmi les trois présentes dans l'ABox.

Afin de décrire notre approche en mode pas à pas à partir des données matricielles fournies en entrée, nous allons utiliser le même jeu de données à deux dimensions (figure 6.6, (a)) tout au long de cette section. Nous supposons dans notre scénario que nous disposons d'une TBox contenant **deux concepts thématique C1 et C2**. L'objectif est d'opérer de la reconnaissance d'instances sur la base de **quatre classes** thématiques qui intéressent l'expert. Nous illustrons dans l'exemple l'apport de notre approche hybride, combinant les connaissances disponibles (même incomplètes) avec le clustering par contraintes.

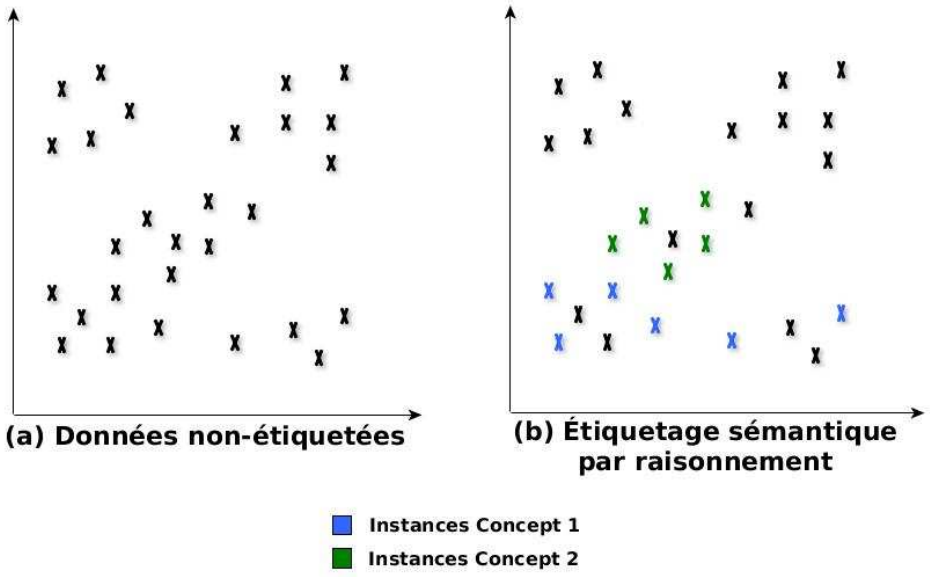


FIGURE 6.6 – Étiquetage sémantique des instances par raisonnement

À la suite des deux premières étapes (transformation en instances OWL et raisonnement), un ensemble d'instances de C1 et de C2 sont identifiées par inférence comme l'illustre la figure 6.6 (b).

6.2.4 Génération automatisée des contraintes à partir des données étiquetées par l'ontologie

Une fois le raisonnement effectué, il est alors temps de procéder à la génération des contraintes. Les instances appartenant au même concept sont ainsi liées entre elles par des contraintes dites *must-link*. Il en va ainsi pour toutes les instances de C1 et de C2 qui sont respectivement unies par des liens *must-link*. À l'opposé, chaque instance de C1 est liée à chaque instance de C2 par un lien *cannot-link* et vice-versa. Cette démarche est généralisée à toutes les instances appartenant à différents concepts (figure 6.8 (c)). Les liens *must-link* et *cannot-link* sont à assimiler à la définition de nouvelles contraintes posées sur les instances et vont dans ce sens, servir par la suite à guider le processus de clustering.

6.2.5 Clustering guidé par contraintes

A ce stade, nous obtenons au travers du raisonnement et du processus de génération des contraintes un ensemble d'instances liées deux à deux par des contraintes et étiquetées sémantiquement. Le jeu d'instances ainsi obtenu, vient alimenter le clustering sous contrainte qui lui est opéré sur la totalité des données disponibles.

Comme déjà introduit, deux variantes se distinguent dans la prise en compte des contraintes. La variante dite *dure ou hard constrained* impose

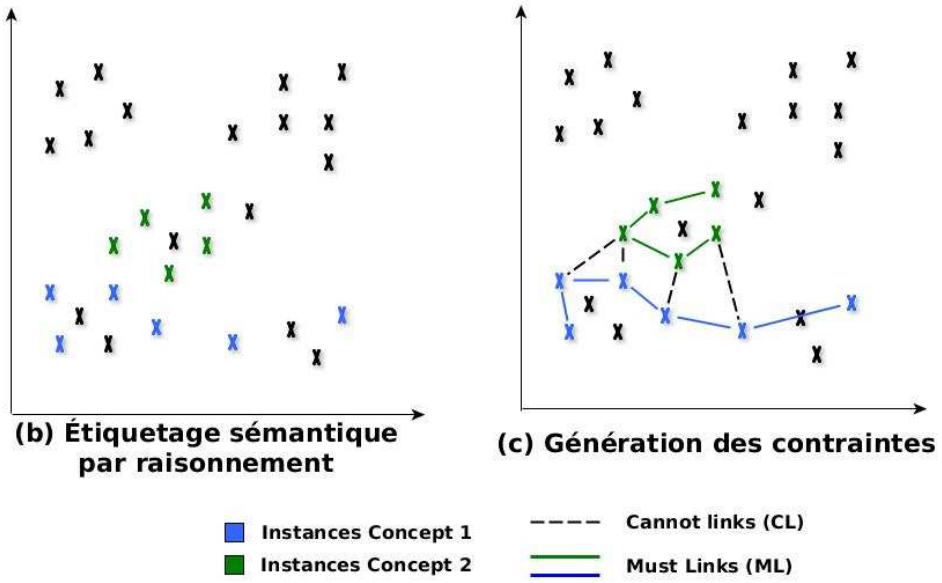


FIGURE 6.7 – Génération des contraintes must-link et cannot-link à partir des résultats du raisonnement

une satisfaction rigoureuse des paires de contraintes définies sur les instances et peut poser des problèmes de faisabilité ; la variante dite *souple ou soft constrained* est moins drastique et tolère certains manquements en cherchant à satisfaire au mieux les contraintes posées. Dans notre contexte d'étude, les contraintes sont produites après reconnaissances d'instances par le raisonneur et le processus proposé est complétement automatisé. Nous avons donc préféré relâcher certaines contraintes afin de faire œuvrer ces contraintes plutôt comme des préférences pour guider le clustering et ainsi toujours pouvoir obtenir un résultat qui maximise la qualité du partitionnement. À cet effet, nous avons fait le choix de l'algorithme PCKMEANS [BASU, BANERJEE et MOONEY, 2004] qui étend l'algorithme des k-moyennes pour tirer parti de paires de contraintes souples entre objets. Notons :

- k le nombre de clusters
- ML l'ensemble des contraintes $ml(x_i, x_j)$ (*must-link*) générées
- CL l'ensemble des contraintes $cl(x_i, x_j)$ (*cannot-link*) générées
- $W = w_{ij}$ et $\overline{W} = \overline{w}_{ij}$ les poids attribués respectivement aux contraintes ML et CL

Le clustering par contraintes PCKMEANS est basé sur la minimisation de la fonction objective suivante :

$$R_{pckm} = \frac{1}{2} \sum_{x_i \in \chi} \|x_i - \mu_{Cl_i}\|^2 + \sum_{(x_i,x_j) \in \mathcal{ML}} w_{ij} 1[Cl_i \neq Cl_j] + \sum_{(x_i,x_j) \in \mathcal{CL}} \overline{w}_{ij} 1[Cl_i = Cl_j] \tag{6.2}$$

Où Cl_i ($Cl_i \in \mathbb{C}_{l_{i=1}}^{k}$) est le cluster d'appartenance de l'instance x_i, et où $w_{ij}1[Cl_i \neq Cl_j]$ et $\overline{w}_{ij}1[Cl_i = Cl_j]$ correspondent respectivement aux coûts de la violation des contraintes $ml(x_i, x_j) \in ML$ et $cl(x_i, x_j) \in CL$. 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_{Cli} ayant comme centroïde μ_{Cli}. La minimisation de cette fonction objective est résolue au travers de l'algorithme suivant [BASU, BANERJEE et MOONEY, 2004] :

Paramètres : Données $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$
Nombre des clusters : k
Contraintes *must-link* : $\mathcal{ML} = \{ml(x_i, x_j)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$
Contraintes *cannot-link* : $\mathcal{CL} = \{cl(x_i, x_j)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$

Résultat : Partitionnement : $\mathbb{C}_l = \{Cl_1, ..., Cl_k\}$

(1). Initialisation des clusters $\mathbb{C}_l = Cl_1, ..., Cl_k$;
Créer λ *ensembles de voisinage :* $\{N_p\}_{p=1}^{\lambda}$ *à partir de* $\mathcal{ML}$ *et* $\mathcal{CL}$;
Trier les ensemble N_p *suivant leurs tailles*;
pour $j \in \{1, .., \lambda\}$ **faire**
  Initialiser $\{\mu_{Cl_j}\}_{j=1}^{\lambda}$ *par la moyenne des instances de* $\{N_p\}_{p=1}^{\lambda}$;
fin
si $\exists$ *instance* x_i *liée par cannot-link à tout* $\{N_p\}_{p=1}^{\lambda}$ **alors**
  Initialiser $\mu_{Cl_{\lambda+1}}$ *avec* x_i;
fin
Initialiser aléatoirement les centroïdes restants ;
(2). Affectation des instances;
pour chaque $x_i \in \mathcal{X}$ **faire**
  Affecter x_i *à* $Cl_j \in \mathbb{C}_l$ ***Tel que*** *;*

$$d(x_i, \mu_{Cl_j}) = argmin(\sum_{j=1}^{k} \|x_i - \mu_{Cl_j}\|^2 + \sum_{(x_i,x_j)\in\mathcal{ML}} w_{ij}1[Cl_i \neq Cl_j] + \sum_{(x_i,x_j)\in\mathcal{CL}} \overline{w}_{ij}1[Cl_i = Cl_j]) \quad (6.3)$$

fin
(3). Mise à jour des centroïdes;
pour chaque $\mu_{Cl_j} \in \{\mu_{Cl_j}\}_{j=1}^{k}$ **faire**
  Calculer les centroïdes des clusters $\mathbb{C}_l$;

$$\mu_{Cl_j} = \frac{1}{\mid Cl_j \mid} \sum_{x_i \in Cl_j} x_i \quad (6.4)$$

fin
(4). Répéter (2) et (3) jusqu'à convergence ;

Algorithme 5 : Adaptation de l'algorithme PCKMEANS à notre approche

6.2.6 Capitalisation des résultats et propagation de l'étiquetage sémantique

Une fois le clustering effectué, nous propageons l'étiquetage sémantique des instances obtenues 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.

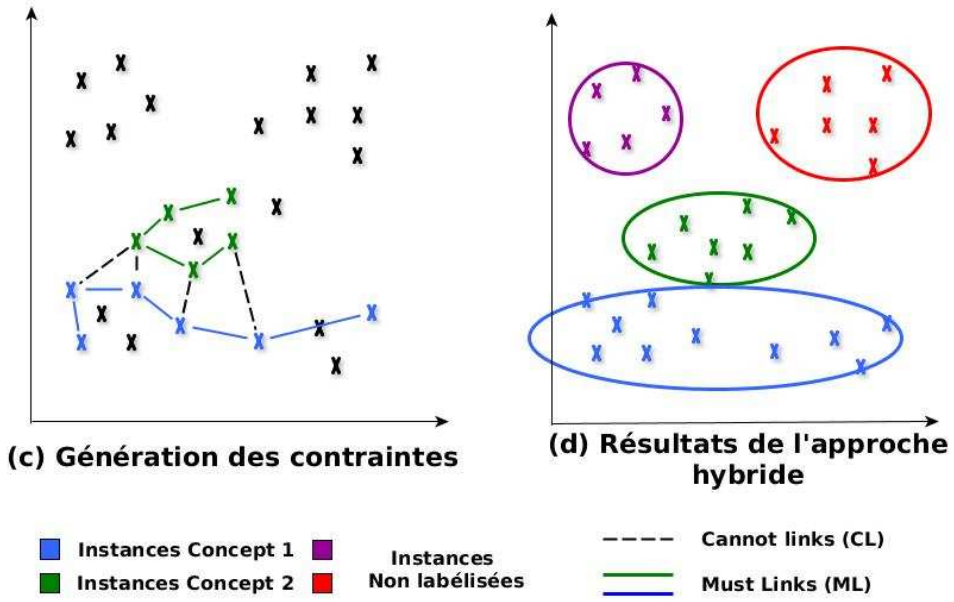


FIGURE 6.8 – Clustering par contraintes avec étiquetage sémantique des clusters par l'ontologie

6.3 Mise en oeuvre

Le principe général de notre proposition, ainsi que l'enchaînement de ses différentes étapes sont maintenant définis. D'un point de vue pratique, plusieurs déclinaisons techniques peuvent être envisagées pour implémenter notre proposition. Cependant, quelques impératifs doivent être respectés pour l'obtention des résultats attendus.

Dans cette section, nous allons présenter plus en détails les réalisations logicielles venant valider l'approche. Les spécificités de la mise en œuvre de ces réalisations dans le cadre de la classification d'images satellites sont également abordées.

6.3.1 Données : Images LANDSAT

Les données que nous exploitons sont des images LANDSAT 5 TM (chapitre 4, section 4.3.3), qui ont une résolution de 30 mètres et sont composées de sept bandes spectrales. Les images LANDSAT sont disponibles en ligne gratuitement et libres d'utilisation. De plus, l'acquisition des images LANDSAT se fait depuis maintenant plus de quarante ans, offrant ainsi de pouvoir disposer de séries temporelles sur le long terme. Les fenêtres radiométriques des bandes spectrales des images LANDSAT facilitent également la détection de différentes classes d'occupation du sol. La facilité d'usage et la perspective de pouvoir analyser les évolutions des paysages dans le temps nous ont encouragé à choisir ce type d'images pour la mise en application de notre approche. Notre proposition est toutefois générique et peut s'appliquer à tout type d'image voire à tout type de donnée, elle est conditionnée uniquement par la disponibilité de connaissances expertes formalisées.

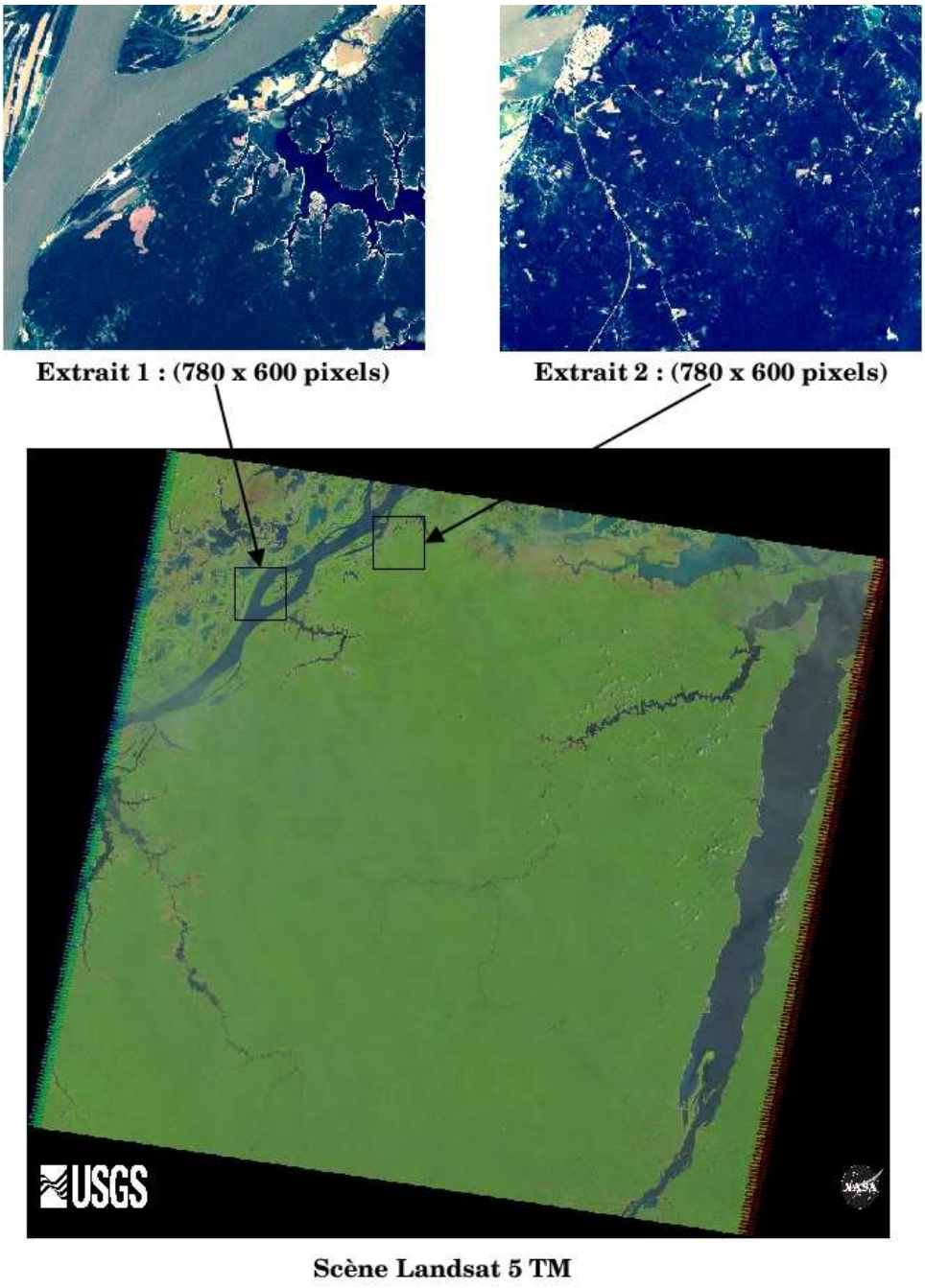


FIGURE 6.9 – Scène LANDSAT 5 TM (WRS 228/062) de la région *Rio Tapajós* acquise le 29/10/2011, avec deux extraits contenant chacun de (780 × 600) pixels
©USGS/NASA Landsat

Nous prenons pour exemple des extraits d'images provenant de deux différentes scènes LANDSAT 5 TM. Une scène concerne la rivière *Rio Tapajós* au Brésil et l'autre le Sud de la France.

La figure 6.9 représente la première des deux scènes. Cette scène peut être téléchargée en ligne sur le portail *EarthExplorer* de la NASA.

FIGURE 6.10 – Scène LANDSAT 5 TM (WRS 196/030) de la région *Occitanie* acquise le 13/10/2011, avec un extrait de (780 × 600) pixels sur la *Grande Motte*
©USGS/NASA Landsat

La deuxième scène porte sur la région *Occitanie* du Sud de la France, elle a été acquise le 13/10/2011 et est représentée ici (figure 6.10) avec un extrait de 780 × 600 sur la station balnéaire de la *Grande Motte*.

L'objectif est de disposer d'extraits d'images qui contiennent à la fois des surfaces en eau et des surfaces végétales avec la présence d'autres types de couverture de sol à l'exemple de sol nu ou de surfaces bâties. Nous avons fait le choix de deux zones d'étude différentes (Brésil et France) pour avoir une première indication sur la généricité de l'approche.

6.3.2 Calibration radiométrique des images satellites

Un des grands avantages de notre approche est de pouvoir s'affranchir d'échantillons de données d'apprentissage qui sont toujours très coûteux à acquérir. Toutefois, des prétraitements sont nécessaires pour calibrer les images satellites avant de pouvoir appliquer l'approche méthodologique.

Ces prétraitements nécessitent également un travail supplémentaire qui n'est pas négligeable et qui n'est pas à négliger pour obtenir des images comparables.

La matrice des vecteurs numériques sous-tendant l'image est un conteneur de valeurs numériques sans véritable sémantique. Ces valeurs représentent les rayonnements de la surface capturées par les instruments à bord du satellite et dépendent de multiples paramètres externes, comme la position du satellite par rapport à la surface, l'angle de vision, les conditions météorologiques (ensoleillement, ...) ou encore les perturbations atmosphériques. Ces facteurs externes peuvent induire de fortes variations dans les valeurs numériques acquises par les capteurs.

Il est donc essentiel de tenir compte de ces paramètres externes qui sont renseignés dans les lots de métadonnées associés aux images de manière à normaliser les grandeurs physiques attachées aux pixels et pouvoir ainsi traiter des séries d'images de manière indifférenciée.

Nous nous intéressons ici à la correction radiométrique, et nous réalisons une calibration sur les images en nous appuyant sur les informations contenues dans les métadonnées. L'objectif principal est la transformation des comptes numériques des pixels en grandeurs physiques, à savoir dans notre cas étalonner les valeurs des pixels pour obtenir leurs mesures de *réflectance au dessus de l'atmosphère*, et ce pour les différentes bandes spectrales de l'image. Les métadonnées fournissent, en particulier, des informations précises sur les caractéristiques des capteurs, les conditions d'acquisition et la position du satellite. Ces métadonnées jouent un rôle fondamental dans le processus de calibration radiométrique des pixels. La calibration en réflectance au dessus de l'atmosphère (**TOA**) s'effectue en deux étapes (figure 6.11). La première étape transforme les valeurs numériques en *luminance*, en appliquant l'équation 6.5 à partir des caractéristiques spécifiées dans les métadonnées. La deuxième étape permet d'obtenir les valeurs de réflectance à partir de celles de la luminance en appliquant l'équation 6.6 à partir cette fois-ci des caractéristiques renseignant les conditions d'acquisition de l'image.

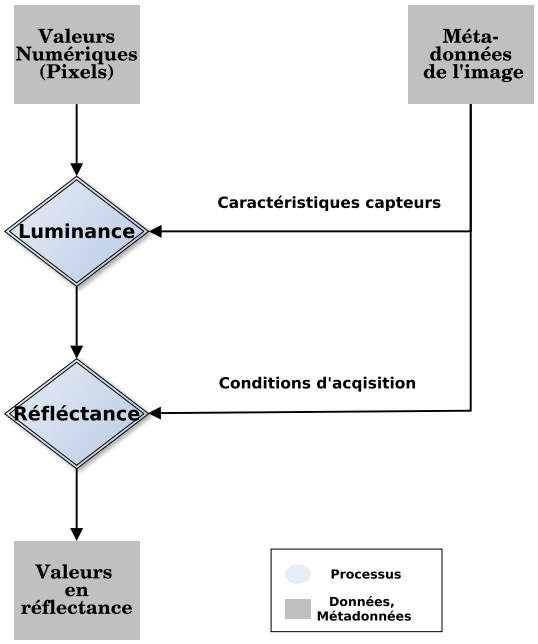


FIGURE 6.11 – Processus de calibration radiométrique en réflectance au dessus de l'atmosphère. Les deux étapes s'appuient sur les métadonnées de l'image pour opérer l'étalonnage des valeurs

La luminance désigne en physique la puissance du flux électromagnétique émis par unité de surface et par unité d'angle solide ($W.m^{-2}.sr^{-1}$). Pour les images d'observation de la Terre, la luminance est calculée en tenant compte des caractéristiques des capteurs du satellites. L'équation 6.5 permet d'obtenir la luminance pour les images LANDSAT TM[CHANDER, MARKHAM et HELDER (2009)]. Les valeurs des variables sont fournies dans le fichier des métadonnées de l'image, ces valeurs sont spécifiques pour chacune des bandes spectrales.

$$L_\lambda = (\frac{LMAX_\lambda - LMIN_\lambda}{\mathcal{Q}_{calmax} - \mathcal{Q}_{calmin}})(\mathcal{Q}_{cal} - \mathcal{Q}_{calmin}) + LMIN_\lambda \qquad (6.5)$$

Avec :

- $\mathcal{Q}_{cal}$: Valeur numérique du pixel dans la bande
- $\mathcal{Q}_{calmin}$: Valeur minimale quantifiée du pixel correspondant à $LMIN_\lambda$
- $\mathcal{Q}_{calmax}$: Valeur maximale quantifiée du pixel correspondant à $LMAX_\lambda$
- $LMIN_\lambda$: Luminance du capteur rapportée à $\mathcal{Q}_{calmin}$
- $LMAX_\lambda$: Luminance du capteur rapportée à $\mathcal{Q}_{calmax}$

L'équation s'appuie sur les caractéristiques des capteurs pour fournir la luminance des pixels de l'image. C'est une étape intermédiaire à l'obtention de la réflectance. En s'appuyant sur la luminance et en prenant compte des conditions d'acquisition comme la distance entre le soleil et la surface et l'angle de vision du satellite par rapport à la Terre. La réflectance des

images LANDSAT 5 TM est obtenue suivant l'équation suivante [CHANDER et MARKHAM, 2003 ; CHANDER, MARKHAM et BARSI, 2007] :

$$\rho_\lambda = \frac{\pi . L_\lambda . d^2}{ESun_\lambda . \cos\theta_S} \tag{6.6}$$

Avec :

- π : Constante mathématique, approximativement égale à $3,14159$
- L_λ : Luminance calculée (équation 6.5)
- d : Distance entre le soleil et la surface de la Terre à la date d'acquisition
- $ESun_\lambda$: Luminance émise par le soleil sur la bande concernée
- θ_S : L'angle zénithal solaire

Le portail EarthExplorer [4] fournit plusieurs types de produits pour la même image. Ainsi les utilisateurs finaux peuvent disposer des images brutes (pixels exprimés en comptes numériques) ou prétraitées en luminance (pixels exprimés en luminance). Suivant le produit choisi, la calibration en réflectance TOA se fait en appliquant successivement la calibration en luminance puis en réflectance ou bien en appliquant uniquement la calibration en réflectance.

Les prétraitements présentés ici sont pour une grande part spécifiques aux images LANDSAT. Il faut cependant garder à l'esprit que les images satellites sont à prétraiter dans tous les cas de figure de manière à pouvoir disposer de produits normalisés et rendre notre méthode reproductible. Les pixels sont maintenant prêts à être analysés et enrichis sémantiquement.

6.3.3 Ontologie du domaine pour les images d'observation de la Terre

L'interprétation d'images satellites présuppose de disposer d'un socle préalable de connaissances expertes du domaine. Nous avons construit à cet effet une ontologie formalisant partiellement des connaissances de la Télédétection et permettant la réduction du fossé sémantique pour deux classes d'occupation du sol en particulier. Le rôle de cette ontologie est de montrer un exemple d'application réelle et significatif de notre approche.

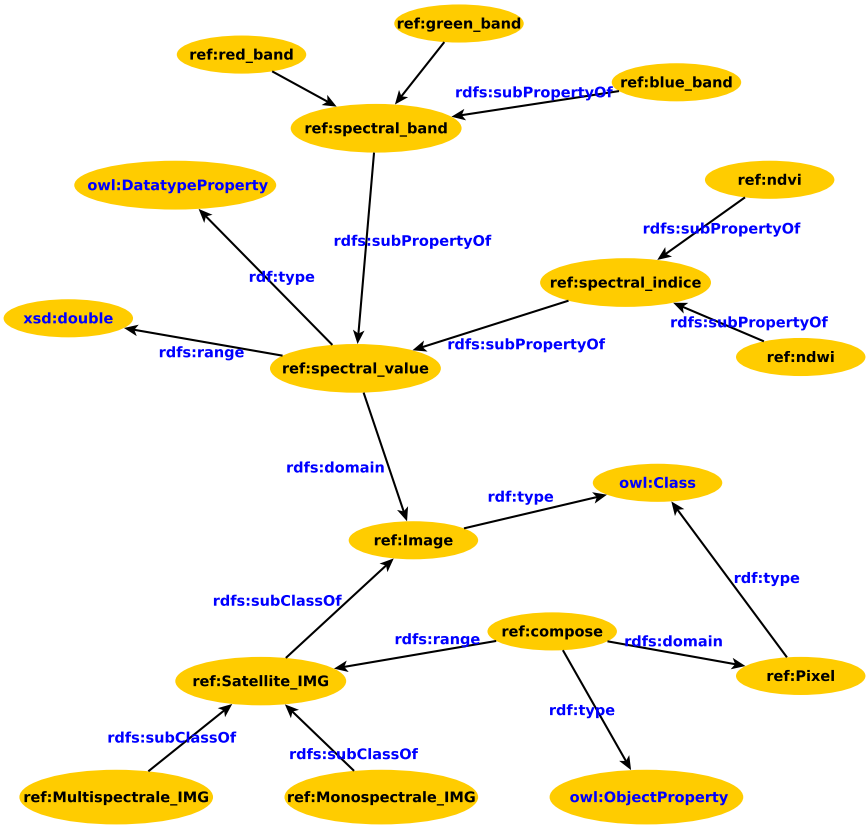
Pour la modélisation de l'ontologie du domaine, nous adoptons les trois niveaux d'abstraction retenus. Dans ce cadre, nous exposons ici les connaissances expertes qui correspondent aux deux niveaux d'abstraction supérieurs, l'ABox étant générée à la volée à partir de l'image.

Conceptualisation de référence d'images pour l'observation de la Terre

Nous avons vu que la conceptualisation de référence regroupe les connaissances du domaine relativement étendues et indépendantes des besoins spécifiques qui peuvent évoluer suivant le contexte applicatif.

En Télédétection, les concepts de référence peuvent concerner la modélisation de la structure de l'image, du satellite, des capteurs ou encore les

4. EarthExplorer : https ://earthexplorer.usgs.gov



Préfixes:

rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

owl : http://www.w3.org/2002/07/owl#

rdfs : http://www.w3.org/2000/01/rdf-schema#

ref : http://www.coclicoknowledge.net/Reference#

FIGURE 6.12 – Extrait de notre conceptualisation de référence du domaine des images de Télédétection

relations entre ces différentes entités. Cette conceptualisation va aussi formaliser les domaines de définition associés aux propriétés des pixels et des images.

Dans notre conceptualisation, *Image Satellite*, *Pixel* et *Segment* sont des concepts OWL. Les relations entre les concepts comme des *ObjectProperty* et les propriétés représentant les bandes et les indices spectraux comme des *DatatypeProperty*. Leurs types et domaines de définition sont quant à eux spécifiés à l'aide du langage RDFS (*rdfs :domain* et *rdfs :range*), ce qui permet de contrôler les valeurs que peuvent prendre ces propriétés.

L'indice de végétation normalisé *NDVI* par exemple est une propriété qui ne peut être calculée que sur une image multispectrale et ce quel que soit le contexte d'application. Nous illustrons un extrait de notre conceptualisation dans la figure 6.12.

Connaissances contextuelles sur les classes d'occupation du sol

Dans les connaissances contextuelles, nous formalisons les concepts du domaine avec une vision thématique. Cette partie va ainsi définir des concepts qui vont à la fois spécialiser les concepts de la conceptualisation de référence et au même temps utiliser les opérateurs logiques et les domaines concrets pour réduire le fossé sémantique.

Les connaissances contextuelles vont ainsi faire le lien entre des concepts de référence, comme *Pixel*, et des concepts thématiques, comme *Pixel_Eau* qui va être défini à l'aide des propriétés de bas niveau et d'intervalles numériques.

Caractérisation des concepts thématiques Les concepts thématiques sont construits à partir des signatures spectrales qui permettent de discriminer les différentes couvertures du sol à partir des bandes spectrales de l'image satellite.

La signature spectrale se révèle en effet un moyen fiable de caractériser les types de surface terrestre. Une signature spectrale désigne la quantité d'énergie émise ou réfléchie en fonction de la longueur d'onde [DÉVELOPPEMENT DURABLE, 2008] et correspond à la courbe de la réflectance sur le spectre électromagnétique.

La figure 6.13, empruntée à MCCLOY et SEVERIENS (2012), montre les signatures de quelques surfaces naturelles dans les spectres visibles, proche et moyen infrarouge. Ces mesures sont prises directement depuis la Terre.

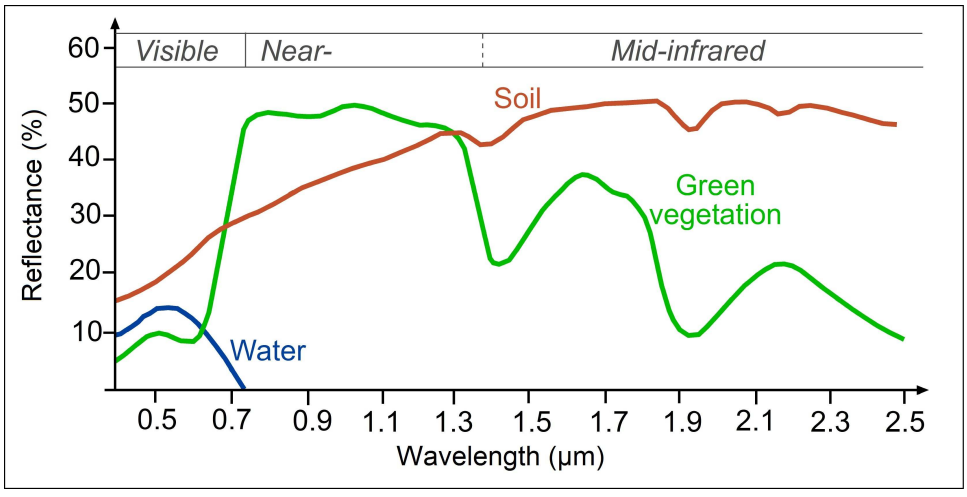


FIGURE 6.13 – Signatures spectrale de l'eau, de la végétation verte et du sol sur différents spectres électromagnétiques

Les images satellites fournissent quant à elles des télémesures, sur des fenêtres réduites du spectre électromagnétique et résumées par une seule valeur (moyenne du signal reçu sur la fenêtre spectrale). En projetant par exemple les fenêtres spectrales des capteurs LANDSAT 5 TM (figure 6.14), nous voyons que l'information potentiellement disponible ne couvre pas

tout le spectre. Différentes perturbations subies par les signaux avant d'arriver aux capteurs du satellite s'ajoutent aussi aux difficultés d'observer les surfaces à distance.

La figure 6.14 montre néanmoins que les fenêtres spectrales fournissent une information pertinente sur la nature des classes d'occupation du sol. Reste à établir de manière formelle les liens entre ces bandes et les concepts d'intérêt. Le calcul d'indices spectraux figurent parmi les moyens utilisés par les experts pour faciliter ce passage vers le qualitatif.

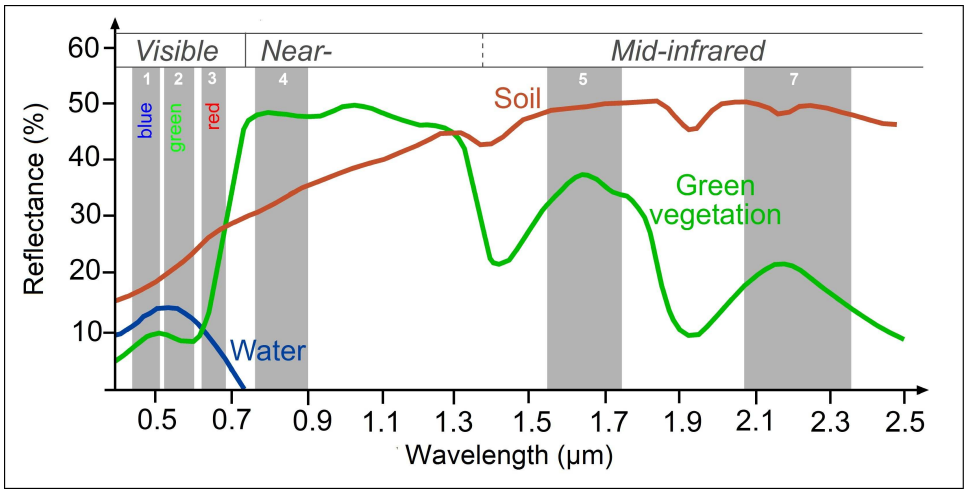


FIGURE 6.14 – Signatures spectrales de l'eau, de la végétation verte et du sol avec la représentation des fenêtres spectrales des bandes spectrales LANDSAT 5 TM

La définition d'indices spectraux en Télédetection a pour mérite de faciliter la discrimination d'une classe d'occupation du sol. Le principe se résume à calculer de nouvelles variables à partir des bandes spectrales, qui vont aider à faire apparaître un ou plusieurs traits caractéristiques dans les données. Le *NDVI* fait parti des indices les plus utilisés.

Le NDVI (Normalized Difference Vegetation Index) est un indice de végétation normalisé calculé à partir des bandes rouge (r) et proche infrarouge (pir). Introduit initialement par ROUSE JR et al. (1974) et exploité par TUCKER (1979) pour la discrimination de la végétation. L'indice met en évidence l'une des caractéristiques spécifiques à la végétation, qui est sa forte absorption dans le rouge et sa forte réflectance dans le proche infrarouge (figure 6.14). Le *NDVI* est un ratio calculé par l'expression suivante :

$$NDVI = \frac{B_{pir} - B_r}{B_{pir} + B_r} \tag{6.7}$$

L'indice *NDVI* peut également être utile à la détection de l'eau. En nous appuyant sur les travaux de spécialistes en Télédétection des classes d'occupation du sol [ZHU et WOODCOCK, 2012; BARALDI et al., 2006], nous

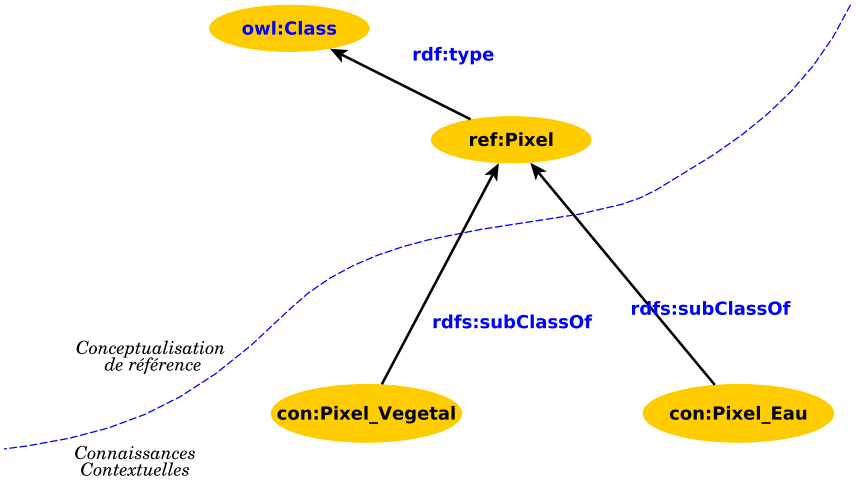
avons défini deux concepts thématiques dans le contexte des images LANDSAT. Il s'agit des concepts *Pixel_Eau* et *Pixel_Végétal* :

$$Pixel\_Eau \equiv Pixel \wedge ((\exists nir\_band_{\{<0.05\}} \wedge \exists ndvi._{\{<0.01\}}) \vee (\exists nir\_band_{\{<0.11\}} \wedge \exists ndvi_{\{<0.001\}})) \tag{6.8}$$

$$Pixel\_Vegetal \equiv Pixel \wedge ((\exists ndvi_{\{>0.3\}}) \vee (\exists rvi_{\{>2.5\}})) \tag{6.9}$$

Pour identifier au mieux les pixels *végétaux*, BARALDI et al. (2006) expliquent que la corrélation entre la bande rouge et proche infrarouge peut aussi être exploitée en complément du *NDVI*. Ce rapport est quantifié à l'aide de l'indice de végétation par quotient *RVI*, introduit par JORDAN (1969) et défini selon l'expression :

$$RVI = \frac{B_{pir}}{B_r} \tag{6.10}$$



Préfixes:

rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

owl : http://www.w3.org/2002/07/owl#

rdfs : http://www.w3.org/2000/01/rdf-schema#

ref : http://www.coclicoknowledge.net/Reference#

con: http://www.coclicoknowledge.net/Contextual#

FIGURE 6.15 – Extrait des concepts définis par spécialisation dans nos connaissances contextuelles

Les concepts thématiques qui apparaissent dans la figure 6.15, représentent une vue partielle des connaissances contextuelles. La représentation des primitives exprimées par les expressions 6.8 et 6.9 dans des graphes RDF donne des représentations par trop difficiles à lire visuellement.

Comme précisé au début de la section, le langage adopté pour l'opérationnalisation de l'ontologie est OWL. Toutes les connaissances expertes

(de référence et contextuelles) sont donc exprimées dans ce langage.

6.4 Expérimentations

Nous avons décrit les connaissances expertes formalisées, les données exploitées ainsi que les prétraitements appliqués. Cette section décrit le protocole expérimental adopté ainsi que les résultats obtenus dans l'analyse et l'interprétation des images satellites.

6.4.1 Protocole expérimental

Chacune des images utilisées dans nos expérimentations est composée de sept bandes spectrales et contient 468.000 pixels. Aucun échantillon de pixels préalablement 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 entre autres la formalisation de deux concepts thématiques : l'eau 6.8 et la végétation 6.9.

Plusieurs frameworks ont été utilisés afin d'implémenter notre plateforme de test. Un processus dédié pour les pré-traitements des images satellites, leurs transformations et le calcul des indices spectraux a été mis en place à partir de la librairie *Orfeo ToolBox* [5].

Les extraits des images brutes sont ainsi calibrés et les indices spectraux *NDVI* et *RVI* sont également calculés. En sortie, chaque jeu de de données $\mathcal{X}$ est composé de 468.000 instances $x_i \in \mathbb{R}^9$. Chaque instance (pixel) est décrite par 9 variables, 6 variables correspondent aux mesures de réflectance sur les principales bandes spectrales de l'image, une septième variable donne la température et les deux dernières variables correspondent aux deux indices spectraux calculés.

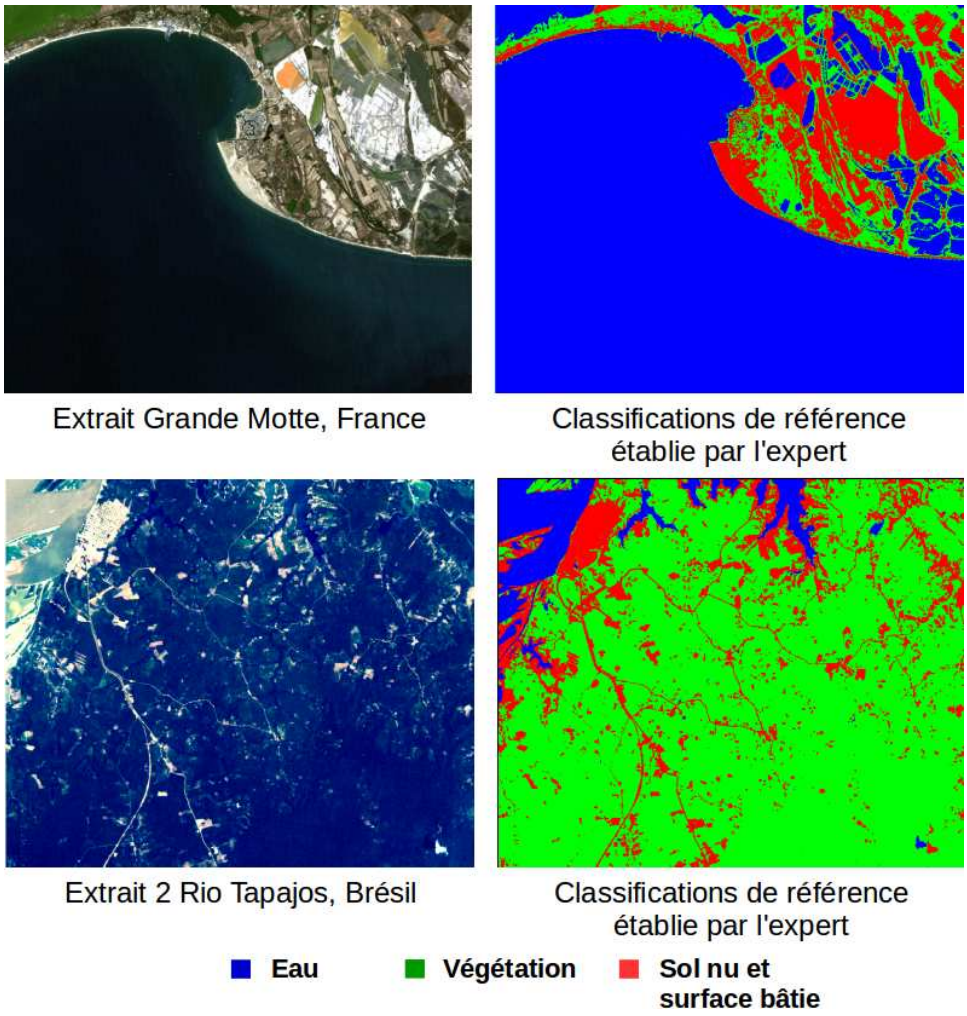
Pour ce qui concerne la base de connaissances, la projection des données en instances OWL est implémentée en Java en s'appuyant sur la librairie *OWL API* et l'algorithme 4 que nous avons proposé. Pellet [SIRIN et al., 2007] est le raisonneur OWL utilisé afin de matérialiser le type déduit des pixels à partir des définitions des concepts de l'ontologie (réalisation). Il a été choisi pour son support complet du raisonnement sur les ensembles de types prédéfinis du langage XML Schema (*datatypes xsd*).

La génération des contraintes ainsi que l'algorithme de clustering PCK-MEANS sont aussi implémentés en Java.

6.4.2 Classifications de référence

Pour évaluer nos résultats, nous nous appuyons sur des classifications de référence que nous avons établi manuellement. Nous disposons de quatre jeu de données étiquetés. Les trois premiers extraits portent sur la rivière de *Rio Tapajós* en Amazonie. Le dernier extrait concerne quant à lui la ville de la *Grande Motte* au sud de la France.

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



Pour obtenir ces classifications de référence, nous avons effectué un étiquetage manuel (photo-interprétation) d'environ 80% des pixels pour chaque ensemble de données à l'aide du logiciel d'analyse d'images ENVI [6]. Une classification supervisée SVM est appliquée par la suite pour obtenir l'étiquetage des pixels restants et la qualité de la classification est ensuite vérifiée.

Ces classifications de référence sont basées sur trois classes identifiées par l'expert. *La végétation*, *l'eau* et une troisième classe regroupant *sol nu et surface bâtie*. La figure représente visuellement deux extraits parmi les quatre classifications de référence disponibles.

6.4.3 Résultats

L'objectif ici est de présenter les résultats de mise en œuvre de notre approche et de mettre en évidence les avantages de l'exploitation simultanée

6. Envi : http ://www.exelisvis.co.uk/ProductsServices/ENVIProducts/ENVI.aspx

de l'ontologie pour l'étiquetage sémantique des pixels sans aucune intervention experte et des contraintes générées automatiquement pour guider le clustering.

Nous rappelons que notre ontologie contient deux concepts thématiques. Connaissant *a priori* le nombre de classes, nous avons donc fixé le nombre de clusters à k = 3 pour la phase du clustering dans toutes nos expérimentations.

Pour chacun des jeux de données analysés, 30 mille contraintes sont générées aléatoirement à partir des données étiquetées par raisonnement et injectés lors de la phase du clustering. Les poids des contraintes $\overline{w}$ et w ont été fixés à 1. La valeur des poids et le nombre de contraintes ont été choisis expérimentalement après plusieurs tests.

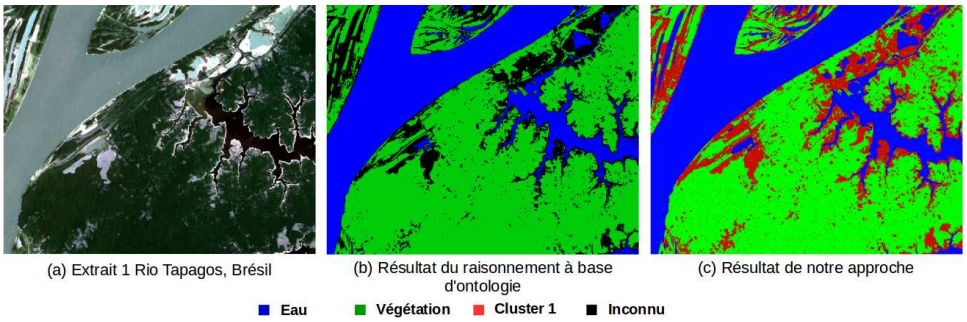


FIGURE 6.16 – Résultat de l'application de notre approche sur un extrait de la rivière *rio Tapajos*, Brésil

La figure 6.16 illustre les résultats de la mise en œuvre sur le premier extrait des images d'Amazonie traitées. Nous pouvons apporter différents commentaires.

En observant les imagettes (b) et (c) dans la figure, deux aspects importants peuvent être facilement observés. Le premier concerne l'amélioration de l'étiquetage des pixels par notre approche par rapport à l'ontologie comme le montre la partie en haut à gauche des imagettes (b) et (c). Sur cette portion de l'image, nous pouvons visualiser que des pixels eau (en bleu) n'ont pas été étiquetés par le raisonnement mais l'ont bien été par le clustering par contraintes. Ainsi, l'approche hybride a complété l'attachement sémantique des pixels aux concepts de l'ontologie. Ce mécanisme naturel à l'enrichissement souligne la capacité de notre approche à pallier l'incomplétude des définitions des concepts.

Le deuxième aspect est relatif à la découverte de nouveaux clusters non décrits par l'ontologie. Nous observons dans cette figure la combinaison entre des clusters sémantiquement étiquetés par les concepts thématiques (Végétation et Eau), et un cluster induit par le clustering (cluster 1).

Les expérimentations montrent ainsi la capacité de notre approche à utiliser efficacement les connaissances, même quand elles sont incomplètes, tout en permettant l'émergence de nouvelles classes thématiques non spécifiées.

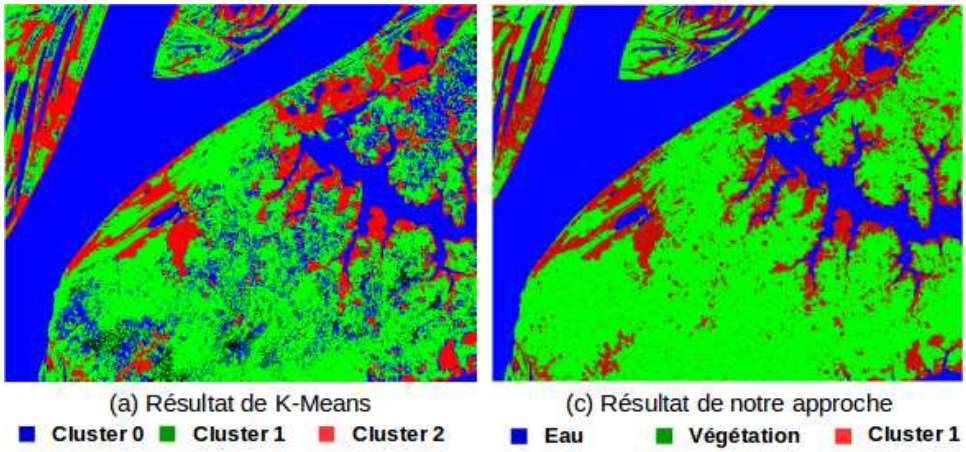


FIGURE 6.17 – Comparaison des résultats de K-Means avec ceux de notre approche sur l'extrait 2 de la région d'Amazonie, Brésil

La figure 6.17 présente les résultats d'un K-Means (a) et ceux de notre approche (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 remarquer que le clustering 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 une partie de ces problèmes et aide le clustering à bien distinguer l'eau de la végétation.

Une autre différence capitale réside dans l'interprétation sémantique des clusters. En effet, dans notre approche l'interprétation des instances des concepts présents dans l'ontologie est automatique, tandis que dans le cadre du clustering, l'intervention de l'expert demeure obligatoire pour étiqueter les clusters obtenus.

Images	Clustering		Ontologie			Approche proposée	
	Prec.	F-Mes.	% Étiqueté	Prec.	F-Mes.	Prec.	F-Mes.
Amazonie 1	0.8899	0.8764	83,6	0.8359	0.8360	**0.9445**	**0.9296**
Amozonie 2	0.8701	0.8598	81,26	0.8125	0.8126	**0.9271**	**0.9181**
Amazonie 3	0.8889	0.9241	90,33	0.9031	0.9020	**0.9299**	**0.9304**

TABLE 6.1 – 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 (chapitre 3, section 3.2.4) par rapport à la classification de référence disponible (section 6.4.2).

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, à la différence de notre approche où l'étiquetage des pixels se fait automatiquement.

Le tableau présente les mesures obtenues sur les trois extraits d'Amazonie en se basant sur les classifications de référence. Concernant les résultats obtenus par l'ontologie, nous rappelons que l'ontologie adopte l'hypothèse du monde ouvert et qu'elle contient la formalisation de deux classes sur trois (Eau et Végétation). Par conséquent, l'ontologie ne classifie pas toutes les instances (83,6 % pour le premier extrait). En calculant les indices de qualité uniquement sur la partie étiquetée, nous avons obtenu des mesures de l'ordre de 0.99. En revanche, en ramenant les calculs sur toutes les instances, on obtient les résultats rapportés dans le tableau 1, puise qu'aucune instance de la troisième classe n'est étiqueté par raisonnement. D'où l'intérêt d'utiliser le clustering pour obtenir une classification complète de l'image.

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 tend à prouver qu'en complément de l'apport sémantique, notre approche améliore globalement la qualité du clustering.

Résultats sur la région du sud de la France

Jusqu'ici, nous avons appliqué notre approche sur différentes images qui concernent la région de *Rio Tapajós*. Dans ce qui suit, les expérimentations menées sur une image du sud de la France sont présentés et discutés.

Le même protocole expérimental a été reconduit, avec en entrée une TBox à deux concepts thématiques et une image non étiquetée et calibrée. Les deux indices *RVI* et *NDVI* sont calculés et ajoutés aux bandes spectrales présentes dans l'image. Nous obtenons donc en entrée de l'approche un jeu de données $\mathcal{X} = \{x_i\}_{i=1}^{468.000}$, avec $x_i \in \mathbb{R}^9$.

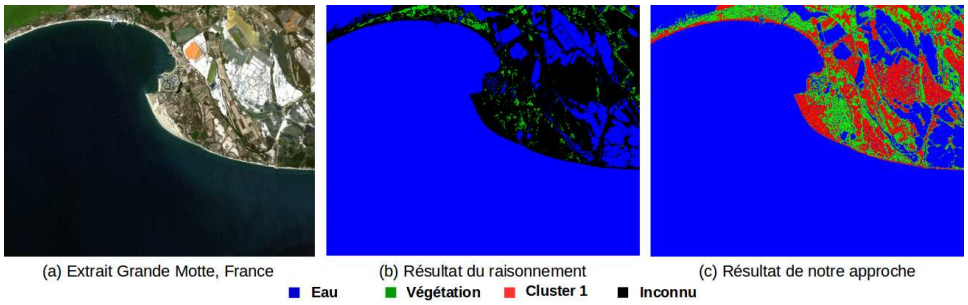


FIGURE 6.18 – Résultat de l'application de notre approche sur un extrait de la *Grande Motte*, France

La figure 6.18 montre les résultats obtenus par notre approche, avec la projection de l'étape intermédiaire du raisonnement dans l'image (b). Sur cette image, les résultats du raisonnement montrent une bonne détection

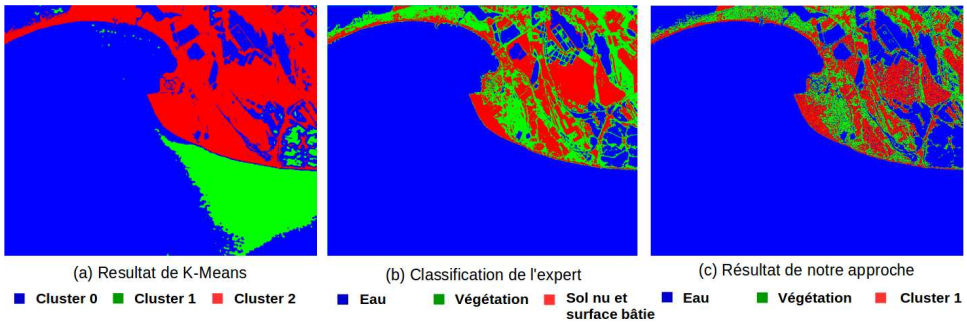


FIGURE 6.19 – Comparaison de K-Means et de notre approche par rapport à la classification de référence sur l'extrait de la *Grande Motte*, France

des pixels eau, tandis que les pixels de végétation ont été cette fois-ci peu détectés par l'ontologie. Cela montre que la définition du concept thématique végétation n'englobe pas toutes les possibilités de l'apparition des pixels de végétation.

En général, les experts ont tendance à définir les concepts thématiques en s'appuyant sur des intervalles de valeurs sûrs, ce qui explique la précision élevée de l'étiquetage par ontologie, mais cela pénalise les résultats de la classification, comme le montre les résultats de l'étiquetage par raisonnement des pixels de végétation sur la figure et ceux de l'eau sur la figure 6.16.

Avec notre proposition, cet inconvénient est résolu grâce au clustering par contraintes. En effet, même avec un nombre réduit de pixels végétation étiquetés, notre approche a pu détecter une grande partie des pixels de végétation réstants (figure), donnant une illustration supplémentaire de la capacité de notre approche à exploiter des connaissances incomplètes. La figure 6.19 donne une comparaison des résultats de notre approche et de K-Means avec la classification de l'expert. Cette comparaison montre un aspect intéressant de notre proposition. Il s'agit de la capacité à construire des groupes respectant la vision de l'utilisateur spécifiée dans l'ontologie. Quand on applique le K-Means à 3 clusters sur l'image, deux clusters contenant tout les deux des pixels eau se forment. Cela s'explique par la quantité importante des pixels eau qui causent un problème de déséquilibre de clusters pour K-Means, s'ajoute à cela les problèmes d'initialisation propre à la classification non-supervisée.

Ces deux problèmes sont évités avec notre approche, puise que l'initialisation se fait avec les pixels étiquetées par l'ontologie et que le respect des contraintes aide à la formation de clusters ayant des tailles relativement différentes.

6.5 Discussions

Nous avons présenté dans ce chapitre 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'un ontologie. En proposant une approche d'une nature à la fois déductive et inductive, nous avons pu obtenir une méthode robuste face à l'incomplétude des connaissances mais qui permet au même temps de prendre 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 significatifs que ce soit au niveau de l'étiquetage sémantique des clusters ou celui de l'amélioration de la qualité du clustering.

L'application de notre approche sur des extraits d'images capturées à différents endroits (Brésil et France) en utilisant la même ontologie a montré son potentiel de généricité.

6.6 Valorisation scientifique

Les travaux exposés dans ce chapitre ont fait l'objet de trois publications. Deux dans des conférences francophones avec comité de sélection. La troisième publication a concerné toute l'approche et s'est faite dans la conférence internationale *IEEE DSAA, International Conference on Data Science and Advanced Analytics* dont le taux d'acceptation est de 20.1 %.

— *On the Use of Ontology as a priori Knowledge into Constrained Clustering*,
IEEE International Conference on Data Science and Advanced Analytics, **DSAA 2016** - Montréal, Canada

— Génération de contraintes pour le clustering à partir d'une ontologie - Application à la classification d'images satellites,
Conférence Extraction et Gestion des Connaissances, EGC 2016 - Reims, France (Nominé pour le prix du meilleur article académique)

— Approche hybride à base d'ontologie pour le clustering par contraintes,
Conférence internationale francophone sur la science de la données, AAFD & SFC 2016 - Marrakech, Maroc

Chapitre 7

Optimisation du raisonnement à base d'ontologie par clustering topographique

Nous allons présenter dans ce chapitre, notre seconde contribution qui allie à la fois du clustering topographique à base de cartes SOM et du raisonnement pour faciliter l'interprétation automatisée de grands volumes de données quantitatives.

À cet effet, notre approche optimise les activités de raisonnement en réduisant la taille de l'ABox, en d'autres termes, en ne définissant qu'un individu par prototype de la carte SOM construite par clustering topographique. Cette vision très synthétique mais cependant juste des données en entrée permet d'une part de réduire considérablement le temps nécessaire à la reconnaissance d'instances par le raisonneur à base des logiques de description et d'autre part de fournir un étiquetage sémantique qui s'appuie sur les connaissances du domaine à la carte SOM.

Sommaire

"L'ignorance mène à la peur, la peur mène à la haine et la haine conduit à la violence. Voilà l'équation." **Averroès**

7.1 Introduction et motivations

Les algorithmes de clustering sont incontournables dès lors qu'il s'agit d'extraire des connaissances à partir de données non-étiquetées. Dans ce sens, les cartes de Kohonen figurent parmi les approches de clustering les plus efficaces lorsque les données sont en nombre, en particulier pour la compression et la visualisation des données.

Le clustering SOM permet en effet de représenter les données au travers d'une carte spatialement organisée. Les données ou encore instances sont catégorisées en des *mini-clusters* (neurones), et la carte apprise est organisée de manière à ce que les instances proches dans l'espace de représentation des données se situent dans des neurones proches dans le carte.

Cependant, SOM reste un algorithme d'apprentissage non-supervisé, l'étiquetage et l'interprétation des résultats en sortie du clustering demeurent des activités manuelles. Ce passage obligé peut parfois être problématique, notamment quand une forte connaissance du domaine est exigée comme en télédétection. La mise en place d'un système d'extraction de connaissances automatisé nécessite donc d'alléger le coût de la tâche d'expertise en sortie de l'activité de clustering.

Le parti-pris est de voir comment un langage de représentation de connaissances à l'exemple du langage OWL 2 (chapitre 2), et des raisonneurs associés peuvent contribuer à dégager automatiquement de nouvelles connaissances en raisonnant sur les résultats en sortie de clustering.

Une des limites actuelles des algorithmes d'inférence qui existent pour OWL, réside dans leur incapacité à raisonner sur des ontologies à très larges ABox [HORROCKS et al., 2004 ; BAADER, 2003]. L'explication en est que l'inférence de nouvelles connaissances passe par de l'expansion de règles qui ont tendance à faire augmenter considérablement le modèle qui représente l'ontologie en mémoire vive. Un tel inconvénient freine donc fortement l'utilisation du raisonnement dans des applications du monde réel.

Dans ce chapitre, nous explorons à la fois les potentialités des cartes de Kohonen et des ontologies OWL 2 afin de proposer une méthode originale allant vers la définition d'un système d'interprétation de données automatisé et performant.

Notre principale motivation est en effet de développer une approche hybride capable de tirer bénéfice des points forts du raisonnement ontologique et du clustering SOM, tout en essayant de s'affranchir de leurs points faibles tout relatifs.

7.2 Raisonnement sur une base de connaissances de grande taille

Nous proposons une approche que nous pensons originale qui permet à la fois d'optimiser le raisonnement à base de logiques de description sur des

boîtes assertionnelles de grande taille et d'automatiser l'étiquetage sémantique des vecteurs référents des cartes topologiques. L'optimisation du raisonnement est permise par une compression préalable des données. L'étiquetage automatique est quant à lui rendu possible en exploitant le raisonnement à base d'ontologie.

A cet effet, notre méthode s'appuie en premier lieu sur la capacité des cartes de Kohonen à représenter les données au travers d'un nombre réduit de prototypes. Dans l'absolu, la base de connaissances aurait du contenir une boîte assertionnelle de très grande taille à même d'organiser toutes les données en entrée. La carte topologique permet de faire une économie d'échelle, la boîte assertionnelle n'est peuplée que d'individus définis à partir des prototypes et qui représentent un condensé du contenu des données d'origine. Le raisonnement à base des logiques de description combiné à une méthode de propagation d'étiquetage aux instances pour fournir une interprétation automatisée des cartes et des données d'entrées.

La figure 8.3 offre une vue générale de l'approche proposée. L'enchaînement séquentiel des étapes y est illustré avec en ligne de mire l'allègement du coût du raisonnement et l'automatisation de l'étiquetage des résultats du clustering et des données.

À partir d'un ensemble d'instances non-étiquetées $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$, la première étape consiste à réduire le nombre d'instances à l'aide du clustering topographique SOM. L'algorithme permet de représenter les données au travers une carte auto-organisée topographiquement (en gris dans la figure) où chaque neurone est représenté par un prototype $\{w_j\}_{j=1}^{k} \in \mathbb{R}^d$ résumant les données qui lui sont affecté.

Une fois les prototypes obtenus, la deuxième étape consiste à les transformer en un ensemble d'assertions qui vont venir alimenter une boîte assertionnelle notée ABox' [1], créant ainsi pour chaque prototype une instance OWL caractérisée par les propriétés de la connaissance formalisée (Tbox) en s'appuyant sur un processus semi-automatique.

Les représentations sémantiques des prototypes ainsi que les concepts formalisés (TBox) sont par la suite fournis au raisonneur. Dans notre figure, nous supposons que la TBox fournit trois concepts thématiques (colorés en rouge, vert et bleu sur la figure). Le raisonneur va par la suite permettre d'effectuer de la reconnaissance d'instances (en d'autres termes, faire appel au service d'inférence nommé *réalisation*) à partir de l'ontologie définie à partir de l'union de la TBox et de l'ABox'. Les individus de l'ABox' qui répondent à la description des concepts thématiques définis sont alors typés par ces concepts. La TBox ne contient qu'une partie des concepts pouvant être formalisés et en conséquence les individus ne sont pas tous typés.

Une fois le raisonnement effectué, nous sélectionnons dans l'ABox' les individus étiquetés et par prolongement les prototypes étiquetés. Les prototypes sont par ailleurs associés aux données qui sont à leur voisinage. Il reste alors à propager l'étiquetage sémantique à ces données. L'ABox peut être en partie reconstruite à partir de l'ABox'. De la même manière, la carte

1. Pour la différencier de ABox qui devrait contenir toutes les instances correspondant dans l'absolu à toutes les données en entrée

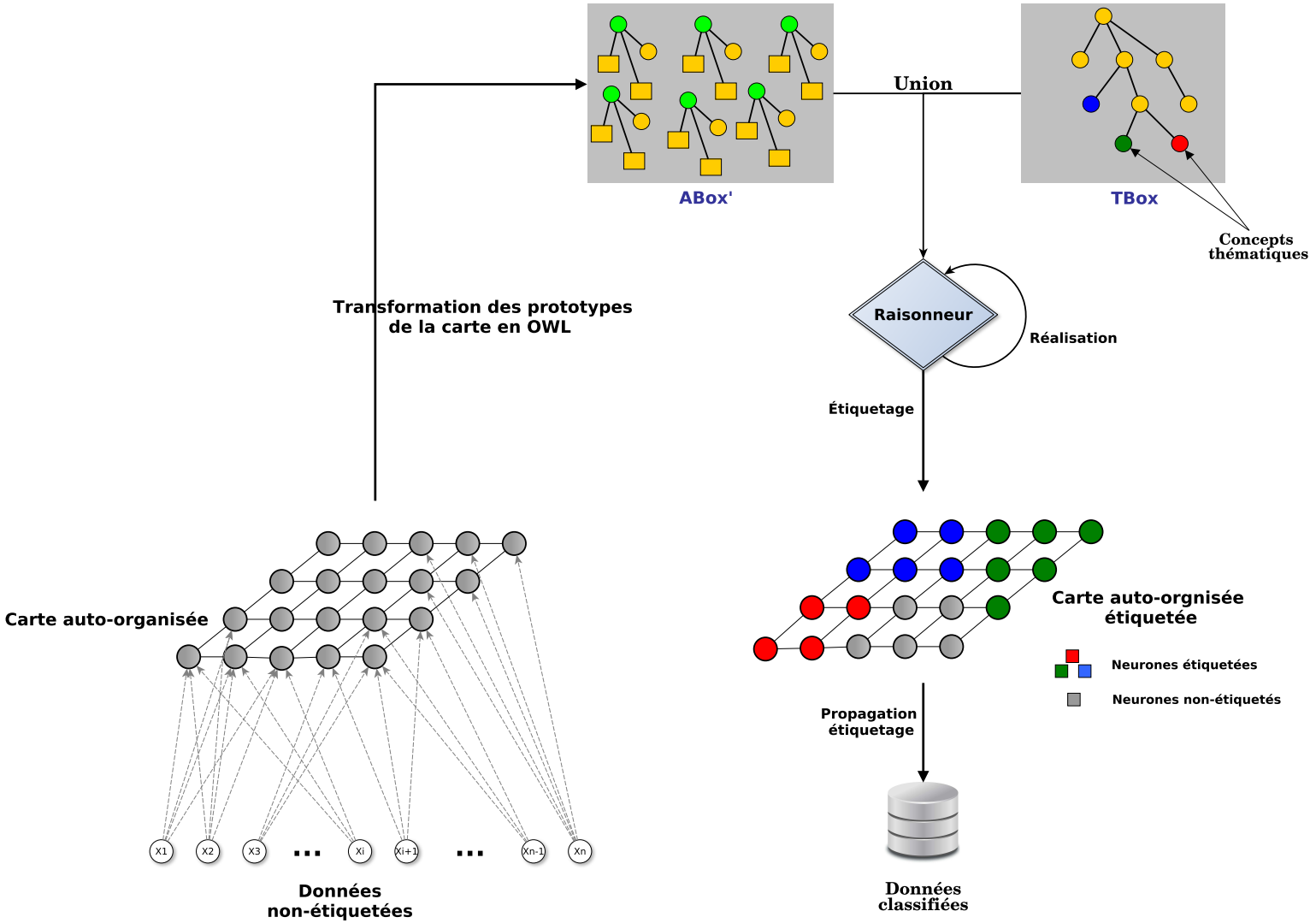


FIGURE 7.1 – Vue générale de notre proposition pour l'étiquetage à base d'ontologie optimisé par clustering topographique

produite après application de l'algorithme SOM est enrichie par le résultat de l'étiquetage sémantique par raisonnement.

Dans notre figure, nous pouvons ainsi voir qu'une partie de la carte est colorée par les couleurs dévolues aux concepts thématiques (rouge, vert et bleu) signifiant leur étiquetage et qu'une autre partie n'a pas été étiquetée (en gris). La partie en gris de la carte représente les neurones des prototypes ne correspondant à aucun concept thématique, comme le raisonnement s'opère dans l'hypothèse du monde ouvert (*OWA*), seuls les prototypes qui obéissent complètement aux règles définies par les concepts thématiques sont étiquetés.

Nous avons décrit l'enchaînement de notre approche et les différents rôles dévolus à chaque étape. Nous revenons maintenant plus en détail sur d'une part la mise en œuvre du clustering et d'autre part la construction de l'ontologie et la mise en jeu de mécanismes inférentiels. Le passage de l'une à l'autre de ces activités sont également décrits.

7.2.1 Clustering à base des cartes auto-organisatrices

Le premier pan d'activité de l'approche concerne le clustering des données à base de l'algorithme SOM pour en produire une synthèse et obtenir un nombre réduit de prototypes représentatifs. Nous utilisons la version classique de l'algorithme proposé par Kohonen [KOHONEN, 1990 ; KOHONEN, 2013] pour en construire une carte.

L'algorithme SOM prend en entrée un ensemble de données non-étiquetées $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$ et fournit en sortie une carte $\mathcal{M}$ (typiquement de deux ou trois dimensions) spatialement organisée. Cette carte peut être vue comme un graphe non orienté composé de neurones interconnectés, où chaque neurone est représenté par un prototype $w_j \in \mathbb{R}^d$ de la même dimension que les instances d'entrée.

La carte peut avoir une forme hexagonale ou rectangulaire, suivant la configuration choisie. Pour une carte à deux dimensions par exemple, la forme rectangulaire permet à chaque neurone d'avoir quatre voisins tandis que la forme hexagonale permet à chaque neurone d'être connecté à six neurones.

Chaque neurone de la carte influence son voisinage. Ainsi, plus deux neurones sont à proximité et plus leur influence mutuelle est grande. Le contrôle de l'influence d'un neurone sur son voisinage se fait par une fonction pondérée, positive et symétrique notée $\mathcal{K}$, où ($\mathcal{K} \geq 0$ et $\lim_{|y|\to\infty} \mathcal{K}(y) = 0$).

$\mathcal{K}$ est une fonction gaussienne (équation 7.1) qui permet de quantifier l'influence entre deux neurones notés m_j et m_k :

$$\mathcal{K}_{m_j,m_k} = \frac{1}{\lambda(t)} \exp(-\frac{d_1^2(m_j, m_k)}{\lambda^2(t)}) \tag{7.1}$$

Où $\lambda(t)$ est une fonction dite de température (équation 7.2) qui réduit l'influence des neurones les plus éloignés de la carte au fur et à mesure

de l'avancement de l'apprentissage (en d'autres termes du nombre d'itérations). La fonction $\lambda(t)$ varie entre λ_0 (température initiale) et λ_f (température finale), avec t_{max} pour le nombre total des itérations :

$$\lambda(t) = \lambda_0(\frac{\lambda_0}{\lambda_f})^{\frac{t}{t_{max}}} \tag{7.2}$$

$d_1(m_j, m_k)$ est la distance Manhattan calculée entre deux neurones m_j et m_k, dont les coordonnées sur la carte correspondent respectivement à (i_1, j_1) et (i_2, j_2) :

$$d_1(m_j, m_k) = |i_1 - i_2| + |j_1 - j_2| \tag{7.3}$$

Si on considère $X = \{x_i\}_{i=1}^n$ un ensemble d'instances décrites dans l'espace euclidien $\mathbb{R}^d$. Chaque neurone de la carte a un prototype (vecteur de référence) $w_j \in \mathbb{R}^d$ qui lui est associé. Ce prototype caractérise les clusters associés aux neurones de la carte. On note $\mathcal{W} = \{w_j\}_{j=1}^k$ l'ensemble des prototypes de la carte $\mathcal{M}$ et $\chi(x_i)$ la fonction qui assigne chaque instance x_i au plus proche neurone de la carte.

L'objectif de SOM devient la minimisation de la fonction objective suivante :

$$R_{SOM} = \sum_{i=1}^{n}\sum_{j=1}^{k} \mathcal{K}_{j,\chi(x_i)} \|x_i - w_j\|^2 \tag{7.4}$$

La figure 7.2 illustre visuellement une carte SOM accompagnée de la fonction objective et celle du voisinage.

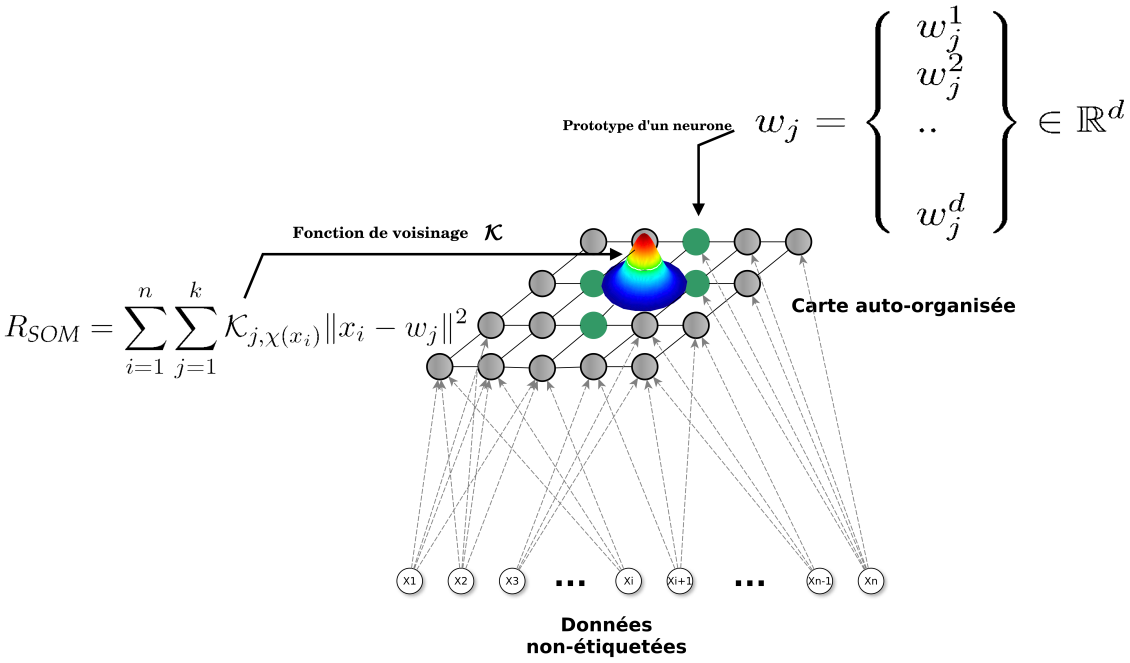


FIGURE 7.2 – Illustration d'un carte SOM

Cette fonction objective peut être minimisée de différentes manières [KOHONEN, 2013]. Nous avons choisi d'utiliser la version proposée par KOHONEN (1990) dans notre approche. L'algorithme 6 détaille ces différentes étapes :

Paramètres : Données $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$
Taille de la carte k et sa forme
λ_0, ϵ_0 la température et le pas d'apprentissage initiaux
et t_{max} le nombre maximum d'itérations

Résultat : Carte auto-organisée : $\mathcal{M}$
Ensemble des prototypes : $\mathcal{W} = \{w_j\}_{j=1}^{k}$

(1). *Initialisation aléatoire des prototypes* $\mathcal{W} = \{w_j\}_{j=1}^{k}$;

tant que $t \leq t_{max}$ **faire**
  (2). *Phase de compétition*;
  pour chaque *donnée* x_i *choisie aléatoirement dans* X **faire**
    Choisir le neurone $m_j^* \in \mathcal{M}$ *gagnant,* ***Tel que***;

$$d(x_i, w_j^*) = argmin_{1 \leq j \leq k}.\|x_i - w_j\|^2 \tag{7.5}$$

  fin
  (3). *Phase de coopération*;
  pour chaque $w_j \in \mathcal{W}$ **faire**

$$w_{j(t+1)} = w_{j(t)} - \epsilon_t.\mathcal{K}_{m_j, m_j^*(t)}.\|w_{j(t)} - x_i\|^2 \tag{7.6}$$

  fin
  (4). *Phase d'adaptation*;
  Réduire le pas d'apprentissage et la fonction de température;

$$\epsilon_t = \epsilon_0 \exp(-\frac{t}{t_{max}}) \tag{7.7}$$

$$\lambda_t = \lambda_0 \exp(-\frac{t}{t_{max}}) \tag{7.8}$$

fin

Algorithme 6 : Algorithme SOM

A l'issue de la phase d'apprentissage, les prototypes obtenus correspondent à une représentation réduite mais néanmoins fidèle aux données d'entrées.

De plus, SOM est une méthode peu gourmande en terme de coût de calcul, ce qui rend son application adaptée dans un contexte de grandes masses de données.

7.2.2 Raisonnement et étiquetage sémantique des données

Une fois la carte produite, nous sélectionnons le prototype produit par chaque neurone. Ces prototypes sont par la suite transformés en instances OWL. Cette transformation est assurée par le même processus semi-automatique introduit lors de notre première contribution (chapitre 6). Nous rappelons ici ses différentes étapes :

Paramètres :
Ensemble des prototypes : $\mathcal{W} = \{w_i\}_{i=1}^{k}$ décrits par : $V = \{v_j\}_{j=1}^{d}$
TBox : $\mathcal{T} =< \mathcal{N}_\mathcal{C}, \mathcal{N}_\mathcal{P} >$
Résultats :
ABox' : $\mathcal{A}' = \{a_i\}_{i=1}^{n}$
Méthode :
Pour tout p_k in $\mathcal{N}_\mathcal{P}$ **and** v_j in V **Faire**
Boolean Requête = Correspondance entre p_k et v_j
Si Requête.valide() **Alors**
$map(\mathcal{N}_\mathcal{P}, V)$.ajout($p_k$,$v_j$)
Fin Si
Fin Pour
Pour tout $w_i \in \mathcal{W}$ **Faire**
$a_i := createOWLInstance()$;
Pour tout $p_k \in map(\mathcal{N}_\mathcal{P}, V)$ **Faire**
$a_i.addProperty(p_k)$
$a_i.setPropertyType(p_k, \mathcal{T}.getPropertyType(p_k))$
$a_i.setPropertyValue(p_k, w_i.getValueOf(v_j))$
Fin Pour
return a_i : Axiomes OWL représentants w_i
$\mathcal{A}.add(a_i)$
Fin Pour

Algorithme 7 : Transformation semi-automatique des prototypes en OWL

L'algorithme présenté prend en entrée les prototypes et la TBox. En se basant sur les propriétés $\mathcal{N}_\mathcal{P}$ définies par les connaissances expertes et l'ensemble des variables descriptives des prototypes V. Notre processus suggère à l'utilisateur de lier chaque variable à la propriété qui lui correspond. Il est aussi possible de fournir à notre processus un mapping préalable en complément de la TBox et des données. Dans ce cas, La création de l'ABox se fait sans aucune intervention humaine. Une fois le mapping réalisé, le processus va générer des axiomes OWL qui vont venir enrichir les instances OWL et les décrire sémantiquement. Ces axiomes attachent les propriétés adéquates aux instances, alimentent leurs valeurs par les valeurs des prototypes $\mathcal{W}$, puis les typent en s'appuyant sur les définitions de ces propriétés dans la TBox.

Ce processus produit des instances OWL représentant les prototypes obtenus par l'algorithme SOM. Les instances ainsi définies forment l'ABox'

dans notre approche. C'est une sorte de représentation réduite de ce qu'aurait pu être une ABox représentant toutes les données d'entrées. L'ABox' est ensuite associée à la TBox contenant la définition des concepts thématiques dans un raisonneur.

En invoquant le service d'inférence de type *réalisation*, le raisonneur va inférer les prototypes représentés dans l'ABox' qui satisfont les définitions des concepts. Comme nous l'avons expliqué auparavant (chapitre 2), la *réalisation* consiste à trouver pour une instance donnée a_i dans une ontologie, le concept C_i le plus spécifique [BAADER, 2003].

Nous utilisons dans notre approche le raisonneur Pellet [SIRIN et al., 2007] à cette fin. Son implémentation efficace des spécifications d'OWL 2 et son support du raisonnement sur les domaines concrets [LUTZ, 2003] font de lui un raisonneur adapté à nos besoins.

En étiquetant les prototypes par les concepts thématiques, le raisonnement fournit un étiquetage de la carte SOM sans l'intervention manuelle de l'expert. C'est un étiquetage modulaire, qui ne nécessite pas d'instances étiquetées et qui s'adapte aux concepts présents dans la TBox

Nous propageons par la suite l'étiquetage sémantique aux instances captées par les neurones dont les prototypes ont été typés. Cela se fait en s'appuyant sur les résultats du clustering SOM. En effet, chaque instance $x_i \in \mathcal{X}_{i=1}^{n}$ est attribuée à un neurone $m_j \in \mathcal{M}_{j=1}^{k}$ avec la fonction $\chi(x_i) \Rightarrow m_j$.

7.3 Validation expérimentale

Cette section présente les expérimentations menées pour valider notre proposition et les résultats obtenus. L'objectif de ces expérimentations est double. Il vise en même temps à démontrer la capacité de notre approche à étiqueter automatiquement les cartes auto-organisées obtenues par l'algorithme SOM, et à souligner son efficacité pour optimiser le raisonnement et permettre le raisonnement sur de grandes masses de données.

Nous avons mené nos expérimentations sur deux jeux de données différents. Le premier porte sur la classification de familles de vins, à partir du jeu nommé *Wine dataset* [2] qui est emprunté au répertoire *UCI* pour la mise à disposition de données d'apprentissage artificiel. Le deuxième concerne un extrait d'image satellite LANDSAT 5 TM sur la région de l'Amazone au Brésil.

7.3.1 Expérimentations sur le *wine dataset*

Le *wine dataset* est composé de 178 instances décrites au travers de 13 variables numériques $x_i \in \mathbb{R}^{13}$. Ces variables décrivent la quantité des constituants du vin analysé (alcool, intensité de la couleur, acidité, ...).

Notre approche prend en entrée les instances non-étiquetées provenant du jeu de données et une ontologie qui décrit trois concepts thématiques.

2. Wine dataset : http ://archive.ics.uci.edu/ml/datasets/Wine

Les concepts thématiques ont été obtenus en appliquant une approche similaire à celle proposée par SHEEREN et al. (2006) (chapitre **??**). Dans notre cas, nous appliquons l'algorithme d'arbre de décision, proposé par BREIMAN et al., pour apprendre des règles de classification à partir du jeu de données et de sa classification de référence. Ces règles sont par la suite transformées manuellement en ontologie OWL.

Notre objectif ici n'est pas d'explorer l'induction de règles de classification pour la construction d'ontologies, mais plutôt d'obtenir des concepts qui vont venir illustrer une connaissance du domaine lors de nos expérimentations et permettre l'utilisation du raisonnement dans le cadre de notre approche sur des données non-étiquetées.

Chaque concept de notre ontologie est formalisé avec le langage OWL 2 en utilisant les domaines concrets pour définir les intervalles d'appartenance aux concepts sur les propriétés décrivant les instances. 7.9 donne un extrait de la définition du premier concept thématique :

$$\begin{aligned} Wine\_Vineyard\_1 \equiv Wine \wedge (\exists flavanoids_{\{\geq 2.165\}} \\ \wedge \exists magnesium_{\{<135.5\}} \wedge \exists proline_{\{\geq 755\}}) \end{aligned} \tag{7.9}$$

Nous suivons les étapes spécifiées dans la figure 7 lors de son application sur le jeu de données.

Premièrement, nous appliquons l'algorithme SOM sur les instances non-étiquetées pour obtenir la carte résumant les données. SOM exige la spécification à priori de la taille de la carte (donc du nombre de neurones). Nous fixons la taille de la carte à 6×11. Nous choisissons une configuration rectangulaire à deux dimensions pour la forme de la carte. La taille de la carte à été choisie expérimentalement.

Une fois le clustering appliqué, nous exploitons les prototypes de la carte. Ces prototypes sont ensuite transformés en instances OWL par notre processus semi-automatique et l'ABox' est construite.

Nous injectons l'ABox' avec les connaissances expertes dans le raisonneur Pellet. Nous obtenons donc une base de connaissances avec 66 instances OWL (correspondant aux prototypes transformés) et trois concepts thématiques.

Le raisonneur type par la suite les prototypes avec les concepts à l'aide des mécanismes inférentiels de *réalisation* sur toutes les assertions (ABox' + TBox).

Pour les prototypes typés, l'étiquetage est automatiquement propagé aux instances associées en s'appuyant sur les résultats du clustering obtenus dans la précédente étape.

L'évaluation de nos résultats s'est appuyée sur la classification de référence fournie avec le jeu de données. Nous avons mesuré la qualité de nos résultats en se basant sur deux indices. Le premier est la pureté de la carte. La pureté évalue la qualité du clustering par rapport aux données étiquetées attribuées aux neurones. Un cluster est considéré comme pur s'il ne contient que des instances d'une seule classe. Et impur si, au contraire, il

contient des instances appartenant à différentes classes. Le deuxième indice utilisé est l'exactitude (section 3.2.4).

Pour nos expérimentations sur ce jeu de données, la pureté obtenue est de 96,62% et l'exactitude est de 92,42% sur les prototypes étiquetés. Ce résultat est très encourageant au vu de la qualité des résultats obtenus. Il montre la faisabilité de l'approche et sa capacité d'étiquetage automatique des neurones de la carte et des données non-étiquetées avec notre méthode combinant SOM avec le raisonnement à base d'ontologie.

Le jeu de données est composé de 178 instances dans ces expérimentations. L'intérêt ici est de pointer la capacité de notre approche à préserver la qualité du raisonnement. Dans le second lot d'expérimentations, nous allons nous pencher sur le potentiel d'optimisation du raisonnement sur de grandes masses de données de notre approche.

7.3.2 Interprétation d'images satellite

Nous présentons dans cette partie l'application de notre approche à l'interprétation d'une image satellite. L'image utilisée est un extrait d'une scène LANDSAT 5 TM (figure 6.9). C'est une image que nous avons déjà exploitée lors de nos premières expérimentations (chapitre 6) et qui porte sur la rivière *Rio de Tapajos*, Brésil. L'image est calibrée en réflectance et les indices *NDVI* et *RVI* sont calculés en s'appuyant sur la librairie *Orfeo Toolbox*.

À l'issue des prétraitements, nous obtenons un ensemble d'instances non étiquetées $\mathcal{X} = \{x_i\}_{i=1}^{468.000} \in \mathbb{R}^9$.

La TBox contient les connaissances que nous avons formalisées et présentées dans la chapitre précédent (chapitre 6). Ces connaissances contiennent deux concepts thématiques, dont nous rappelons ci-dessous la définition :

$$\begin{aligned} Pixel\_Eau \equiv Pixel \wedge ((\exists nir\_band_{\{<0.05\}} \wedge \exists ndvi._{\{<0.01\}}) \\ \vee (\exists nir\_band_{\{<0.11\}} \wedge \exists ndvi_{\{<0.001\}})) \end{aligned} \tag{7.10}$$

$$\begin{aligned} Pixel\_Vegetal \equiv Pixel \wedge ((\exists ndvi_{\{>0.3\}}) \\ \vee (\exists rvi_{\{>2.5\}})) \end{aligned} \tag{7.11}$$

Nous avons menés nos expérimentations en appliquons les différentes étapes spécifiées par notre approche. L'ensemble $\mathcal{X}$ des pixels non étiquetés est dans un premier temps compressé à l'aide de l'algorithme SOM (algorithme 6). La taille de la carte a été fixée à 20×20 lors de nos expérimentations. Cette taille a été choisie expérimentalement après plusieurs essais.

Les prototypes obtenus sont par la suite transformés par notre processus semi-automatique (algorithme 7). Les instances OWL obtenues composent ainsi l'ABox' réduite. Cette boite assertionnelle, composée uniquement de 400 prototypes résumant les pixels en entrée, est couplée avec la TBox puis fournie au raisonneur Pellet [SIRIN et al., 2007].

À l'aide du service d'inférence *réalisation*, les prototypes OWL respectant les règles définies dans les concepts thématiques Eau ou Végétation sont automatiquement étiquetés. En se basant sur les résultats de SOM,

nous propageons cet étiquetage aux instances captées par les neurones étiquetées. Cela permet d'obtenir une carte SOM et des données étiquetées suivant les connaissances formalisées de l'expert.

La figure 7.3 rapporte les résultats obtenus sur l'image étudiée.

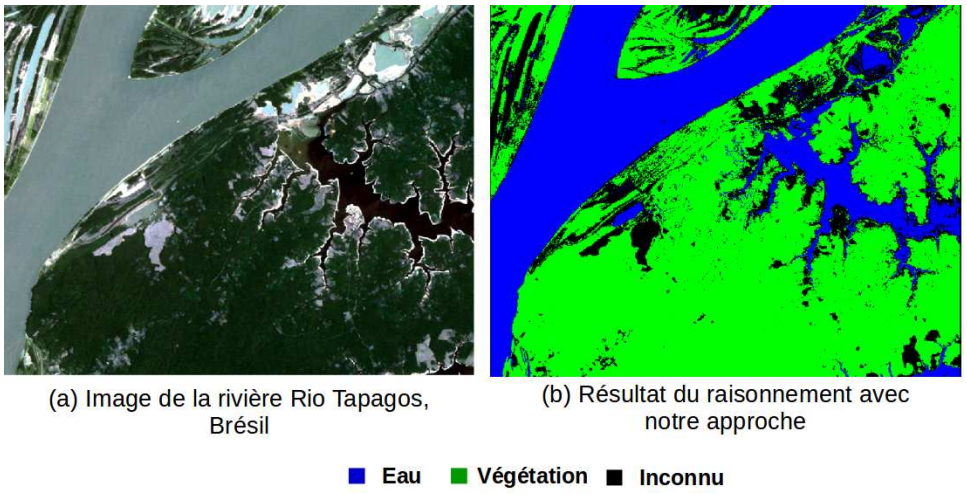


FIGURE 7.3 – Application de notre approche sur un extrait d'image non-étiqueté

Le taux d'étiquetage obtenu par notre approche est de 76% de la carte, soit 304 prototypes. Sur la partie étiquetée, nous avons obtenu une pureté de 98,45%. Ces résultats montrent la capacité de notre approche à préserver la qualité du raisonnement tout en améliorant sa rapidité et en garantissant son passage à l'échelle.

Le raisonnement a été effectué uniquement sur 400 prototypes et non sur les 468.000 pixels de l'image. Nous pouvons voir visuellement que les résultats sont très satisfaisants. La plupart des pixels végétation et eau sont correctement étiquetés. Les pixels en noir représentent les pixels attachés à des neurones dont les prototypes n'ont pas été étiquetés par raisonnement. Ces pixels correspondent pour la plupart aux pixels de la troisième classe sol nu et bâtiment.

Une partie des pixels non étiquetés en haut à gauche de la figure sont des pixels eau. Ces pixels correspondent pour la plupart aux mêmes pixels qui n'ont pas été étiquetés en raisonnant sur toute l'image, comme le montre la figure 7.4 de comparaison. Cette incapacité d'étiquetage peut donc être amputée à l'incomplétude de la définition du concept par l'expert et non à une défaillance dans notre approche.

Nous rappelons que l'étiquetage des prototypes est basé sur les connaissances et se fait sans intervention de l'expert. Dans le cadre non-supervisé classique de SOM, l'étiquetage est fait par l'expert, qui dans ce cas aurait du s'investir à étiqueter manuellement les 400 prototypes ou utiliser d'autres méthodes de classification en plus du *clusteing* SOM.

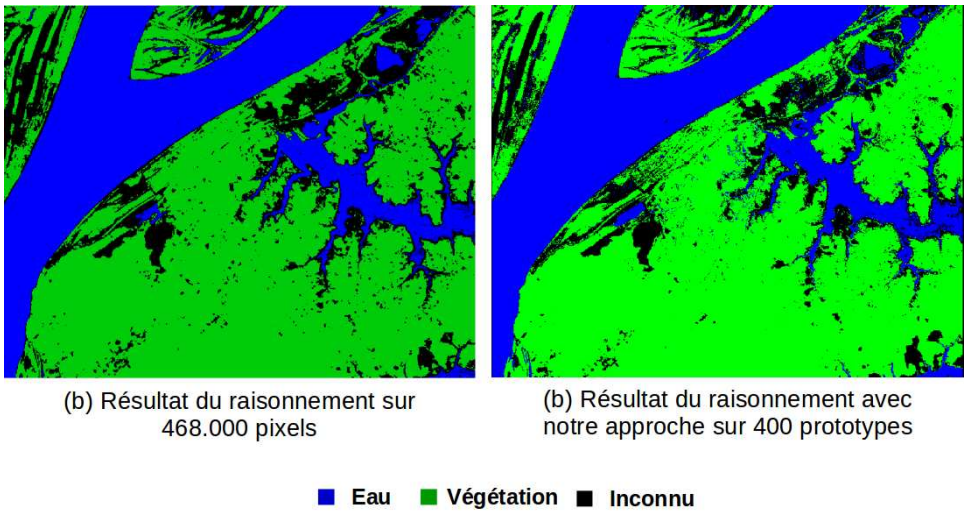


FIGURE 7.4 – Comparaison des résultats de notre approche avec les résultats de raisonnement à base d'instances

La figure 7.4 affiche les résultats de notre approche et les résultats du raisonnement sur toutes les instances de l'image. Il faut noter que le raisonnement sur toute l'image est rendu possible par une séparation des instances de l'ABox et un raisonnement séquentiel sur plusieurs ontologies, où chaque ontologie est constituée de la même TBox et d'une seule instance OWL à chaque fois.

Actuellement, sans séquencement du raisonnement, les raisonneurs ne sont pas capable de raisonner sur des ABox de cette grandeur. Si on se réfère aux chiffres fournis par HORROCKS et al. (2004), les raisonneurs ne peuvent pas raisonner sur des ABox à plus de 200.000 instances. ANDRÉS (2013) ont quant à eux rapporté que le raisonneur Fact++ [TSARKOV et HORROCKS, 2006] ne pouvait aller au delà de 22.500 instances quand les domaines concrets étaient exploités.

Bien que la capacité du raisonnement soit liée, en plus de la taille de l'ABox, à d'autres éléments d'une égale importance comme la taille de la TBox, la complexité des constructeurs utilisés ; notamment la présence des domaines concrets, ainsi que la capacité matérielle du système embarquant le raisonneur. Il est prouvé que la complexité du raisonnement évolue exponentiellement avec l'ajout d'axiomes dans l'ontologie [BAADER, 2003]. Dans ce cadre, notre proposition peut être vue comme un moyen efficace pour permettre le passage à l'échelle du raisonnement sur des ABox de très grandes tailles.

La séparation des instances de l'ABox pour le raisonnement sur les pixels implique l'adoption d'une hypothèse importante. Celle que la définition des concepts thématiques s'appuie uniquement sur les propriétés propres aux instances et ne dépend pas de leurs relations avec leurs relations avec d'autres instances de l'ABox.

Dans le cadre des images satellites, cela peut se traduire par la définition des classes d'occupation du sol en s'appuyant sur les réponses radiométriques, les indices spectraux ou les indices de texture, et non sur le voisinage des pixels. Cet hypothèse est souvent adopté par les algorithmes d'apprentissage. Bien que dans un contexte complètement différent, les instances dans l'algorithme SOM par exemple sont évaluées uniquement à partir de leurs valeurs propres représentées dans l'espace $\mathbb{R}^d$.

En plus de permettre le raisonnement sur de plus fortes volumétries de données ; le temps du raisonnement est également fortement réduit, comme l'indique le tableau 7.1. Ces mesures ont été prises sur la même machine et avec la même configuration à chaque expérimentation. Nous utilisons la version Pellet 2.3.1, la version 1.7 du Java SE, la JVM a un espace alloué de mémoire vive de 4 Gb et tourne sur un système d'exploitation Ubuntu 14.04 LTS avec un processeur Intel Core i7-3632QM CPU @ 2.20GHz $\times$ 8.

Sur notre extrait, raisonner directement sur les 468.000 pixels en une seule fois n'est pas faisable. En adoptant le raisonnement séquentiel, le raisonnement est rendu possible et le temps d'exécution enregistré est de 9 min 36sec. Avec notre proposition, nous divisons le temps de raisonnement de plus de 570 fois. Cela s'explique par notre raisonnement sur 400 instances uniquement avec la même TBox. Le temps d'exécution de SOM dans notre cas s'élève à 19 secondes. Le temps d'exécution de l'approche devient alors de 20 secondes. Il faut noter que dans le cadre du raisonnement, la complexité du raisonnement augmente avec la taille de toute l'ontologie, ce qui inclut la TBox aussi. Tandis que dans le cadre des cartes de Kohonen, elle dépend de la taille des données (ABox) et de la taille de la carte, et non de la TBox.

Image	Raisonnement classique	Raisonnement séquentiel	Raisonnement avec notre approche
Amazonie 1	Infini	9 mn 36 s	19 s (clustering) + 1 s (raisonnement)

TABLE 7.1 – Temps du raisonnement des différentes méthodes sur l'extrait de l'image satellite étudiée

7.4 Discussions

Dans ce chapitre, nous avons présenté une nouvelle approche combinant les cartes auto-organisatrices avec le raisonnement à base d'ontologie pour fournir une interprétation automatique des résultats du clustering et permettre le raisonnement sur de grandes quantités de données.

L'idée principale de notre proposition est de s'appuyer sur un processus inductif pour réduire la complexité du raisonnement déductif. Cette combinaison "intelligente" donne une approche hybride capable d'un coté, d'assurer le passage à l'échelle du raisonnement et une réduction importante du temps d'exécution en exploitant les prototypes comme une représentation fiable et résumée des données, ainsi que le processus de propagation de l'étiquetage. Et de l'autre coté, d'inférer l'étiquetage sémantique des cartes

SOM et des données en se basant sur les connaissances expertes sans avoir recours à des données étiquetées.

Nous avons appliqué notre approche sur deux jeux de données et les résultats obtenus sont encourageants. L'application de notre approche pour l'interprétation d'images satellites a montré de bonnes performances sur un problème du monde réel. Un gain important en temps de calcul (tableau 7.1), accompagné d'une conservation de la qualité des résultats (figure 7.4) ont été constatés.

Ces premières expérimentations menées ont permis la validation de notre proposition et la démonstration de son intérêt. Afin d'établir plus en profondeur les différentes facettes de l'approche, d'autres expérimentations peuvent être menées en utilisant d'autres jeux de données avec des connaissances plus riches.

7.5 Valorisation scientifique

Cette contribution a donné lieu à deux publications scientifiques :

— "*Towards Ontology Reasoning for Topological Cluster Labeling*"
International Conference on Neural Information Processing, **ICONIP 2016** - Kyoto, Japan (*Rang : A*)

— "*Intégration des connaissances ontologiques en apprentissage topologique non supervisé*"
Société Francophone de Classification, SFC 2014 - Rabat, Maroc

Chapitre 8

Conclusion et perspectives

La fouille exploratoire de données recouvre de multiples méthodes très efficaces pour partitionner les données et ainsi créer de nouvelles formes de valeurs riches de sens pour les thématiciens. Il n'en demeure pas moins la nécessité d'une implication humaine et experte très chronophage à la fois pour la mise en œuvre des méthodes et techniques d'apprentissage mais aussi pour l'analyse des données.

Le point de départ de la thèse porte à cet effet sur le constat d'une sorte de paradoxe entre le nombre toujours grandissant d'images satellites à traiter et le peu d'experts disponibles à même de mener à bien les travaux d'analyse. Les données s'accumulent et ne sont donc que très partiellement analysées faute d'experts. L'idée principale est donc de proposer des approches qui pourraient permettre à terme d'alléger la tâche des experts et donc de mieux exploiter la mine informationnelle que constituent les images. Le postulat est alors d'envisager les ontologies comme des conteneurs sémantiques à même de faciliter l'extraction de connaissances sans toutefois trop solliciter les experts.

Il s'agit dans cette dernière partie du mémoire de dresser une synthèse sur les contributions apportées. Quelques perspectives, pouvant faire l'objet de travaux ultérieurs, sont également proposées et discutées.

Sommaire

"La victoire sur soi est la plus grande des victoires", **Platon**

8.1 Synthèse des travaux

Différents aspects du processus d'extraction des connaissances ont été explorés. L'idée de départ de nos travaux était d'entrevoir les possibilités offertes par l'utilisation mutuelle des ontologies comme supports de connaissances formalisées et de l'apprentissage automatique pour l'analyse exploratoire des données. L'objectif in fine est la proposition de méthodes originales exploitant ces deux axes et permettant la levée de verrous liés au **fossé sémantique**, à l'**incomplétude des connaissances** et à la **prise en compte des connaissances dans le clustering**.

8.1.1 Contributions

Le premier axe investi a concerné le développement d'une approche hybride qui exploite une ontologie pour réutiliser les connaissances explicitées par raisonnement sous forme de contraintes dans une activité de clustering semi-supervisé. Nous avons ainsi ainsi pu définir méthodologiquement une approche qui combine avec succès raisonnement déductif et apprentissage inductif afin de guider une analyse experte d'une portion de l'image. Une partie de l'expertise est déléguée à l'ontologie qui contient des connaissances formalisées mais incomplètes. Les connaissances formalisées retranscrivent en partie la vision du thématicien et viennent donc guider le clustering et l'orienter en fonction des attendus.

Notre approche permet d'apporter de la modularité à savoir que le contenu de l'ontologie peut être adaptée à la vision du thématicien qui a commandité l'analyse de l'image. Ainsi non seulement les contraintes sont définies automatiquement sans implication préalable du thématicien mais aussi différents jeux de contraintes peuvent être générés pour satisfaire les besoins de l'analyse. Une telle approche va vers une réduction du fossé sémantique. L'expert reste cependant sollicité et garde le contrôle de l'analyse. Il a ainsi à décider du nombre de clusters que le clustering semi-supervisé aura à traiter. Il a également à évaluer et à valider les résultats obtenus et en particulier à analyser les clusters non étiquetés qui correspondent à des classes qui ne sont pas décrites dans l'ontologie de départ. Nous avons donc appliqué notre approche à l'analyse d'images satellites. Il est à noter qu'une telle approche pourrait s'adapter à toute problématique et ne se limite pas à l'imagerie satellite. Les résultats obtenus ont permis la validation expérimentale de notre approche.

En nous appuyant sur une même ontologie élémentaire décrivant les classes thématiques d'eau et de végétation, nous avons analysé plusieurs extraits provenant d'images satellites différentes (scènes portant sur le Brésil et sur la France), nous avons montré le fort potentiel de généricité de notre approche et avons permis la levée d'un verrou important du clustering par contraintes, qui nécessite classiquement la définition de jeux de contraintes pour chacun des jeux de données analysés.

La comparaison de notre méthode avec du clustering non-supervisé et de la classification à base de raisonnement à l'aide de différents indices

de qualité a mis en exergue les bonnes performances de notre approche et confirmé les avantages d'une telle approche mixte.

Notre seconde contribution procède de manière inverse en faisant d'abord appel à du clustering topologique pour alléger le calcul des inférences s'appliquant à l'ontologie. Nous avons ainsi exploré les moyens de réduire le coût du calcul de la tâche de raisonnement appelée reconnaissance des instances de l'ontologie tout en cherchant à automatiser l'étiquetage des clusters obtenus au préalable par analyse exploratoire. À cet effet, nous avons développé une approche originale à base de cartes SOM et de raisonnement à base de logiques de description.

L'idée principale a été d'exploiter dans un premier temps le clustering pour extraire un nombre réduit de données représentées par les prototypes, puis d'utiliser le raisonnement sur ces prototypes, combiné avec un processus de propagation de l'étiquetage sémantique aux instances, afin de permettre une interprétation automatisée des données et d'assurer le passage à l'échelle du raisonnement.

Nous avons mené des expérimentations sur des images satellites et nos résultats ont montré la potentialité de notre approche à fournir une interprétation automatisée des clusters et des données. La comparaison de notre proposition avec le raisonnement classique a fait ressortir sa capacité à permettre le raisonnement sur de très grandes quantités de données. Une deuxième comparaison de notre approche avec le raisonnement séquentiel, opéré par la séparation des instances, a permis une amélioration conséquente du temps du raisonnement sans perte dans la qualité des résultats.

L'étude des méthodes et outils de traitement des images satellites et la compréhension du domaine de la télédétection ont nécessité un fort investissement. Les connaissances acquises ont été capitalisées pour modéliser une ontologie du domaine de la télédétection capable de réduire le fossé sémantique ainsi que la production des jeux de données de référence.

8.2 Perspectives

Nos travaux de thèse sont des travaux exploratoires qui s'emploient à créer des approches innovantes qui se situent entre les domaines de l'ingénierie des connaissances et de l'apprentissage numérique.

Nos propositions ouvrent la voie vers l'utilisation mutuelle et l'exploitation intelligente des ontologies et du clustering et peuvent apporter des éléments de réponse à une prise en charge efficace et fiable du processus d'extraction de connaissances. La mise en place de nos approches et leurs expérimentations nous ont poussé à réfléchir à des perspectives de recherche.

Nous discutons dans ce qui suit de nouvelles contributions possibles qui nous paraissent les plus prometteuses.

8.2.1 Perspectives pour la classification à base d'ontologie et de clustering par contraintes

En ce qui concerne notre première proposition pour la classification à base d'ontologie et de clustering semi-supervisé [CHAHDI et al., 2016c; CHAHDI et al., 2016b; CHAHDI et al., 2016a]. Nous pensons que l'enrichissement de l'ontologie et une méthode de sélection des contraintes constituent des pistes prometteuses.

Dans notre approche, les instances utilisées pour la génération des contraintes sont choisies aléatoirement parmi celles étiquetées par raisonnement. Cette manière de procéder est celle adoptée par la grande majorité des travaux portant sur le clustering par contraintes [DAVIDSON et BASU, 2007]. Cependant, cette sélection aléatoire peut parfois fournir des contraintes inutiles ou répétitives DAVIDSON, WAGSTAFF et BASU (2006). Face à ce constat, nous pensons que le développement d'une technique de sélection de contraintes améliorerait les résultats du clustering. Cela permettrait également d'abaisser le nombre de contraintes introduites dans le clustering et donc d'alléger le coût du calcul.

Une des pistes de sélection des contraintes pourrait se baser sur des travaux liés à l'apprentissage actif [FU, ZHU et LI, 2013]. L'apprentissage actif a été largement étudié pour améliorer les performances de la classification supervisée en sélectionnant parmi les instances étiquetées celle qui permettraient le meilleur apprentissage. Similairement, les techniques utilisées pour sélectionner les exemples d'apprentissage pourraient être adaptées afin de sélectionner les contraintes les plus pertinentes à fournir au clustering dans notre approche.

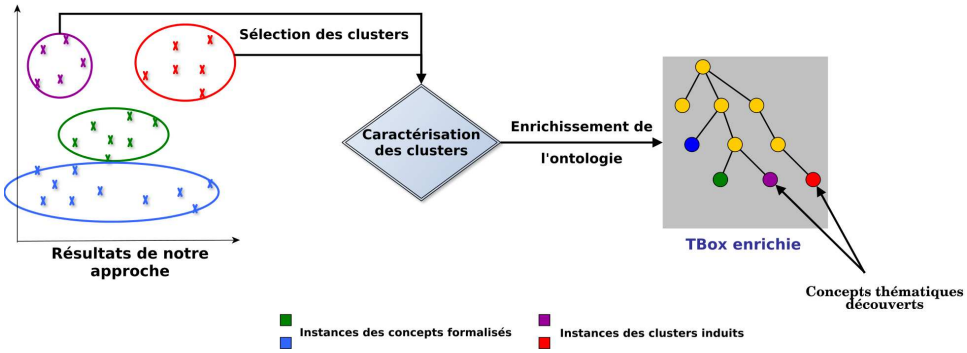


FIGURE 8.1 – Enrichissement des concepts

Une deuxième perspective concerne l'enrichissement de l'ontologie par d'autres concepts thématiques. Il permettrait une plus large adoption de notre approche par les télédetecteurs. Deux pistes de travail peuvent être envisagées afin d'aboutir à cet enrichissement.

La première serait l'implication d'experts en télédétection pour la modélisation de nouveaux concepts thématiques. L'interaction entre un informaticien et un télédétecteur n'est pas une chose aisée. Elle demande une grande implication des deux cotés. D'un coté, l'expert en télédétection doit

être capable d'identifier les indicateurs pouvant discriminer précisément les concepts thématiques qu'il cible. De l'autre coté, l'informaticien doit savoir lever les ambiguïtés sur ces indicateurs et les formaliser en concepts venant peupler l'ontologie.

La deuxième piste pouvant permettre l'enrichissement de l'ontologie porterait sur la caractérisation des clusters non spécifiés par l'expert et découverts par notre approche. En effet, en appliquant notre méthode, nous avons la possibilité d'obtenir de nouveaux clusters par induction. L'un des moyens d'intégrer ses clusters dans l'ontologie serait de développer une approche interactive où l'utilisateur serait à même d'étiqueter les clusters découverts et où les variables discriminantes seraient extraites automatiquement avec des méthodes d'induction de règles. Le développement d'une telle approche faciliterait de manière significative l'acquisition et la formalisation des concepts thématiques.

La figure 8.1 montre l'organisation possible de l'approche enrichie. Les résultats de notre approche en sont le point de départ, et contiennent des instances étiquetées par les concepts de l'ontologie (colorés en bleu et vert) et des instances non-étiquetées affectées aux clusters induits (colorés en rouge et violet). Les clusters induits seront par la suite caractérisés afin d'obtenir des règles de classification. Finalement, ces règles vont servir à la définition des nouveaux concepts thématiques nommés par l'expert.

8.2.2 Perspectives pour l'optimisation du raisonnement par clustering topographique

Dans le cadre de notre deuxième proposition qui traite de l'optimisation du raisonnement et l'étiquetage sémantique des clusters par raisonnement [CHAHDI et al., 2016d ; CHAHDI et al., 2014], une des amélioration possibles réside dans la prise en compte des résultats du raisonnement pour la remise en cause de la distribution des neurones de la carte.

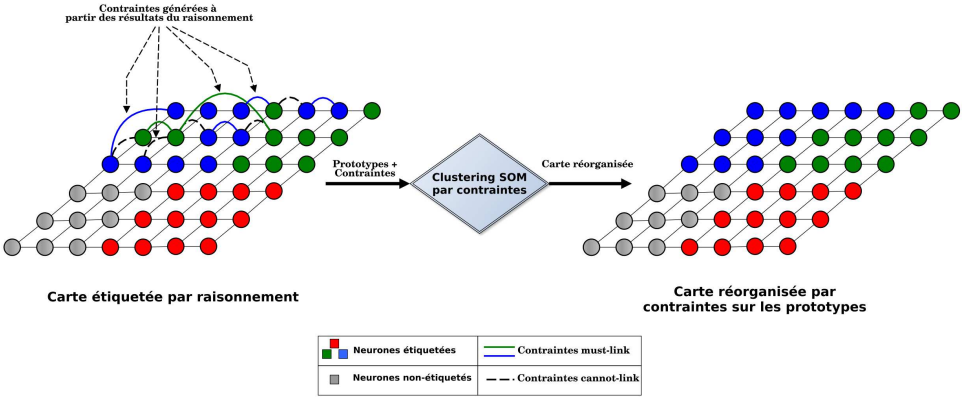


FIGURE 8.2 – Vue générale d'une prise en compte des résultats du raisonnement pour la réorganisation des résultats de SOM

Au delà de l'étiquetage sémantique de la carte et des données, les connaissances expertes fournissent une indication forte sur le partitionnement souhaité par l'expert. Dans ce sens, un moyen efficace d'introduire cette vision pourrait être la mise en place de contraintes au niveau des neurones de la carte résultante.

Ces contraintes ne porteraient pas sur les instances mais sur la distribution des prototypes dans la carte SOM. Elles pourront alors être introduites par une modification de la fonction de voisinage $\mathcal{K}$ de l'algorithme SOM. Nous rappelons que cette fonction intervient lors de la phase de coopération (algorithme 6) afin d'influencer les prototypes des neurones par leurs voisinages.

La figure 8.2 schématise une éventuelle prise en compte des résultats de l'étiquetage pour la réorganisation de la carte. Nous supposons dans cet exemple qu'à l'issue de l'étiquetage de la carte avec notre proposition, des neurones étiquetés se retrouvent non regroupés dans la même région de la carte. En posant des contraintes topographiques sur les neurones bleu et vert. Nous appliquons une deuxième fois le clustering SOM par contraintes au niveau des neurones pour réorganiser la carte et produire une carte respectant l'étiquetage établi par l'ontologie.

8.2.3 Perspectives à long terme

Pour conclure, il nous semble que les approches développées sont à même de faciliter l'analyse et l'interprétation d'images satellites, sans demander de surcharge de travail aux experts. Notre première approche permet la classification des images satellites à partir d'une ontologie du domaine et du clustering semi-supervisé. Notre deuxième contribution propose quant à elle d'exploiter dans un premier temps le clustering SOM pour obtenir une représentation réduite des données d'entrée, puis d'utiliser le raisonnement sur cette représentation réduite, combiné avec un processus de propagation d'étiquetage, pour étiqueter toutes les instances, permettant ainsi une optimisation du raisonnement sur de grande masse de données.

En intégrant nos deux propositions, un système automatique pour l'étiquetage sémantique des images, l'optimisation du raisonnement et l'induction de nouveaux concepts par clustering pour l'enrichissement de l'ontologie pourrait voir le jour.

La figure détaille le déroulement général d'une telle proposition. La première étape consisterait à compresser les données à l'aide du clustering SOM et obtenir un ensemble de prototypes représentatifs. Ces prototypes seront ensuite transformés en instances OWL pour former une ABox réduite dans une deuxième étape. L'ABox construite ainsi que la TBox seront transmises au raisonneur et le service d'inférence réalisation sera appliqué afin d'étiqueter par raisonnement les prototypes répondant aux définitions des concepts thématiques (bleu et vert) ; une fois la carte partiellement étiquetée obtenue. La quatrième étape envisagée est le clustering par contraintes sur les prototypes de la carte. Dans notre figure, nous supposons une application d'un clustering à base de quatre clusters (k = 4).

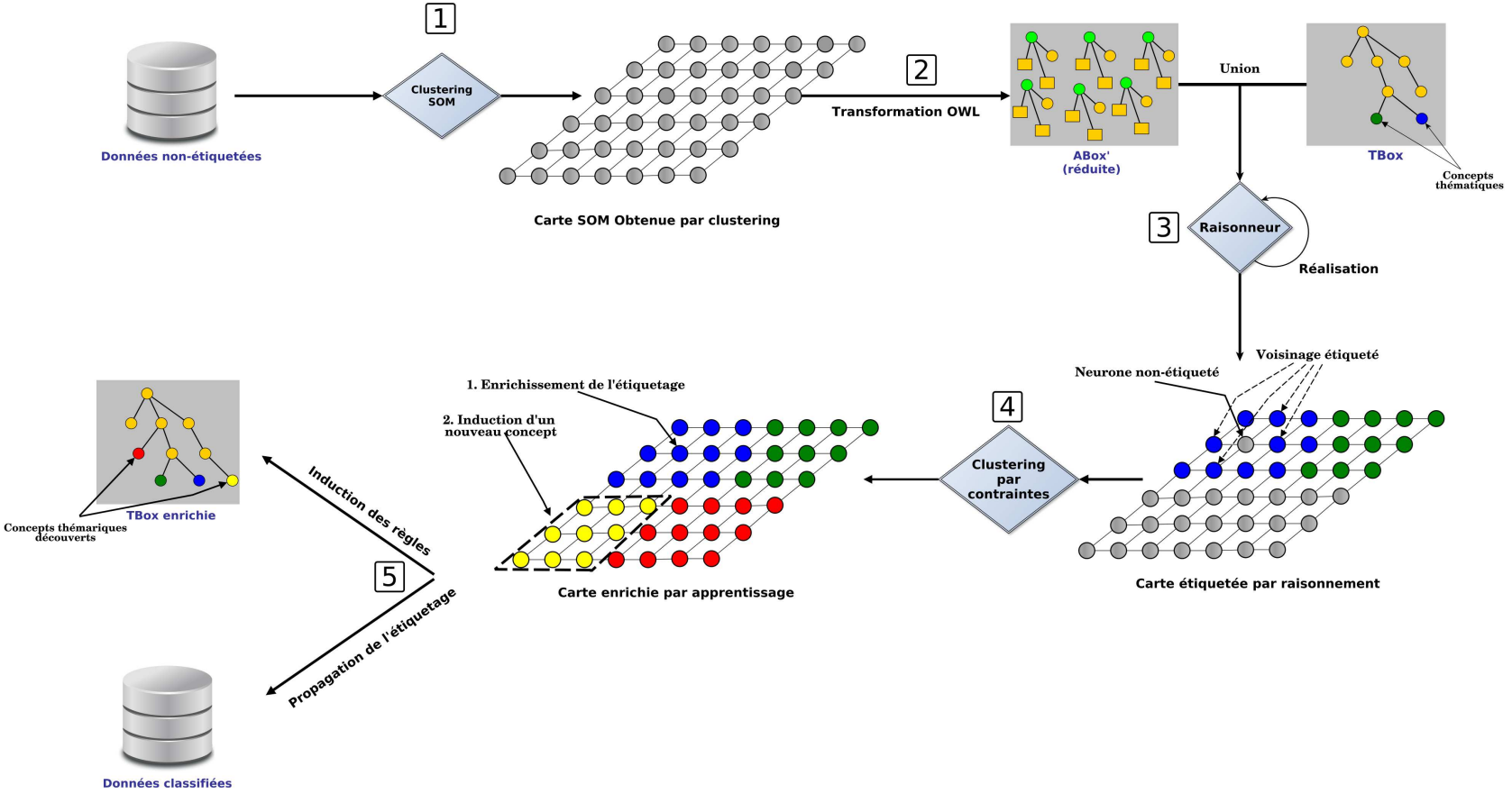


FIGURE 8.3 – Vue générale d'une approche globale pour la classification des données et l'enrichissement des connaissances formalisées

Cette étape permettra la classification des prototypes suivant quatre clusters. Dans notre exemple, deux clusters (bleu et vert) représentent des concepts formalisés dans l'ontologie et deux autres (rouge et jaune) représentent des clusters induits par apprentissage. Finalement, deux processus seront exploités lors d'une cinquième étape pour aboutir à une classification de toutes les données et à un enrichissement de l'ontologie. L'un concernera la propagation de l'étiquetage sur les instances en se basant sur les résultats du SOM. L'autre processus correspondra à la caractérisation des nouveaux clusters et à l'induction des règles pour la formalisation des nouveaux concepts thématiques dans la TBox de l'ontologie.

8.3 Valorisation

Les travaux réalisés et les résultats obtenus ont été partagés avec la communauté scientifique afin de permettre leur diffusion au plus grand nombre. Nous listons ci-dessous les publications effectuées pendant notre thèse :

Publications internationales

— *On the Use of Ontology as a priori Knowledge into Constrained Clustering*,
IEEE International Conference on Data Science and Advanced Analytics, **DSAA 2016** - Montréal, Canada (*Taux d'acceptation : 20.1%*),
Hatim Chahdi, Nistor Grozavu, Isabelle Mougenot, Laure Berti-Equille et Younès Bennani

— *Towards Ontology Reasoning for Topological Cluster Labeling*,
International Conference on Neural Information Processing, **ICONIP 2016** - Kyoto, Japan (*Rang : A*),
Hatim Chahdi, Nistor Grozavu, Isabelle Mougenot, Younès Bennani et Laure Berti-Equille

— *Automated Topological Co-Clustering Using Fuzzy Features Partition*,
International Joint Conference on Neural Networks, **IJCNN 2016** - Vancouver, Canada (*Rang : A*),
Nistor Grozavu, Guenael Cabanes, **Hatim Chahdi** et Nicoleta Rogovshi

— *Application profile for Earth Observation images*,
8^th Metadata and Semantics Research Conference, **MTSR 2014** - Karlsruhe, Allemagne
Jean-Christophe Desconnets, **Hatim Chahdi** et Isabelle Mougenot

— A DCAP to Promote Easy-to-Use Data for Multiresolution and Multitemporal Satellite Imagery Analysis,
International Conference on Dublin Core and Metadata Applications, **DC 2015** - Sao Paulo, Brésil,
Isabelle Mougenot, Jean-Christophe Desconnets et **Hatim Chahdi**

Publications nationales

— Génération de contraintes pour le clustering à partir d'une ontologie - Application à la classification d'images satellites,
Conférence Extraction et Gestion des Connaissances, EGC 2016 - Reims, France (Nominé pour le prix du meilleur article académique),
Hatim Chahdi, Nistor Grozavu, Isabelle Mougenot, Laure Berti-Equille et Younès Bennani

— Approche hybride à base d'ontologie pour le clustering par contraintes,
Conférence internationale francophone sur la science de la données,
AAFD & SFC 2016 - Marrakech, Maroc,
Hatim Chahdi, Nistor Grozavu, Isabelle Mougenot, Laure Berti-Equille et Younès Bennani

— Intégration des connaissances ontologiques en apprentissage topologique non supervisé,
Société Francophone de Classification, SFC 2014 - Rabat, Maroc,
Hatim Chahdi, Nistor Grozavu, Younès Bennani, Isabelle Mougenot et Laure Berti-Equille

Bibliographie

ABBURU, Sunitha (2012). « A Survey on Ontology Reasoners and Comparison ». In : *International Journal of Computer Applications* 57.17.

ALLAB, Kais et Khalid BENABDESLEM (2011). « Constraint Selection for Semi-supervised Topological Clustering ». In : *Machine Learning and Knowledge Discovery in Databases*. Springer, p. 28–43.

ANAND, Saket et al. (2014). « Semi-Supervised Kernel Mean Shift Clustering ». In : *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.6, p. 1201–1215.

ANDRÉS, Samuel (2013). « Ontologies dans les images satellitaires : interpretation sémantique des images ». Thèse de doct. Université Montpellier II - Sciences et Techniques du Languedoc.

ANDRES, Samuel, Damien ARVOR et Christelle 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*. IEEE, p. 825–832.

ATIF, Jamal et al. (2007). « From Generic Knowledge to Specific Reasoning for Medical Image Interpretation Using Graph based Representations. » In : *IJCAI*, p. 224–229.

BAADER, Franz (2003). *The Description Logic Handbook : Theory, Implementation, and Applications*. Cambridge university press.

BAADER, Franz et Philipp HANSCHKE (1991). « A Scheme for Integrating Concrete Domains into Concept Languages ». In : *Proceedings of the 12th International Joint Conference on Artificial Intelligence, IJCAI*. Sydney, New South Wales, Australia, p. 452–457.

BAGHDADI, Nicolas et al. (2015). « The Theia Land Data Centre ». In : *Remote Sensing Data Infrastructures (RSDI) International Workshop*. La grande motte, France.

BANNOUR, Hichem et Céline HUDELOT (2011). « Towards ontologies for image interpretation and annotation ». In : *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*. IEEE, p. 211–216.

— (2012). « Building Semantic Hierarchies Faithful to Image Semantics ». In : *International Conference on Multimedia Modeling*. Springer, p. 4–15.

— (2014). « Building and Using Fuzzy Multimedia Ontologies for Semantic Image Annotation ». In : *Multimedia Tools and Applications* 72.3, p. 2107–2141.

BARALDI, Andrea et Palma BLONDA (1999). « A Survey of Fuzzy Clustering Algorithms for Pattern Recognition. I ». In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29.6, p. 778–785.

BARALDI, Andrea et al. (2006). « Automatic Spectral Rule-based Preliminary Mapping of Calibrated Landsat TM and ETM+ Images ». In : *Geoscience and Remote Sensing, IEEE Transactions on* 44.9, p. 2563–2586.

BASU, Sugato, Arindam BANERJEE et Raymond MOONEY (2002). « Semi-Supervised Clustering by Seeding ». In : *Proceedings of 19th International Conference on Machine Learning, ICML*. Citeseer.

BASU, Sugato, Arindam BANERJEE et Raymond J MOONEY (2004). « Active Semi-Supervision for Pairwise Constrained Clustering ». In : *Proceedings of the 2004 SIAM International Conference on Data Mining*. T. 4. Society for Industrial et Applied Mathematics. Lake bueno vista, FL, USA, p. 333–344.

BASU, Sugato, Mikhail BILENKO et Raymond J MOONEY (2004). « A Probabilistic Framework for Semi-Supervised Clustering ». In : *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, p. 59–68.

BASU, Sugato, Ian DAVIDSON et Kiri WAGSTAFF (2008). *Constrained Clustering : Advances in Algorithms, Theory, and Applications*. CRC Press.

BECKETT, Dave et Brian MCBRIDE (2004). « RDF/XML Syntax Specification (Revised) ». In : *W3C recommendation* 10.

BELLAL, Fazia, Khalid BENABDESLEM et Alexandre AUSSEM (2008). « SOM Based Clustering with Instance-Level Constraints ». In : *Proceedings of the 16th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN*. Bruges, Belgium, p. 313–318.

BELLET, Aurélien, Amaury HABRARD et Marc SEBBAN (2013). « A Survey on Metric Learning for Feature Vectors and Structured Data ». In : *arXiv preprint, arXiv :1306.6709*.

BENNETT, Kristin, Ayhan DEMIRIZ et al. (1999). « Semi-Supervised Support Vector Machines ». In : *Advances in Neural Information processing systems*, p. 368–374.

BERKHIN, Pavel (2006). « A Survey of Clustering Data Mining Techniques ». In : *Grouping multidimensional data*. Springer, p. 25–71.

BERNERS-LEE, Tim, Roy FIELDING et Larry MASINTER (2005). « RFC 3986 : Uniform Resource Identifier (URI) : Generic Syntax ». In : *The Internet Society*.

BERNERS-LEE, Tim, James HENDLER, Ora LASSILA et al. (2001). « The Semantic Web ». In : *Scientific American* 284.5, p. 28–37.

BILENKO, Mikhail, Sugato BASU et Raymond J MOONEY (2004). « Integrating Constraints and Metric Learning in Semi-Supervised Clustering ». In : *Proceedings of the 21th International Conference on Machine learning, ICML*. ACM. San Francisco, CA, USA, p. 11.

BIRON, Paul, Ashok MALHOTRA et al. (2004). « XML Schema Part 2 : Datatypes ». In : *World Wide Web Consortium Recommendation REC-xmlschema-2-20041028*.

BLASCHKE, Thomas, Stefan LANG et Geoffrey HAY (2008). *Object-Based Image Analysis : Spatial Concepts for Knowledge-Driven Remote Sensing Applications*. Springer Science & Business Media.

BLOCH, Isabelle (2005). « Fuzzy Spatial Relationships for Image Processing and Interpretation : a Review ». In : *Image and Vision Computing* 23.2, p. 89–110.

BLOEHDORN, Stephan et al. (2005). « An Ontology-Based Framework for Text Mining ». In : 20.1, p. 87–112.

BOICU, Mihai et al. (2001). « Ontologies and the Knowledge Acquisition Bottleneck ». In : *Proceedings of the 17th International Joint Conference on Artificial Intelligence, IJCAI*. Seattle, WA, USA.

BOSER, Bernhard E, Isabelle M GUYON et Vladimir N VAPNIK (1992). « A Training Algorithm for Optimal Margin Classifiers ». In : *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. ACM, p. 144–152.

BRACHMAN, Ronald J et Hector J LEVESQUE (1984). « The Tractability of Subsumption in Frame-based Description Languages ». In : *Proceedings of the 4th National Conference on Artificial Intelligence, AAAI*. T. 84. Austin, TX, USA : AAAI Press, p. 34–37.

BRAY, Tim, Dave HOLLANDER et Andrew LAYMAN (1999). *Namespaces in XML*.

BRAY, Tim et al. (1998). « Extensible Markup Language (XML) ». In : *World Wide Web Consortium Recommendation REC-xml-19980210* 16, p. 16.

BRAY, Tim et al. (2008). *Extensible Markup Language (XML) 1.0*.

BREIMAN, Leo et al. (1984). *Classification and Regression Trees*. CRC press.

BRICKLEY, Dan et R.V. GUHA (2004). *RDF Vocabulary Description Language 1.0 : RDF Schema*. W3C Recommendation. W3C.

CALOZ, Régis et Claude COLLET (2001). *Précis de télédétection : traitements numériques d'images de télédétection*. T. 3. Presses de l'Université du Québec.

CHAHDI, Hatim et al. (2014). « Intégration des connaissances ontologiques en apprentissage topologique non-supervisé ». In : *Actes de la 21ème rencontre de la société francophone de classification, SFC*. Marrakech, Maroc.

CHAHDI, Hatim et al. (2016a). « Approche hybride à base d'ontologie pour le clustering par contraintes ». In : *Actes de la conférence internationale francophone sur la science de la données, AAFD & SFC*. Marrakech, Maroc.

— (2016b). « Génération de contraintes pour le clustering à partir d'une ontologie - Application à la classification d'images satellites ». In : *Extraction et Gestion des Connaissances, EGC*. Reims, France.

— (2016c). « On the Use of Ontology as A Priori Knowledge into Constrained Clustering ». In : *Proceedings of the IEEE 3rd International Conference on Data Science and Advanced Analytics, DSAA*. Montreal, Canada.

CHAHDI, Hatim et al. (2016d). « Towards Ontology Reasoning for Topological Cluster Labeling ». In : *Proccedings of The 23rd International Conference on Neural Information Processing*. Kyoto, Japan.

CHANDER, Gyanesh et Brian MARKHAM (2003). « Revised Landsat-5 TM Radiometric Calibration Procedures and Postcalibration Dynamic Ranges ». In : *IEEE Transactions on Geoscience and Remote Sensing* 41.11, p. 2674–2677.

CHANDER, Gyanesh, Brian L MARKHAM et Julia A BARSI (2007). « Revised Landsat-5 Thematic Mapper Radiometric Calibration ». In : *IEEE Geoscience and Remote Sensing Letters* 4.3, p. 490–494.

CHANDER, Gyanesh, Brian L MARKHAM et Dennis L HELDER (2009). « Summary of Current Radiometric Calibration Coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI Sensors ». In : *Remote Sensing of Environment* 113.5, p. 893–903.

CHAPELLE, Olivier, Bernhard SCHÖLKOPF et Alexander ZIEN, éds. (2006). *Semi-Supervised Learning*. Cambridge, MA : MIT Press.

CONSORTIUM, World Wide Web (2006). *Namespaces in XML 1.1*.

CORNUÉJOLS, Antoine et Laurent MICLET (2011). *Apprentissage artificiel : concepts et algorithmes*. Editions Eyrolles.

CORTES, Corinna et Vladimir VAPNIK (1995). « Support-Vector Networks ». In : *Machine learning* 20.3, p. 273–297.

COVOES, Thiago F, Eduardo R HRUSCHKA et Joydeep GHOSH (2013). « A Study of K-Means-Based Algorithms for Constrained Clustering ». In : *Intelligent Data Analysis* 17.3, p. 485–505.

CRISTIANINI, Nello, John SHAWE-TAYLOR et Jaz S KANDOLA (2001). « Spectral Kernel Methods for Clustering ». In : *Advances in Neural Information Processing Systems*, p. 649–655.

CULLEN, J et A BRYMAN (1988). « The Knowledge Acquisition Bottleneck : Time for Reassessment ? » In : *Expert Systems* 5.3, p. 216–225.

DAVIDSON, Ian et Sugato BASU (2007). « A Survey of Clustering with Instance Level Constraints ». In : *ACM Transactions on Knowledge Discovery from Data*, p. 1–41.

DAVIDSON, Ian et SS RAVI (2005a). « Agglomerative Hierarchical Clustering with Constraints : Theoretical and Empirical Results ». In : *Knowledge Discovery in Databases : PKDD 2005*. Springer, p. 59–70.

— (2005b). « Towards Efficient and Improved Hierarchical Clustering with Instance and Cluster Level Constraints ». In : *State University of New York, Albany, Tech. Rep*.

DAVIDSON, Ian, Kiri WAGSTAFF et Sugato BASU (2006). « Measuring Constraint-Set Utility for Partitional Clustering Algorithms ». In : *Proceedings of the 10th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD*. Berlin, Germany.

DAVIES, David L et Donald W BOULDIN (1979). « A Cluster Separation Measure ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2, p. 224–227.

DEMPSTER, Arthur P, Nan M LAIRD et Donald B RUBIN (1977). « Maximum Likelihood from Incomplete Data Via the EM Algorithm ». In : *Journal of the Royal Statistical Society. Series B, Methodological*, p. 1–38.

DERIVAUX, Sébastien et al. (2006). « Watershed Segmentation of Remotely Sensed Images Based on a Supervised Fuzzy Pixel Classification ». In : *Proceedings of the IEEE International Geosciences And Remote Sensing Symposium, IGARSS*. Denver, Colorado, USA, p. 3712–3715.

DESCONNETS, Jean-Christophe, Hatim CHAHDI et Mougenot MOUGENOT (2014). « Application Profile for Earth Observation Images ». In : *Proceedings of the 8th Metadata and Semantics Research Conference, MTSR*. Karlsruhe, Germany, p. 68–82.

DIRECTIVE, INSPIRE (2007). « 2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) ». In : *Official Journal of European Union*.

DURAND, Nicolas et al. (2007). « Ontology-Based Object Recognition for Remote Sensing Image Interpretation ». In : *19th IEEE International Conference on Tools with Artificial Intelligence, ICTAI*. T. 1. IEEE, p. 472–479.

DÉVELOPPEMENT DURABLE, Université Virtuelle Environnement et (2008). *Éléments de physique du rayonnement*.

EITER, Thomas et al. (2014). « Using Openstreetmap Data to Create Benchmarks for Description Logic Reasoners ». In : *Informal Proceedings of the 3rd International Workshop on OWL Reasoner Evaluation, ORE*.

ESTER, Martin (2009). « Density-Based Clustering ». In : *Encyclopedia of Database Systems*. Springer, p. 795–799.

ESTER, Martin et al. (1996). « A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise ». In : *The 2nd International Conference on Knowledge Discovery and Data Mining, KDD*. T. 96. 34. Portland, OR, USA, p. 226–231.

EYBEN, Florian, Sabine BUCHHOLZ et Norbert BRAUNSCHWEILER (2012). « Unsupervised Clustering of Emotion and Voice Styles for Expressive TTS ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, p. 4009–4012.

FALOMIR, Zoe et Ana-Maria OLTEŢEANU (2015). « Logics Based on Qualitative Descriptors for Scene Understanding ». In : *Neurocomputing* 161, p. 3–16.

FALOMIR, Zoe et al. (2011). « Describing Images Using Qualitative Models and Description Logics ». In : *Spatial Cognition & Computation* 11.1, p. 45–74.

FELZENSZWALB, Pedro F et Daniel P HUTTENLOCHER (2004). « Efficient Graph-Based Image Segmentation ». In : *International Journal of Computer Vision* 59.2, p. 167–181.

FILIPPONE, Maurizio et al. (2008). « A Survey of Kernel and Spectral Methods for Clustering ». In : *Pattern Recognition* 41.1, p. 176–190.

FISHER, Douglas H (1987). « Knowledge Acquisition via Incremental Conceptual Clustering ». In : *Machine learning* 2.2, p. 139–172.

FORESTIER, Germain et al. (2012). « Knowledge-Based Region Labeling for Remote Sensing Image Interpretation ». In : *Computers, Environment and Urban Systems* 36.5, p. 470–480.

FU, Yifan, Xingquan ZHU et Bin LI (2013). « A Survey on Instance Selection for Active Learning ». In : *Knowledge and Information Systems*, p. 1–35.

GARDINER, Tom, Ian HORROCKS et Dmitry TSARKOV (2006). *Automated Benchmarking of Description Logic Reasoners*.

GÓMEZ-ROMERO, Juan et al. (2011). « Ontology-Based Context Representation and Reasoning for Object Tracking and Scene Interpretation in Video ». In : *Expert Systems with Applications* 38.6, p. 7494–7510.

GÓRRIZ, Juan Manuel et al. (2006). « Hard C-Means Clustering for Voice Activity Detection ». In : *Speech Communication* 48.12, p. 1638–1649.

GRANDVALET, Yves et Yoshua BENGIO (2004). « Semi-Supervised Learning by Entropy Minimization ». In : *Advances in Neural Information Processing Systems*, p. 529–536.

GRANELL, Carlos et al. (2009). « Handbook of Research on Geoinformatics ». In : sous la dir. d'University of PITTSBURGH. Idea Group Publishing. Chap. V Spatial Data Infrastructures, p. 36–41.

GRIRA, Nizar, Michel CRUCIANU et Nozha BOUJEMAA (2004). « Unsupervised and Semi-Supervised Clustering : A Brief Survey ». In : *A Review of Machine Learning Techniques for Processing Multimedia Content* 1, p. 9–16.

GROUP, W3C OWL Working et OTHERS (2009). « OWL 2 Web Ontology Language Document Overview ». In : *W3C recommendation*.

GRUBER, Thomas R (1993). « A Translation Approach to Portable Ontology Specifications ». In : *Knowledge Acquisition* 5.2, p. 199–220.

GUHA, Sudipto, Rajeev RASTOGI et Kyuseok SHIM (1998). « CURE : an Efficient Clustering Algorithm for Large Databases ». In : 27.2, p. 73–84.

— (1999). « ROCK : A Robust Clustering Algorithm for Categorical Attributes ». In : *Proceedings of the 15th International Conference on Data Engineering*. IEEE, p. 512–521.

HALKIDI, Maria, Yannis BATISTAKIS et Michalis VAZIRGIANNIS (2001). « On Clustering Validation Techniques ». In : *Journal of Intelligent Information Systems* 17.2-3, p. 107–145.

HAN, Jiawei, Jian PEI et Micheline KAMBER (2011). *Data mining : concepts and techniques*. Elsevier.

HARE, Jonathon S et al. (2006). « Mind the Gap : Another Look at the Problem of the Semantic Gap in Image Retrieval ». In : *Electronic Imaging 2006*. International Society for Optics et Photonics, p. 607309–607309.

HARNAD, Stevan (1990). « The Symbol Grounding Problem ». In : *Physica D : Nonlinear Phenomena* 42.1-3, p. 335–346.

HAY, Geoffrey J et G CASTILLA (2008). « Geographic Object-Based Image Analysis (GEOBIA) : A New Name for a New Discipline ». In : *Object-Based Image Analysis*. Springer, p. 75–89.

HAYES, Patrick J (1979). « The Logic of Frames ». In : *Frame Conceptions and Text Understanding* 46, p. 61.

HORROCKS, Ian et al. (2002). « DAML+OIL : A Description Logic for the Semantic Web ». In : *IEEE Data Eng. Bull.* 25.1, p. 4–9.

HORROCKS, Ian et al. (2004). « The Instance Store : DL Reasoning with Large Numbers of Individuals ». In : *Proceedings of the 2004 Description Logic Workshop, DL*. British Columbia, Canada, p. 31–40.

HOTHO, Andreas, Alexander MAEDCHE et Steffen STAAB (2002). « Ontology-Based Text Document Clustering ». In : *KI* 16.4, p. 48–54.

HOTHO, Andreas, Steffen STAAB et Gerd STUMME (2003). « Ontologies Improve Text Document Clustering ». In : *Proceedings of the 3rd IEEE International Conference on Data Mining, ICDM*. IEEE. Melbourne, Florida, USA, p. 541–544.

HU, Xiaohua et al. (2009). « Exploiting Wikipedia as External Knowledge for Document Clustering ». In : *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. Paris, France, p. 389–396.

HUDELOT, Céline, Jamal ATIF et Isabelle BLOCH (2008). « Fuzzy Spatial Relation Ontology for Image Interpretation ». In : *Fuzzy Sets and Systems* 159.15, p. 1929–1951.

IMAGE, SPOT (1988). « SPOT User's Handbook ». In : *Centre National d'Etude Spatiale (CNES) and SPOT Image* 1, p. 3.

ISO 19115 - Geographic Information Metadata (2003). ISO. Geneva, Switzerland : International Organization for Standardization.

JAIN, Anil K et Patrick J FLYNN (1996). *Image Segmentation Using Clustering*. IEEE Press, Piscataway, NJ.

JAIN, Anil K, M Narasimha MURTY et Patrick J FLYNN (1999). « Data Clustering : A Review ». In : *ACM Computing Surveys, CSUR* 31.3, p. 264–323.

JAIN, Anil K et al. (2004). « Landscape of Clustering Algorithms ». In : *Proceedings of the 17th International Conference on Pattern Recognition, ICPR*. T. 1. IEEE, p. 260–263.

JENSEN, John R (1986). *Introductory Digital Image Processing : a Remote Sensing Perspective*. Rapp. tech. Univ. of South Carolina, Columbus.

JING, Liping et al. (2006). « Ontology-Based Distance Measure for Text Clustering ». In : *Proceedings of SIAM SDM workshop on text mining*. Bethesda, Maryland, USA.

JOACHIMS, Thorsten (1999). « Transductive Inference for Text Classification Using Support Vector Machines ». In : *Proceedings of 16th International Conference on Machine Learning, ICML*. T. 99. San Francisco, CA, USA, p. 200–209.

JOHNSON, Stephen C (1967). « Hierarchical Clustering Schemes ». In : *Psychometrika* 32.3, p. 241–254.

JORDAN, Carl F (1969). « Derivation of Leaf-Area Index from Quality of Light on the Forest Floor ». In : *Ecology* 50.4, p. 663–666.

KARYPIS, George, Eui-Hong HAN et Vipin KUMAR (1999). « Chameleon : Hierarchical Clustering using Dynamic Modeling ». In : *Computer* 32.8, p. 68–75.

KAZMIERSKI, Mathieu et al. (2014). « GEOSUD SDI : Accessing Earth Observation Data Collections with Semantic-Based Services ». In : *Proceedings of the 17th AGILE Conference on Geographic Information Science, Connecting a Digital Europe through Location and Place*. Castellon, Spain.

KENT, Allen et al. (1955). « Machine Literature Searching VIII. Operational Criteria for Designing Information Retrieval Systems ». In : *American Documentation* 6.2, p. 93–101.

KERGOMARD, Claude (1990). « La télédétection aérospatiale : une introduction ». In : *Cours de télédétection, Ecole Normale Supérieure Paris*.

KHOREVA, Anna et al. (2014). « Learning Must-Link Constraints for Video Segmentation Based on Spectral Clustering ». In : *Pattern Recognition*. Springer, p. 701–712.

KIFER, Michael, Georg LAUSEN et James WU (1995). « Logical Foundations of Object-Oriented and Frame-Based Languages ». In : *Journal of the ACM (JACM)* 42.4, p. 741–843.

KOHONEN, Teuvo (1990). « The Self-Organizing Map ». In : *Proceedings of the IEEE* 78.9, p. 1464–1480.

— (2013). « Essentials of the Self-Organizing Map ». In : *Neural Networks* 37, p. 52–65.

KOZA, John R (1992). *Genetic Programming : On the Programming of Computers by Means of Natural Selection*. T. 1. MIT press.

KRÖTZSCH, Markus, Frantisek SIMANCIK et Ian HORROCKS (2012). « A Description Logic Primer ». In : *arXiv Preprint arXiv :1201.4089*.

KULIS, Brian (2012). « Metric learning : A Survey ». In : *Foundations and Trends in Machine Learning* 5.4, p. 287–364.

KULIS, Brian et al. (2009). « Semi-Supervised Graph Clustering : A Kernel Approach ». In : *Machine Learning* 74.1, p. 1–22.

LIU, Ying et al. (2007). « A Survey of Content-Based Image Retrieval with High-Level Semantics ». In : *Pattern Recognition* 40.1, p. 262–282.

LUTZ, Carsten (2002). « The Complexity of Description Logics with Concrete Domains ». Thèse de doct. Bibliothek der RWTH Aachen.

— (2003). « Description Logics with Concrete Domains–A Survey ». In :

MACQUEEN, James et al. (1967). « Some Methods for Classification and Analysis of Multivariate Observations ». In : *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. T. 1. 14. Oakland, CA, USA, p. 281–297.

MAILLOT, Nicolas, Monique THONNAT et Alain BOUCHER (2004). « Towards Ontology-Based Cognitive Vision ». In : *Machine Vision and Applications* 16.1, p. 33–40.

MAILLOT, Nicolas Eric et Monique THONNAT (2008). « Ontology Based Complex Object Recognition ». In : *Image and Vision Computing* 26.1, p. 102–113.

MANOLA, Franck et Eric MILLER (2004). *RDF Primer*. W3C Recommendation. http://www.w3.org/TR/rdf-primer/. World Wide Web Consortium.

MATENTZOGLU, Nicolas et al. (2015). « A Survey of Current, Stand-Alone OWL Reasoners ». In : *Informal Proceedings of the 4th International Workshop on OWL Reasoner Evaluation*. T. 1387.

MAULIK, Ujjwal et Sanghamitra BANDYOPADHYAY (2002). « Performance Evaluation of some Clustering Algorithms and Validity Indices ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.12, p. 1650–1654.

MCCLOY, Keith R et Thomas SEVERIENS (2012). *Introduction to Categorisation of Objects from Their Data*. URL : http://www.seos-project.eu.

MCLACHLAN, Geoffrey J et Kaye E BASFORD (1988). « Mixture Models. Inference and Applications to Clustering ». In : *Statistics : Textbooks and Monographs* 1.

MILLER, George A (1995). « WordNet : A Lexical Database for English ». In : *Communications of the ACM* 38.11, p. 39–41.

MINSKY, Marvin (1975). *A Framework for Representing Knowledge*.

MOTIK, Boris et al. (2009). « Owl 2 Web Ontology Language : Profiles ». In : *W3C recommendation* 27, p. 61.

MOUGENOT, Isabelle, Jean-Christophe DESCONNETS et Hatim CHAHDI (2015). « A DCAP to Promote Easy-to-Use Data for Multiresolution and Multitemporal Satellite Imagery Analysis ». In : *Proceedings of the International Conference on Dublin Core and Metadata Applications, DC*. Sao Paulo, Brazil, p. 10–19.

MUGNIER, Marie-Laure et Michel CHEIN (1992). « Conceptual Graphs : Fundamental Notions ». In : *Revue d'intelligence artificielle* 6.4, p. 365–406.

MURTAGH, Fionn et Pedro CONTRERAS (2012). « Algorithms for Hierarchical Clustering : An Overview ». In : *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery* 2.1, p. 86–97.

NARDI, Daniele et Ronald J BRACHMAN (2003). « An Introduction to Description Logics ». In : *Description Logic Handbook*, p. 1–40.

NEUMANN, Bernd et Ralf MÖLLER (2008). « On Scene Interpretation with Description Logics ». In : *Image and Vision Computing* 26.1, p. 82–101.

NG, Andrew Y, Michael I JORDAN, Yair WEISS et al. (2002). « On Spectral Clustering : Analysis and an Algorithm ». In : *Advances in Neural Information Processing Systems* 2, p. 849–856.

PELLEG, Dan et Dorit BARAS (2007). « K-Means with Large and Noisy Constraint Sets ». In : *Proceedings of the 18th European Conference on Machine Learning, ECML*. Springer. Warsaw, Poland, p. 674–682.

PENA, José Manuel, Jose Antonio LOZANO et Pedro LARRANAGA (1999). « An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm ». In : *Pattern Recognition Letters* 20.10, p. 1027–1040.

PRIÉ, Yannick et Serge GARLATTI (2004). « Métadonnées et annotations dans le Web sémantique ». In : *Revue I3 Information-Interaction-Intelligence, Numéro Hors-série Web sémantique, Edition Cépaduès* 24.6.

PUISSANT, Anne et al. (2006). « Amélioration des connaissances sur l'environnement urbain : intérêt de l'intégration de règles dans les procédures de classification ». In : *Interactions Nature-Société, analyse et modèles*, p. 3–6.

QUILLIAN, M Ross (1968). *Semantic Memory*. MIT press.

REITER, Raymond (1980). « A Logic for Default Reasoning ». In : *Artificial Intelligence* 13.1, p. 81–132.

RENDÓN, Eréndira et al. (2011a). « A Comparison of Internal and External Cluster Validation Indexes ». In : *Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*. T. 29. San Francisco, CA, USA.

RENDÓN, Eréndira et al. (2011b). « Internal Versus External Cluster Validation Indexes ». In : *International Journal of computers and communications* 5.1, p. 27–34.

RICHARDS, John A et JA RICHARDS (1999). *Remote Sensing Digital Image Analysis*. T. 3. Springer.

RIJSBERGEN, CJ van (1979). « Information Retrieval ». In : *The Information Retrieval Group*.

ROSSE, Cornelius, José LV MEJINO JR et al. (2003). « A Reference Ontology for Biomedical Informatics : The Foundational Model of Anatomy ». In : *Journal of Biomedical Informatics* 36.6, p. 478–500.

ROUSE JR, J_W et al. (1974). « Monitoring Vegetation Systems in the Great Plains with ERTS ». In : *NASA Special Publication* 351, p. 309.

ROUSSEEUW, Peter J (1987). « Silhouettes : a Graphical Aid to the Interpretation and Validation of Cluster Analysis ». In : *Journal of Computational and Applied Mathematics* 20, p. 53–65.

SCHMIDT-SCHAUSS, Manfred et Gert SMOLKA (1991). « Attributive Concept Descriptions with Complements ». In : *Artificial Intelligence* 48.1, p. 1–26.

SCHOBER, Jean-Pierre, Thorsten HERMES et Otthein HERZOG (2004). « Content-Based Image Retrieval by Ontology-Based Object Recognition ». In : *Proceedings of the Workshop on Applications of Description Logics*. Germany.

SEDDING, Julian et Dimitar KAZAKOV (2004). « WordNet-based Text Document Clustering ». In : *Proceedings of the 3rd Workshop on Robust Methods in Analysis of Natural Language Data*. Association for Computational Linguistics. Geneva, Switzerland, p. 104–113.

SEEGER, Matthias (2000). *Learning with Labeled and Unlabeled Data*. Rapp. tech.

SELIM, Shokri Z et Mohamed A ISMAIL (1984). « K-Means-Type Algorithms : a Generalized Convergence Theorem and Characterization of Local Optimality ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1, p. 81–87.

SESTER, Monika (2000). « Knowledge Acquisition for the Automatic Interpretation of Spatial Data ». In : *International Journal of Geographical Information Science* 14.1, p. 1–24.

SHADBOLT, Nigel, Tim BERNERS-LEE et Wendy HALL (2006). « The Semantic Web Revisited ». In : *IEEE Intelligent Systems* 21.3, p. 96–101. ISSN : 1541-1672. DOI : 10.1109/MIS.2006.62.

SHEEREN, D et al. (2006). « Discovering Rules with Genetic Algorithms to Classify Urban Remotely Sensed Data ». In : *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IGARSS*. Denver, Colorado, USA, p. 3919–3922.

SHENTAL, Noam et al. (2004). « Computing Gaussian Mixture Models with EM using Equivalence Constraints ». In : *Advances in Neural Information Processing Systems* 16.8, p. 465–472.

SIRIN, Evren et al. (2007). « Pellet : A Practical OWL-DL Reasoner ». In : *Web Semantics : Science, Services and Agents on the World Wide Web* 5.2, p. 51–53.

SMEULDERS, Arnold WM et al. (2000). « Content-Based Image Retrieval at the End of the Early Years ». In : *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.12, p. 1349–1380.

SOWA, John F (1983). *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley Pub., Reading, MA.

TERRE CCRS, Centre canadien de cartographie et d'observation de la (2013). *Notions fondamentales de Télédétection*.

TSARKOV, Dmitry et Ian HORROCKS (2006). « FaCT++ Description Logic Reasoner : System Description ». In : *Proceedings of the 3rd International Joint Conference on Automated Reasoning*. Springer. Seattle, WA, USA, p. 292–297.

TUCKER, Compton J (1979). « Red and Photographic Infrared Linear Combinations for Monitoring Vegetation ». In : *Remote Sensing of Environment* 8.2, p. 127–150.

VAPNIK, V et A STERIN (1977). « On Structural Risk Minimization or Overall Risk in a Problem of Pattern Recognition ». In : *Automation and Remote Control* 10.3, p. 1495–1503.

VAPNIK, Vladimir (1995). *The Nature of Statistical Learning Theory*.

VAPNIK, Vladimir N et Alexey J CHERVONENKIS (1974). *Theory of Pattern Recognition*. Nauka.

VESANTO, Juha (1999). « SOM-based Data Visualization Methods ». In : *Intelligent data analysis* 3.2, p. 111–126.

VON LUXBURG, Ulrike (2007). « A Tutorial on Spectral Clustering ». In : *Statistics and Computing* 17.4, p. 395–416.

WAGSTAFF, Kiri et Claire CARDIE (2000). « Clustering With Instance-Level Constraints ». In : t. 1097. Austin, Texas, USA.

WAGSTAFF, Kiri et al. (2001). « Constrained K-Means Clustering With Background Knowledge ». In : *Proceedings of the 18th International Conference on Machine Learning, ICML*. T. 1, p. 577–584.

WANG, Xiang, Buyue QIAN et Ian DAVIDSON (2014). « On Constrained Spectral Clustering and its Applications ». In : *Data Mining and Knowledge Discovery* 28.1, p. 1–30.

WARD JR, Joe H (1963). « Hierarchical Grouping to Optimize an Objective Function ». In : *Journal of the American Statistical Association* 58.301, p. 236–244.

WEDEL, Michel et Wagner A KAMAKURA (2012). *Market Segmentation : Conceptual and Methodological Foundations*. T. 8. Springer Science & Business Media.

WILLE, Rudolf (1997). « Conceptual Graphs and Formal Concept Analysis ». In : *Proceeding of the 12th International Conference on Conceptual Structures*. Springer. Seattle, WA, USA, p. 290–303.

WU, Baoyuan et al. (2013). « Constrained Clustering and its Application to Face Clustering in Videos ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA, p. 3507–3514.

XING, Eric P et al. (2003). « Distance Metric Learning with Application to Clustering with Side-Information ». In : *Advances in Neural Information Processing Systems* 15, p. 505–512.

XU, Rui et Donald WUNSCH (2005). « Survey of Clustering Algorithms ». In : *IEEE Transactions on Neural Networks* 16.3, p. 645–678.

YANG, Liu et Rong JIN (2006). « Distance Metric Learning : A Comprehensive Survey ». In : *Michigan State Universiy* 2, p. 78.

YANG, Miin-Shen (1993). « A Survey of Fuzzy Clustering ». In : *Mathematical and Computer Modelling* 18.11, p. 1–16.

ZHU, Xiaojin (2005). *Semi-Supervised Learning Literature Survey*. Rapp. tech. 1530. Computer Sciences, University of Wisconsin-Madison.

ZHU, Xiaojin, Zoubin GHAHRAMANI, John LAFFERTY et al. (2003). « Semi-Supervised Learning using Gaussian Fields and Harmonic Functions ». In : *Proceedings of the 20th International Conference on Machine Learning, ICML*. T. 3. Washington, DC, USA, p. 912–919.

ZHU, Zhe et Curtis E WOODCOCK (2012). « Object-based Cloud and Cloud Shadow Detection in Landsat Imagery ». In : *Remote Sensing of Environment* 118, p. 83–94.