On the Use of Ontology as a priori Knowledge into Constrained Clustering

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

On the Use of Ontology as a priori Knowledge into Constrained Clustering

Hatim Chahdi*†, Nistor Grozavu†, Isabelle Mougenot*, Laure Berti-Equille*‡ and Younès Bennani†
*UMR 228 Espace-Dev , IRD - Université de Montpellier
Maison de la Télédétection, 500 Rue J.F. Breton, 34093 Montpellier, FRANCE
Email: *firstname.lastname*@ird.fr
†UMR 7030 LIPN, CNRS - Université Paris 13
99, avenue J.-B. Clément, 93430 Villetaneuse, FRANCE
Email: *firstname.lastname*@lipn.univ-paris13.fr
‡Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, QATAR
Doha, Qatar

***Abstract*—Recent studies have shown that the use of a priori knowledge can significantly improve the results of unsupervised classification. However, capturing and formatting such knowledge as constraints is not only very expensive requiring the sustained involvement of an expert but it is also very difficult because some valuable information can be lost when it cannot be encoded as constraints. In this paper, we propose a new constraint-based clustering approach based on ontology reasoning for automatically generating constraints and bridging the semantic gap in satellite image labeling. The use of ontology as a priori knowledge has many advantages that we leverage in the context of satellite image interpretation. The experiments we conduct have shown that our proposed approach can deal with incomplete knowledge while completely exploiting the available one.**

I. Introduction

In the last decade, increasingly large volumes of remote sensing images have been made publicly available. The analysis and interpretation of these images are no longer manually feasible but they are mandatory to find actionable solutions to today's environmental and societal issues.

One of the most important methods used in the process of knowledge extraction from satellite images is clustering. Classically used in exploratory and unsupervised settings, clustering aims at partitioning large volumes of non-labeled instances into groups of data based on similarity, density, and proximity. However, in many cases, the quality of such partitioning is relative and highly depends on the user's points-of-interest and/or expertise. Two experts (e.g., hydrologist and urban geographer) who do not have the same topic of interest will evaluate differently the same image clustering/classification output. Another challenging task is related to the interpretation and semantics to associate to the clustering output. In this context, the introduction of knowledge in the process becomes essential, either for guiding the clustering task or for helping the interpretation of the clustering results. Furthermore, in the field of knowledge engineering, ontologies have shown their effectiveness especially in facilitating symbols and semantics anchoring, expressing more and more complex knowledge, as well as performing advanced deductive reasoning. However, the formalization of knowledge remains a bottleneck and some concepts remain difficult or even impossible to define precisely. As an illustrative example, in the context of remote sensing image analysis, experts can more easily define concepts related to vegetation or water rather than concepts related to buildings.

In this paper, we propose a new approach for semantic labeling, automatically combining expert knowledge and clustering with generated constraints. This approach allows to guide and to improve the semi-supervised clustering process based on ontology reasoning. Such approach offers multiple advantages:

- Generalization: The approach can be adapted to different domains;
- Minimal user involvement: Experts are involved only for building the domain ontology that models the knowledge of the domain and the constraints are automatically generated;
- Uncertainty management: Our approach can cope with incomplete and uncertain knowledge bases;
- Adaptive clustering: Clustering is automatically adapted to get as closer as possible to the vision of the expert.

The next sections are organized as following, we will first present related work in Section II. Secondly, we present the core concepts of ontology and description logics reasoning, followed by the description of our approach in Section III. Section IV gives the details of an application of our approach on remote sensing images and summarizes our experimental results. Conclusions and future work conclude the paper in Section V.

II. Related Work

Several studies were conducted to exploit the available domain knowledge. When labeled data is not available or insufficient to perform effective supervised learning, two approaches can be used: knowledge-based classification approaches and semi-supervised clustering approaches. Although these approaches share the same end-goal and use the available knowledge to increase efficiency, they proceed differently. They are also often used at different stages of the knowledge extraction process. The main difference between these two

approaches relies on the type of reasoning they adopt. Most knowledge-based systems use deductive reasoning, whereas semi-supervised clustering approaches are essentially inductive.

A. Knowledge-based systems

Knowledge-based systems have been widely used to reduce the semantic gap and to provide high-level of semantic interpretation [1]. Forestier et al. [2] proposed a method that labels the objects of satellite images using the concepts formalized in a knowledge base. First, Forestier et al. use a segmentation algorithm to obtain the objects from the image. Then, a matching process computes the similarity between the characteristics of the objects and the concepts in the knowledge base. The objects are then labeled with the concept having the highest similarity score. Falomir et al. [3] and Andres et al. [4] use description logic reasoning to label the extracted objects from the images. They also perform segmentation over the images to extract the objects.

These approaches rely on expert knowledge in different ways for the semantic interpretation of the extracted objects. However, none of them is integrated into and actually feeds the clustering or segmentation process. Forestier et al.'s method uses similarity measures of semantic descriptors of the objects extracted from the satellite image, but it does not exploit description logics reasoning. Falomir and Andres' methods use reasoning, but they do only on pre-extracted objects from the images, mainly for a posterior interpretation and they do not guide the clustering process. Overall, only few state-of-art methods have applied logic reasoning to satellite images.

B. Semi-supervised clustering

In the literature, a large body of research has been proposed to introduce and leverage a priori knowledge in clustering [5]–[7]. Typically, several ways have been explored for integrating expert knowledge and supervision into the clustering process. Constraint-based clustering at the instance level is known to be very efficient to guide the cluster formation. Initially introduced by Wagstaff and Cardie [8], knowledge is expressed as two types of links: *must-link* and *cannot-link*. The constraint *must-link* $ml(x_i, x_j)$ specifies that two instances, noted x_i and x_j have to be in the same final cluster, whereas *cannot-link* $cl(x_i, x_j)$ indicates that the two instances should not belong to the same cluster. Both *must-link* and *cannot-link* constraints are transitive.

Several variants of constrained clustering have been proposed in the literature and can be classified into three categories whether they integrate:

- Change of the update step for assigning the instances to the final clusters [9], [10];
- Adjustment of the initialization step of the clustering [11];
- Adaptation of the objective function of the clustering [5].

COP-KMEANS [9] is an adaptation of the k-means algorithm that integrates constraints. An instance is assigned to a cluster only if no constraint is violated. Other techniques modify the initialization step of the clustering algorithm. In the variant of hierarchical clustering based on constraints proposed by Davidson and Ravi [11], transitive closures are computed from the constraints to produce connected instances that are used later by the clustering algorithm.

Algorithms proposed by [9]–[11] have shown that clustering results can be improved by the use of constraints guiding the clustering. However, *hard constrained* clustering variants adopt a *strict enforcement* approach where the algorithm has to find the most feasible clustering output that respects *all* the constraints. Experiments made by Davidson et al. [12] have shown that these algorithms are very sensitive to noise and have issues with inconsistent constraints. *Soft constrained* clustering algorithms have been proposed for *partial enforcement* of the constraints, in order to find the best clustering output that respects the maximum number of constraints. Most of these approaches rely on the modification of the objective function of the clustering adding a penalty weight in case of constraint violation, e.g., *CVQE* [11] and *PCKmeans* [13].

Although constraint-based clustering has received lot of attention in the last years, only few work is focused on automating constraint generation for guiding the clustering process. Current methods rely on manually defined constraints by the expert or user. In many applications, setting constraints can be very expensive and it requires in-depth knowledge about the data and the domain. Another inconvenience of encoding knowledge using only constraints is the loss of class semantics. In this context, we claim that guiding the constraint-based clustering process with formalized knowledge will permit **automatic generation** of constraints relevant to the user/expert points-of-interest.

C. Ontology and clustering

There exist some researches that used jointly ontology and clustering. Most of them are proposed in the context of text mining. Jing et al. [14] proposed a method to improve text clustering using an ontology based distance. The process takes into account the correlations between the terms by exploiting the a priori knowledge contained in Wordnet and other ontologies. Once the correlations computed, the measures were implemented in k-means and experiments have shown improved performance. Hotho et al. [15], [16] exploited the ontology as a support of a priori knowledge at different stages of the knowledge extraction process. The ontology is used in the preprocessing step in order to obtain different text representations. Several k-means is then performed over the different obtained representations, which allow the explanation of the results by selecting the corresponding concepts in the ontology. Theses approaches [14]–[16] used WordNet as a background knowledge, which makes these approaches work well only for topics covered by WordNet. To overcome this problem, Hu et al. [] exploited Wikipedia as an external knowledge and developed approaches that map the text documents to Wikipedia concepts and categories. Once the mappings are established, the documents are clustered based on similarity metric that combines document content information, concept information and categories information.

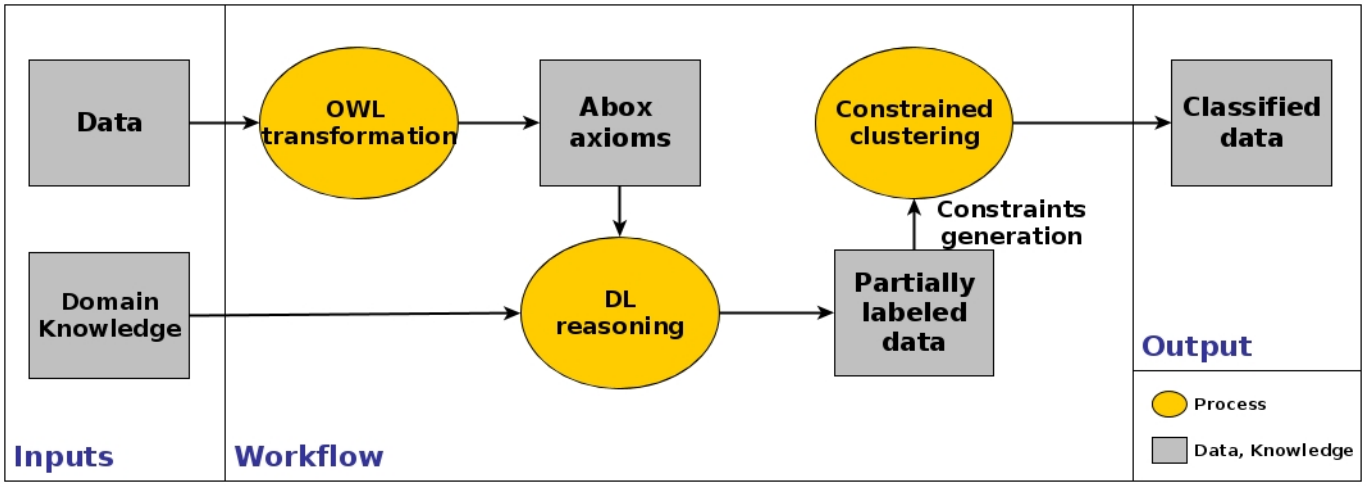

Fig. 1. Overview of Our Approach

Although these approaches are very interesting and use the ontology as a priori knowledge. They are specific to text mining and cannot handle quantitative data. This means that the proposed methods do not deal with the semantic gap problem [17]. An other difference with the presented approaches holds on the relation between the ontology and the clustering. The ontology is first used to match the concepts to the terms present in the text and then replaces the terms by the concepts or introduces additional features to the original data. But no constraints are introduced in the clustering step, and most of the time, a simple clustering is performed.

III. Back-End Ontology for Improving Constraint-based Clustering

A. Preliminaries

We introduce in this section some important elements of the Web Ontology Language (OWL 2) [18] and the description logics (DL) [19]. OWL is a standard language introduced and maintained by the World Wide Web Consortium. The aim of OWL is to give users a simple way to represent rich and complex knowledge, while facilitating the sharing and the publication of this knowledge over the Web. OWL introduces standardized elements with precise meaning and formal semantics. The formal part of OWL is mainly based on DL, which are a family of knowledge representation languages used before OWL to capture a representation of the domain knowledge in a structured way.

In the following, we define an ontology $\mathcal{O}$ as a set of axioms (facts) describing a particular situation in the world from a specific domain point-of-view[1]. Formally, an ontology consists of three sets: the set of classes (concepts) denoted $\mathcal{N}_\mathcal{C}$, the set of properties (roles) denoted $\mathcal{N}_\mathcal{P}$, and the set of instances (individuals) denoted $\mathcal{N}_\mathcal{I}$. Conceptually, it is often divided into two parts: TBox $\mathcal{T}$ and ABox $\mathcal{A}$, where the TBox contains axioms about classes (Domain knowledge) and ABox contains axioms about instances (data), such as:

$$\mathcal{O} = \quad <\mathcal{T},\mathcal{A}> \; = \quad <\mathcal{N}_\mathcal{C},\mathcal{N}_\mathcal{P},\mathcal{N}_\mathcal{I}> \qquad (1)$$

The formalization of the knowledge using formal semantics allows automatic interpretation. This is done by computing the logical consequences of the explicitly stated axioms in $\mathcal{O}$ to infer new knowledge [20]. An interpretation $\mathcal{I}$ of an ontology $\mathcal{O}$ consists of $(\Delta^I,.^I)$, where Δ^I is the domain of I, and $.^I$ the interpretation function of I that maps every class to a subset of Δ^I, every property to a subset of $\Delta^I \times \Delta^I$, and every instance a to an element $a^I \in \Delta^I$.

When quantitative data is available in the domain, extracting high-level semantics from low-level features is difficult. This is a known issue in image analysis [17]. Many approaches have been proposed to tackle this problem also known as the semantic gap and to some extend, ontologies have demonstrated their efficiency to reduce the semantic gap. In OWL, the support of XSD datatypes[2] plays a key role. From a **DL** point-of-view, datatypes can be seen as a *concrete domain* [21] as introduced in OWL. This means that these elements come with a predefined and unique interpretation. This also allows the definition of more complex concepts using qualified datatype restrictions and logical operators.

The interpretation of the ontology is computed using DL reasoners, which provide a set of inference services. Each inference service represents a specific reasoning task. This capability makes OWL very powerful for both knowledge modeling and knowledge processing.

B. Overview of our approach

In this paper, we propose a new method combining deductive reasoning based on **DL** and semi-supervised clustering based on automatically generated constraints. The key idea of

[1] In DL literature, an ontology is considered to be equivalent to a Knowledge Base.

[2] W3C XML Schema Definition Language (XSD) : https://www.w3.org/TR/xmlschema11-2/

our approach is to reason over the available domain knowledge in order to obtain semantically labeled instances, and use these labeled instances to generate constraints that will further guide and enhance the clustering. To enable automatic interpretation of expert knowledge, we use OWL to formalize the domain knowledge and DL reasoning to predict instances types. The exploitation of reasoning keeps our approach generic and modular. As inputs for the semi-supervised clustering, the constraints are generated automatically from the reasoning results and without using any user-labeled data. These elements make our proposition applicable to any problem where domain knowledge, even incomplete, is available.

Figure 1 shows an overview of the proposed approach. The inputs are unlabeled data $X = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$ where each instance x_i is described by a set of attributes $V = \{v_j\}_{j=1}^{d}$, and a formalized domain knowledge representing the TBox $\mathcal{T} =< \mathcal{N}_\mathcal{C}, \mathcal{N}_\mathcal{P} >$ of the ontology. An important aspect about the formalized domain knowledge that can be used in our method is its ability to bridge the semantic gap. The concepts of the TBox have to be defined using low-level properties. The paper's focus is not about domain knowledge formalization, but lot of studies can be found in the literature on this topic [3], [4], [22], [23]. The figure illustrates the steps to guarantee an efficient use of the available knowledge to improve and guide the clustering; these steps can be listed as follows :

- Transformation of the data instances into ABox axioms;
- Reasoning over the constructed Knowledge Base (KB) for instance semantic classification;
- Constraint generation based on the results of the semantic labeling;
- Semi-supervised clustering;
- Capitalization of the results and clusters labeling.

In the rest of the section, we will give a detailed description of these steps. We will also illustrate the effects of our approach using a simple dataset (Figure 2). As an illustrative example, we suppose that the expert is interested in identifying four classes of regions in satellite images and that he has a formalized knowledge about two concepts *C1* and *C2*. The two other concepts are not available (unknown or hard to formalize).

The first step is the data transformation to OWL axioms. We have designed and implemented a semi-automatic process (Algorithm 1) that performs this projection. As shown in the algorithm, our process takes as inputs the TBox of the ontology and the data X to transform. Based on the properties $\mathcal{N}_\mathcal{P}$ of the TBox and the variables V describing the data, the process suggests a mapping between the inputs. Once the mapping is established, our process generates OWL axioms that represent the data. Each instance x_i is represented as an OWL instance a_i (Algorithm 1), where a_i is described by the properties available in TBox, and where these properties get their values from the data. At this point, all the required components are prepared to build up the Knowledge Base (KB).

Algorithm 1 Semi-Automatic Transformation of Data to ABox Assertions

Inputs:
Data $X = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$ described by $V = \{v_j\}_{j=1}^{d}$
Domain Knowledge : $\mathcal{T} =< \mathcal{N}_\mathcal{C}, \mathcal{N}_\mathcal{P} >$
Output:
ABox : $\mathcal{A} = \{a_i\}_{i=1}^{n}$
Method:
for all p_k in $\mathcal{N}_\mathcal{P}$ **and** v_i in V **do**
  Boolean Query = Does p_k correspond to v_i
  if Query.isTrue() **then**
    $map(\mathcal{N}_\mathcal{P}, V)$.add($p_k$,$v_i$)
  end if
end for
for all x_i in X **do**
  $a_i := createOWLInstance()$;
  for all p_k in $map(\mathcal{N}_\mathcal{P}, V)$ **do**
    $a_i.addProperty(p_k)$
    $a_i.setPropertyType(p_k, \mathcal{T}.getPropertyType(p_k))$
    $a_i.setPropertyValue(p_k, x_i.getValueOf(v_k))$
  end for
  return a_i : OWL representation of x_i
  $\mathcal{A}.add(a_i)$
end for

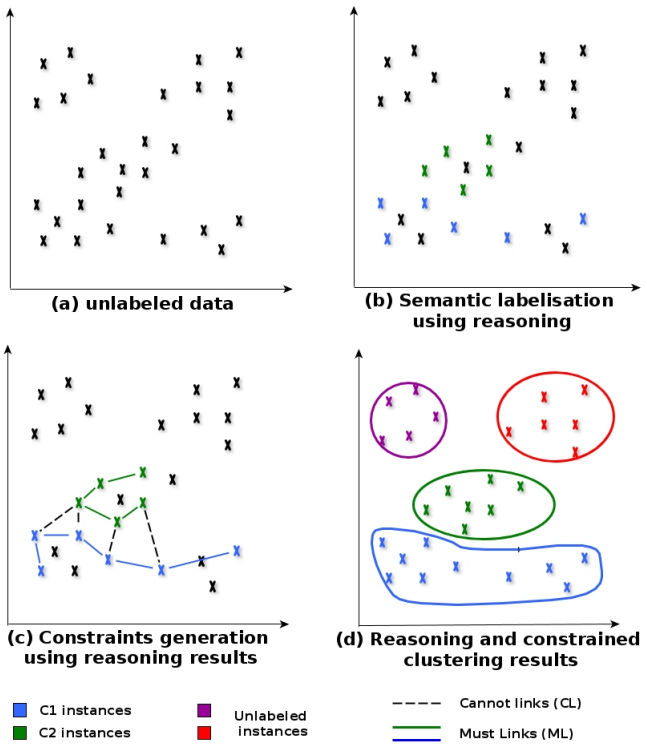

Fig. 2. Illustrative Example (Should be analyzed in color mode)

Note that the **KB** is actually the combination of TBox and ABox, where TBox represents the formalized domain knowledge, while ABox corresponds to the axioms describing

the instances (data). As the OWL language is based on description logic, it allows the use of deductive reasoning to infer new possible knowledge. A set of inferences services can then be used, such as *concept satisfiability*, *classification*, *realization*, etc.

In our method, we are interested in the *realization*, which consists in finding the most specific concept which a given instance belongs to. Performing this reasoning task over the constructed knowledge base allows us to retrieve the instances of the concepts formalized in TBox. Description logic reasoning operates under the *open world assumption*, DL were designed to deal with incomplete information. This means that not all the data will be labeled and only the instances that fit completely in the definition of the concepts will be typed (Figure 2.b).

Once we classify the instances using the ontology, we generate a set of *ML* constraints linking together the identified instances from the same concept and a set of *CL* constraints for the instances identified as belonging to different instances (Figure 2.c). The generated constraints will be used to guide the clustering.

As mentioned above (Section II.B), two variants of constrained clustering exist, hard and soft constrained algorithms. In our case, the constraints are automatically generated based on reasoning over the available knowledge. In this automated process, encoding the results of reasoning as soft constraints is the only way to guarantee the consistency of our approach as the knowledge can contain some approximations and produce some errors. In this step, we choose to use PCKMeans [13] as constrained algorithm, but our method can be applied using most soft constrained algorithms. Compared to the classical k-means algorithm, the objective function of PCKMeans is weighted by the ML and CL constraints :

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

where l_i ($l_i \in h_{h=1}^{k}$) is the cluster assignment of the instance x_i, and $w_{ij}1[l_i \neq l_j]$ and $\overline{w}_{ij}1[l_i = l_j]$ correspond to the cost of the violation of constraints $ml(x_i, x_j) \in ML$ and $cl(x_i, x_j) \in CL$. Note also that 1 is an indicator function with $1[true] = 1$ and $1[false] = 0$. x_i represents the instance affected to the partition χ_{li} with the centroid μ_{li}. Algorithm 2 shows the adapted PCKMeans with automatically generated constraints. Once we apply the constrained clustering, we obtain semantically labeled clusters that respect the expert's vision. Figure 2.c shows the obtained results, we can see that clusters are labeled with the available concepts. The figure also shows how the domain knowledge guided the formation of the clusters based on the automatically generated constraints, even if this knowledge is incomplete.

Algorithm 2 Semi-supervised Clustering with Generated Constraints

Inputs:
Dataset : $X = \{x_i\}_{i=1}^{n} \in \mathbb{R}^d$
Sub-dataset of labeled instance: $X_L = \{(x_i, C_l)\}_{i=1}^{m}$
Where $C_l \in \mathcal{N}_\mathcal{C}$ the set of Classes of the TBox
k : number of clusters
Method:
Generate the ML and CL constraints from X_L
$\lambda = size(\mathcal{N}_\mathcal{C})$
for all Concepts C_l in $\mathcal{N}_\mathcal{C}{}_{l=1}^{\lambda}$ **do**
  Create neighborhood $N_l \in \{N_p\}_{p=1}^{\lambda}$
  Set neighborhood class to C_l
end for
if $\lambda \geq k$ **then**
  initialize $\{\mu_h^{(0)}\}_{h=1}^{k}$ with centroids of $\{N_p\}_{p=1}^{k}$
else if $\lambda < k$ **then**
  initialize $\{\mu_h^{(0)}\}_{h=1}^{\lambda}$ with centroids of $\{N_p\}_{p=1}^{\lambda}$
  if $\exists$ point x cannot-linked to all neighborhoods $\{N_p\}_{p=1}^{\lambda}$ initialize $\mu_{\lambda+1}^{(0)}$ with x
  Initialize remaining clusters randomly
end if
Repeat until *convergence*
assign_cluster: assign each $x_i \in X$ to the cluster h^*, for $h^* = argmin(\frac{1}{2}\|x_i - \mu_h^{(t)}\|^2 + w\sum_{(x_i,x_j)\in ML} 1[l_i \neq l_j] + w\sum_{(x_i,x_j)\in CL} 1[l_i \neq l_j]$
estimate means :
$\{\mu_h^{(t+1)}\}_{h=1}^{k} = \{\frac{1}{\|X_h^{(t+1)}\|}\sum_{x\in X_h^{(t+1)}} x\}_{h=1}^{k}$
$t = t + 1$

IV. Experiments

In this section, we describe the data used in our implementation and the results we obtained. We apply our approach to a real-world application in the classification of Landsat 5 TM images. The Landsat program is a joint NASA/USGS program[3] that freely provides satellite images covering all the earth surface. The Landsat scenes can be downloaded from the USGS Earth Explorer[4]. The Landsat 5 TM scenes have a spatial resolution of 30 meters and seven spectral bands. In our experiments, we extract four images that come from three different scenes. The size of each one of them is of 780x600 pixels, making each dataset having a size of 468.000 pixels. Each instance is described by 9 attributes, the seven bands and two spectral indices (NDVI [24] and NDWI [25]). Three images concern the region of the river Rio Tapajos in the Amazon and one concerns the region of Languedoc Roussillon in the South of France.

In our experiments, the only inputs are the TBox of the ontology containing the expert knowledge, i.e., the formalization of two concepts: Vegetation and Water, and the pixels of the images. We do not use any labeled data and the images are not segmented. During the constrained clustering step of our approach, we fixed k=3 as the number of cluster for PCKMeans and the constraints are automatically generated based on the reasoning results (See Figure 1). The set of must

[3]Landsat Science : http://landsat.gsfc.nasa.gov/
[4]USGS Earth Explorer : http://earthexplorer.usgs.gov/

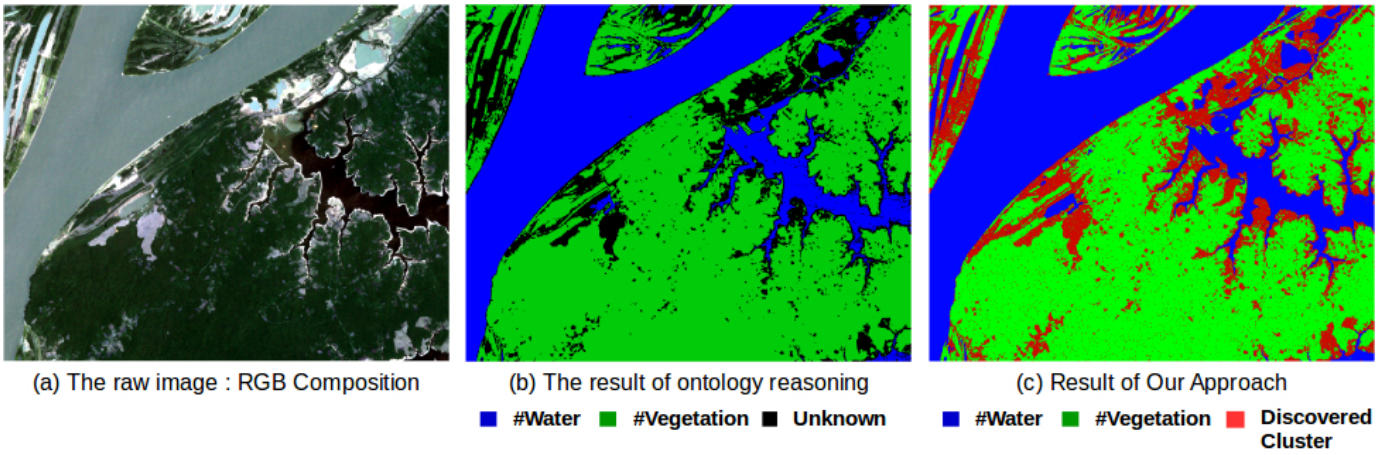

Fig. 3. Application of our approach to an Landsat satellite image Link : http://earthexplorer.usgs.gov/metadata/3119/LT52280622011302CUB01/

link constraints contains a pair of pixels that are both labeled as Vegetation Pixels or Water Pixels. The set of cannot link constraints contains a pair of pixels that one of them is labeled as Water Pixel and the other one as a Vegetation Pixel.

To build the corresponding TBox, several spectral bands and indices were used. The concepts were defined using the seven bands : TM1,...,TM7 and the spectral indices NDVI[5] [24] and NDWI[6] [25]. For example, the water concept is defined as follows :

$Water\_Pixel \equiv Pixel \wedge ((\exists TM4. < 0.05 \wedge \exists ndvi. < 0.01) \vee (\exists TM4. < 0.11 \wedge \exists ndvi. < 0.001))$

a) Implementation notes: Several frameworks have been used to implement our approach. A dedicated process for preprocessing and transformation of the images have been developed using the *Orfeo ToolBox* library[7]. Concerning the semantic layer, the transformation of the data to OWL individuals is ensured by a Java program that uses the *OWL API* [26] and a semi-automatic mapping [Algorithm 1]. Pellet [27] is the DL reasoner that have been chosen to perform the realization task and materialize the deduced type of the pixels. Pellet has a *xsd* datatype oracle that can reason over the qualified datatype restrictions used to define our concepts. Finally, constraint generation and the constrained clustering PCKMeans have been also implemented in Java.

In the rest of this section, we first present the experiments we made with the three images extracted from the two Landsat Scenes of Brazil. Then, we show the results of application of our approach - using the same knowledge - to an image extracted from the Landsat scene of the south of France.

A. Images from Landsat Scenes of Brazil

The experiments we conducted have multiple objectives. First, they show the feasibility of our approach and the advantages of simultaneous exploitation of ontology reasoning and constrained clustering. Secondly, they highlight the capacity of the approach to deal with incomplete domain knowledge. Finally, they demonstrate the effectiveness of the automatic constraints generation from the ontology to supervise clustering.

Figure 3 shows the results of applying our approach to one of the images used in the experiments. For a better understanding, Figure 3 should be analyzed in color mode. Figure 3.a represents the raw image in true colors; Figure 3.b shows the intermediate step of ontology reasoning and the final result of the approach is illustrated in Figure 3.c. We can visually see that our approach improves the results obtained only with reasoning over the ontology. Two important elements are shown in this figure. The first point we can notice is the labeling of water present in the top left corner of the images. These pixels have been semantically labeled using constrained clustering. This shows how our approach can complete the knowledge about the concepts of the ontology. Here, the definition of the experts have not been sufficient to label those pixels, but using the constrained clustering, those pixels have been correctly labeled. The second element is the apparition of the new cluster, which has been identified with the clustering. This cluster, representing the bare soil (as reported later by the expert), is not specified in the ontology but has been detected by the constrained clustering. These figures clearly illustrate how our approach can deal with different paradigms and produces labeled and unlabeled clusters simultaneously.

When we apply k-means and our approach on the same image (Figure 4), we can easily see the improvements offered by the ontology. The first difference is the automatic semantic interpretation. A second observation concerns the confusion between water and vegetation when we use k-means. The expert explains that those errors are due to the nature of the Amazonian forest, where some vegetation grows on wet soils. The ontology is very useful in this case, where the available domain knowledge helps our approach to distinguish the instances of the two concepts (vegetation and water).

To evaluate the quality of the results, we calculate precision and F-measure metrics based on a reference classification

[5]Normalized Difference Vegetation Index
[6]Normalized Difference Water Index
[7]Orfeo ToolBox : https://www.orfeo-toolbox.org/

made by the expert. To obtain this reference classification, a remote sensing expert first labeled manually large parts of the images (about 70%) and then performed a supervised learning with the Envi software[8]. The expert repeats the classification until obtaining a high quality reference classification. We also compare the scores of our approach with the ones obtained by k-means and ontology reasoning. We have to point out that the evaluation of k-means is made after the intervention of the expert to label each cluster, which constitutes an important difference with our approach where the semantic labeling is automatic.

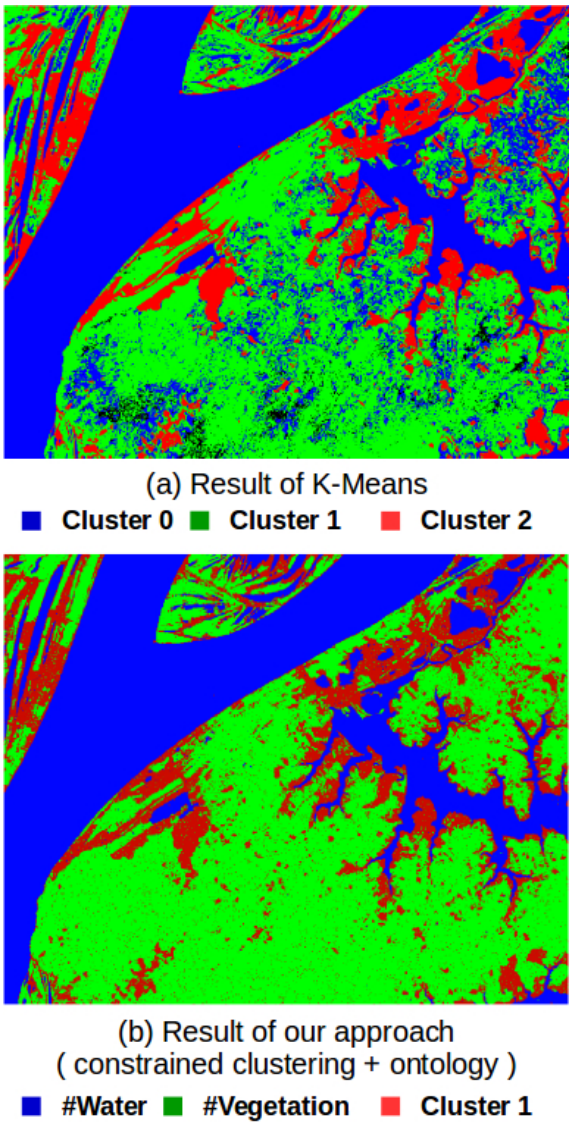

Fig. 4. The results of k-means (a) and our approach (b) applied to the same image

Table 1 shows the performances of the different methods on the three images provided by the expert. Reasoning over the ontology to label the pixels is operated under the open world assumption (OWA) and TBox contains the formalization of the two concepts of Water and Vegetation. This configuration leads to a partial labeling of the instances (83.6 % for Image 1). If we consider only the labeled pixels, the precision of the ontology is high. However, the ontology is not able to label the pixels with the third concept (as no definition is given). It is also not able to label all the vegetation and water pixels as they are not exactly in conformance with the specifications of the concept (Figure 3.b). This disadvantage of the ontology is also a motivation to use our approach, where all the pixels are classified. Table 1 also shows the improvements of the clustering results in our approach when compared to k-means, which demonstrates that, in addition to the semantic interpretation, our approach has better performance.

We also evaluate the obtained results using the Friedman test. A critical Friedman diagram represents a projection of average ranks of classifiers on enumerated axis. The classifiers are ordered from left (the best) to right (the worst) and a thick line connects the classifiers where the average ranks are not significantly different (for the 5% level of significance). From this two tests (Figures 5 and 6), it can be observed that our approach outperforms both the classical clustering and the ontology-based classification as it is situated on the left side of the both figures.

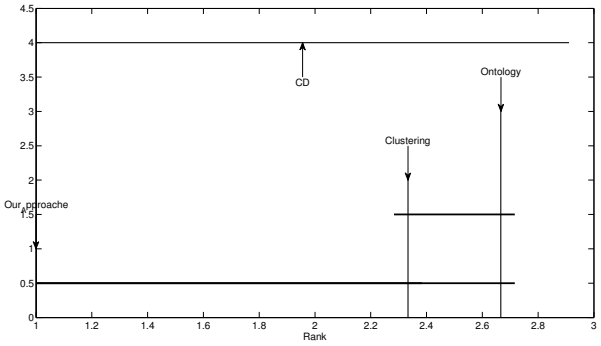

Fig. 5. Friedman Test for Precision Results

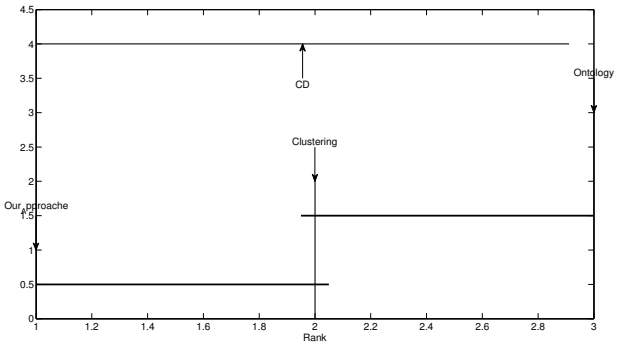

Fig. 6. Friedman Test for F-Measures Results

B. *Image from Landsat Scene of France*

We also applied our approach to an image (size: 468.000 pixels) extracted from a Landsat 5 TM from the south of

[8]Envi : http://www.exelisvis.co.uk/ProductsServices/ENVIProducts/ENVI.aspx

Images	Clustering		Ontology			Our approach	
	Prec.	F-Mes.	% labeled	Prec.	F-Mes.	Prec.	F-Mes.
Image 1	0.8899	0.8764	83,6	0.8359	0.8360	**0.9445**	**0.9296**
Image 2	0.8701	0.8598	81,26	0.8125	0.8126	**0.9271**	**0.9181**
Image 3	0.8889	0.9241	90,33	0.9031	0.9020	**0.9299**	**0.9304**

TABLE I
EXPERIMENTS RESULT ON IMAGES OF THE AMAZONIAN REGION

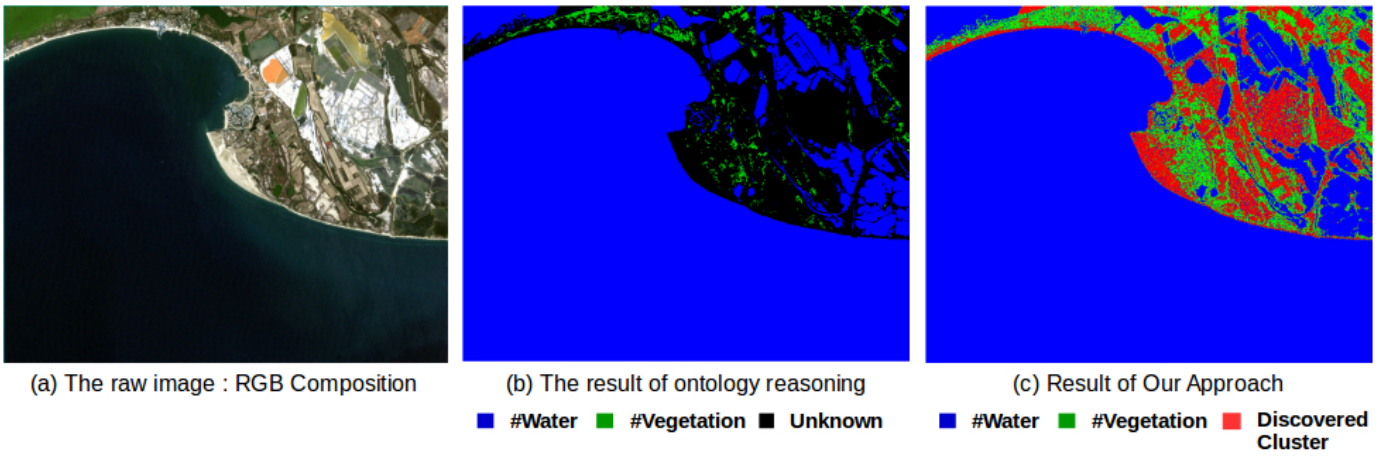

Fig. 7. Application of our approach to an image extracted from a Landsat 5 TM scene of the south of France

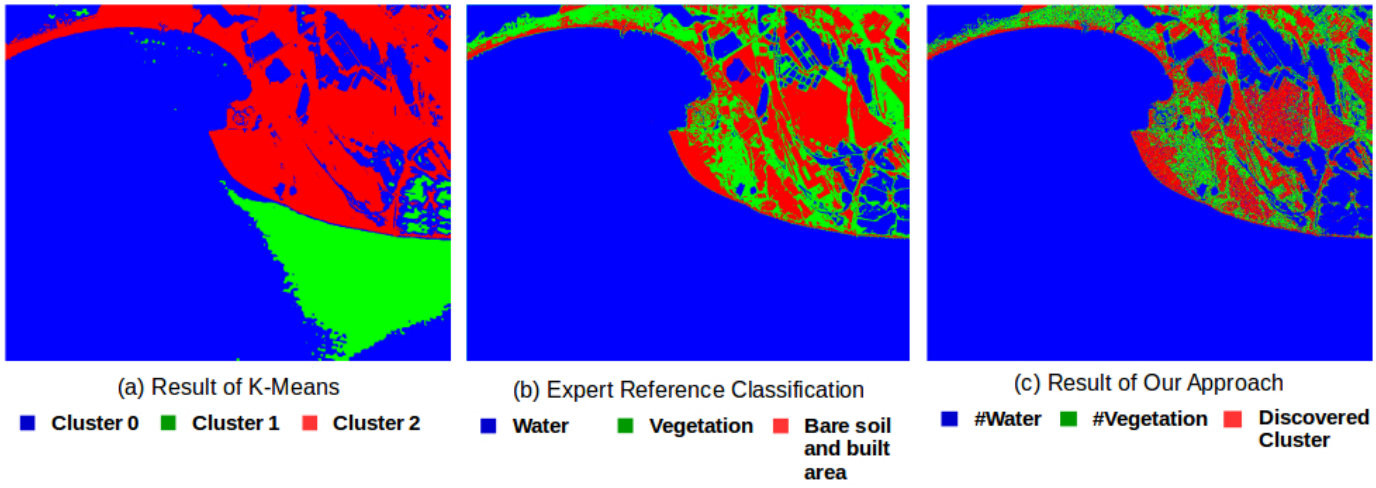

Fig. 8. Results of k-means (a) and our approach (c) compared with an expert reference classification of an excerpt of a Landsat scene of the south of France

France[9]. As mentioned in the beginning of this section (Section V), the same settings were used in this second experiment. The TBox contains two concepts (Water and Vegetation). No labeled instances are used and no expert's intervention is needed during the approach application. Figure 7 shows the different steps of our approach, with the raw image showed in (a), the ontology reasoning result in (b) and the final result of our approach in (c). In Figure 7.b, where the ontology reasoning results are shown, we can see that the water pixels are well detected this time, but that only few vegetation pixels are. This means that the vegetation concept does not cover all the pixels of vegetation present in this image. As the experts tend to define the concepts using only trusted interval values, the concepts cannot cover all the possibilities, as illustrated here (Figure 7.b) with the vegetation concept and in the Figure 3.b with the non detected water pixels in the top left of the image. However, even with this small number of vegetation pixels detected with ontology reasoning, our approach has been able to complete the classification of other vegetation pixels (Figure 7.c). This illustrates the capacity of our approach to work with incomplete knowledge bases.

The figure 8 shows in (a) the result of k-means, (b) the remote sensing expert reference classification made using a manually labeled data and a supervised learning with the Envi

[9]Link : http://earthexplorer.usgs.gov/metadata/3119/LT51960302011286MTI00/

commercial software, and finally (c) the result of our approach. The comparison with k-means highlights an interesting property of our approach, it shows its capacity to efficiently control the clustering via the automatically generated constraints. When we apply k-means with k=3, we obtain two cluster containing water pixels and a cluster with mixed pixels. This is due to the large number of water pixels and to the initialisation problem in unsupervised settings. With our approach, we can see that the obtained results are very close to the reference classification made by the expert. We recall that no labeled data were used and no manual intervention is needed in our case. The hybrid nature of our approach allows the use of incomplete knowledge via deductive reasoning and exploits it to generate constraints for the constrained clustering. The constrained clustering step will then complete the classification of the pixels if the definition of the concepts does not cover all the pixels (Vegetation pixels in Figure 7) and discovers new clusters (The red pixels in Figure 7.c) via its induction nature.

V. Conclusion

We have presented in this paper a new hybrid approach combining reasoning over an ontology and clustering, guided by automatically generated constraints. Combining both deductive and inductive reasoning, our method can exploit the available knowledge although it is incomplete. We have applied our approach to a real-world use case of satellite image classification. The results have shown that our approach improves the quality of the clustering while automating semantic labeling.

As future work, we plan to extend our approach by prioritizing the constraints in the generation process (currently all the constraints have the same weight). Another perspective concerns the enrichment of the ontology by adding new thematic concepts.

Acknowledgment

This work was supported by the French Agence Nationale de la Recherche under Grant ANR-12-MONU-0001.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[2] G. Forestier, A. Puissant, C. Wemmert, and P. Gançarski, "Knowledge-based region labeling for remote sensing image interpretation," *Computers, Environment and Urban Systems*, vol. 36, no. 5, pp. 470–480, 2012.

[3] Z. Falomir, E. Jiménez-Ruiz, M. T. Escrig, and L. Museros, "Describing images using qualitative models and description logics," *Spatial Cognition & Computation*, vol. 11, no. 1, pp. 45–74, 2011.

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

[5] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 59–68.

[6] I. Davidson and S. Basu, "A survey of clustering with instance level constraints," *ACM Transactions on Knowledge Discovery from Data*, pp. 1–41, 2007.

[7] D. T. Truong and R. Battiti, "A survey of semi-supervised clustering algorithms: from a priori scheme to interactive scheme and open issues," 2013.

[8] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," *AAAI/IAAI*, vol. 1097, 2000.

[9] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, 2001, pp. 577–584.

[10] N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall, "Computing gaussian mixture models with em using equivalence constraints," *Advances in neural information processing systems*, vol. 16, no. 8, pp. 465–472, 2004.

[11] I. Davidson and S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *Knowledge Discovery in Databases: PKDD 2005*. Springer, 2005, pp. 59–70.

[12] I. Davidson, K. L. Wagstaff, and S. Basu, *Measuring constraint-set utility for partitional clustering algorithms*. Springer, 2006.

[13] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering." in *SDM*, vol. 4. SIAM, 2004, pp. 333–344.

[14] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang, "Ontology-based distance measure for text clustering," in *Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*, 2006.

[15] A. Hotho, A. Maedche, and S. Staab, "Ontology-based text document clustering," *KI*, vol. 16, no. 4, pp. 48–54, 2002.

[16] A. Hotho, S. Staab, and G. Stumme, "Ontologies improve text document clustering," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003, pp. 541–544.

[17] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. Del Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval," *arXiv preprint arXiv:1503.08248*, 2015.

[18] W. O. W. Group *et al.*, "Owl 2 web ontology language document overview," 2009.

[19] F. Baader, *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2003.

[20] I. Horrocks and U. Sattler, "Ontology reasoning in the shoq (d) description logic," in *IJCAI*, vol. 1, no. 3, 2001, pp. 199–204.

[21] C. Lutz, "Description logics with concrete domains–a survey," 2003.

[22] D. Sheeren, A. Quirin, A. Puissant, P. Gançarski, and C. Weber, "Discovering rules with genetic algorithms to classify urban remotely sensed data," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS2006)*, 2006, pp. 3919–3922.

[23] B. Neumann and R. Möller, "On scene interpretation with description logics," *Image and Vision Computing*, vol. 26, no. 1, pp. 82–101, 2008.

[24] R. DeFries and J. Townshend, "Ndvi-derived land cover classifications at a global scale," *International Journal of Remote Sensing*, vol. 15, no. 17, pp. 3567–3586, 1994.

[25] S. McFeeters, "The use of the normalized difference water index (ndwi) in the delineation of open water features," *International journal of remote sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[26] M. Horridge and S. Bechhofer, "The owl api: A java api for owl ontologies," *Semantic Web*, vol. 2, no. 1, pp. 11–21, 2011.

[27] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical owl-dl reasoner," *Web Semantics: science, services and agents on the World Wide Web*, vol. 5, no. 2, pp. 51–53, 2007.