Towards Ontology Reasoning for Topological Cluster Labeling

Hatim Chahdi[1 2], Nistor Grozavu[2], Isabelle Mougenot[1], Younès Bennani[2], and Laure Berti-Equille[1 3]

[1] Espace-Dev UMR 228, IRD - Université de Montpellier
500 Rue J.F. Breton, 34090 Montpellier, France
email: {firstname.lastname}@ird.fr
[2] LIPN CNRS UMR 7030, CNRS - Université Paris 13
99, av. J-B Clement, 93430 Villetaneuse, France
email: {firstname.lastname}@lipn.univ-paris13.fr
[3] Qatar Computing Research Institute - Hamad Bin Khalifa University
Doha, Qatar

Abstract. In this paper, we present a new approach combining topological unsupervised learning with ontology based reasoning to achieve both : (i) automatic interpretation of clustering, and (ii) scaling ontology reasoning over large datasets. The interest of such approach holds on the use of expert knowledge to automate cluster labeling and gives them high level semantics that meets the user interest. The proposed approach is based on two steps. The first step performs a topographic unsupervised learning based on the SOM (Self-Organizing Maps) algorithm. The second step integrates expert knowledge in the map using ontology reasoning over the prototypes and provides an automatic interpretation of the clusters. We apply our approach to the real problem of satellite image classification. The experiments highlight the capacity of our approach to obtain a semantically labeled topographic map and the obtained results show very promising performances.

1 Introduction and Motivations

Clustering is a very important step in the process of knowledge extraction and discovery. When no labeled instances are available, unsupervised learning aims to discover new structures and group instances according to similarity, density, and proximity. Clustering has multiple applications[10]. We are interested in this work in topological clustering which allows clustering and visualization simultanieosly i.e. the Self-Organizing Map (SOM)[11] algorithm. SOM is an unsupervised artificial neural network that produces a low-dimensional (generally two-dimensional) map from an unlabeled dataset. The nodes of the map represents a summarized version of the original dataset via a set of reference vectors (prototypes) spatially organized. This makes SOM suitable for lot of practical applications, including data visualization, data summarization and compression. However, as an unsupervised learning algorithm, the labeling and interpretation

of the resulting map have to be performed manually, which can be difficult and time consuming. For example, in the field of remote sensing images analysis, where domain expertise of the user have to be very high, the interpretation of the results is not an easy task and only experts can handle it. In addition, the fact that images are encoded using low level features (numerical data) makes the interpretation even harder. This issue is known in the literature as the semantic gap[7].

The field of knowledge representation has been subject to lot of researches last years. These researches, supported by the emergence of the semantic Web[18], have lead to the development of the OWL 2[6] as a standard language of ontology modelling. Based on description logics[1], OWL offers a standardized way to represent rich and complex knowledge. It comes with standard elements with precise meaning and formal semantics. This formal basis gives reasoners[20, 5] the possibility to automatically process the ontologies and propose a set of inferences services that deduce new knowledge by calculating the logical consequences of the present facts, like the instances classes. However, when it comes to use ontology to classify a large number of instances, the reasoners fail to scale[8]. This issue can be very problematic when dealing with real world problems.

One of the motivations of our work is to use ontology reasoning to automate the labeling of a topological map. In our approach, the clusters are labeled using the concepts defined in the ontology. The use of an ontology as a support of the expert knowledge introduces also a modularity to our approach. The reasoning is a procedural process and the results change automatically when the concepts of the ontology change. This allow our approach to give interpretations that automatically meet the interest of the user following the expressed concepts in the ontology. Another motivation is the possibility to scale reasoning over large datasets. The SOM algorithm represents the input data using a finite set of prototypes. By reasoning over the prototypes and not all the input data, the reasoning can be performed over large datasets and in a shorter time.

The next sections are organized as following, we will first present related work in Section 2, followed by some preliminaries about ontology and reasoning in section 3. The description of the different steps of our approach are given in section 4. Section 5 highlights the experiments we made on the UCI wine dataset, it also shows the application of our approach on the real problem of remote sensing images classification. Conclusion and future work conclude the paper in Section 6.

2 Related Work

The problem of cluster labeling has been subject to different interesting researches in the literature[12, 4, 15–17]. These researches have explored different techniques to achieve cluster labeling. A version of SOM dedicated to textual data, called LabelSOM[17] was proposed by Rauber and Merkl. The authors labeled the trained map with a set of features of the data input. In another research, Treeratpituk et al.[21] proposed a method to label hierarchical text clustering. The presented algorithm assigns few labels to the clusters based on the cluster analysis information, the parent cluster and statistics about the corpus. Recently, Li et al.[12] proposed an hybrid approach, combining linguistic and statistical techniques to achieve an automated labeling of the

clusters. Although these approaches are very interesting, they are all dedicated to textual data and cannot work on quantitative data.

When it comes to methods that deal with numerical data, most of the approaches use *a priori* knowledge on the data to propose candidate labels of the clusters[3][4]. The work presented in this paper is focused in proposing an hybrid approach producing a semantic interpretation of the SOM's map based on the ontology reasoning. This makes our approach different from the other approaches present in the literature. Only few approaches are capable to propose a high level labels on quantitative data, and these methods are usually not adapted for topological clustering.

3 Preliminaries about Ontology and Reasoning

Before we present the proposed approach, we introduce in this section some core concepts related to ontology and description logics reasoning. We adopt in our work the Web Ontology Language (OWL 2)[6] as a standard language for ontology formalization. OWL was introduced and is now maintained by the World Wide Web Consortium. The aim of OWL is to give users a simple way to represent rich and complex knowledge. OWL introduces standardized elements with precise meaning and formal semantics. The formal part of OWL is mainly based on description logics[1], which is a family of knowledge representation.

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[4]. 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:

$$\mathcal{O} = < \mathcal{T}, \mathcal{A} > \ = < \mathcal{N}_\mathcal{C}, \mathcal{N}_\mathcal{P}, \mathcal{N}_\mathcal{I} > \tag{1}$$

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 [9]. An interpretation $\mathcal{I}$ of an ontology $\mathcal{O}$ consists of (Δ^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 $a^I \in \Delta^I$.

The interpretation of the ontology is computed using DL reasoners, which provides 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. One of the inference services proposed by the reasoner is *instance checking*. This task can be performed to check if an instance a_i belongs to a concept $C \in \mathcal{T}$ based on the definition of the later. By empowering the OWL 2 modelling capacities, the concepts of the ontology can exploit qualified number restrictions over data properties[13] and logical operators to bridge the semantic gap and permit an efficient instance labeling.

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

4 Hybrid approach : SOM Ontology based Labeling

In this paper, we present a new hybrid approach using the available expert knowledge to semantically label the generated SOM map. Given an unlabeled dataset X and a TBox $\mathcal{T}$ of the ontology O, our goal is, to build a labeled map that reduces data dimension and at the same time gives them a semantic labeling. This can bring an understandable view of the results based on the users point of interest. To achieve this automatic labeling, we propose a two steps approach. The first step performs an unsupervised learning based on the SOM algorithm and generate a spatially organized map that summarizes the data in terms of a set of prototypes. In the second step, we use a dedicated process that transforms the prototypes of the map to OWL instances, inject them in a reasoner with the TBox (formalized expert knowledge) and performs a deductive reasoning that produces a semantic labeling of the prototypes based on the concepts present in the ontology. The rest of this section will detail the two steps of our approach.

4.1 Topological Unsupervised Learning Step

The first step of our approach performs an unsupervised learning over the input dataset using the Self-organizing maps. We used the basic model proposed by Kohonen. It consists of a discrete set $\mathcal{C}$ of cells called "map". This map has a discrete topology defined by an undirected graph, which usually is a regular grid in two dimensions. For each pair of cells (j,k) on the map, the distance $\delta(j,k)$ is defined as the length of the shortest chain linking cells j and k on the grid. For each cell j this distance defines a neighbor cell; in order to control the neighborhood area, we introduce a kernel positive function $\mathcal{K}$ ($\mathcal{K} \geq 0$ and $\lim_{|y|\to\infty} \mathcal{K}(y) = 0$). We define the mutual influence of two cells j and k by $\mathcal{K}_{j,k}$. In practice, as for traditional topological maps we use a smooth function to control the size of the neighborhood as $\mathcal{K}_{j,k} = \exp(\frac{-\delta(j,k)}{T})$. Using this kernel function, T becomes a parameter of the model. As in the Kohonen algorithm, we decrease T from an initilal value T_{max} to a final value T_{min}.

Let $\mathbb{R}^d$ be the euclidean data space and $X = \{\mathbf{x}_i; i = 1, \ldots, N\}$ a set of observations, where each observation $\mathbf{x}_i = (x_i^1, x_i^2, ..., x_i^d)$ is a vector in $\mathbb{R}^d$. For each cell j of the grid (map), we associate a referent vector (prototype) $\mathbf{w}_i = (w_i^1, w_i^2, ..., w_i^d)$ which characterizes one cluster associated to cell i. We denote by $\mathcal{W} = \{\mathbf{w}_j, \mathbf{w}_j \in \mathbb{R}^d\}_{j=1}^{|\mathcal{W}|}$ the set of the referent vectors. The set of parameter $\mathcal{W}$ has to be estimated iteratively by minimizing the classical objective function defined as follows:

$$R(\chi, \mathcal{W}) = \sum_{i=1}^{N} \sum_{j=1}^{|\mathcal{W}|} \mathcal{K}_{j,\chi(\mathbf{x}_i)} \|\mathbf{x}_i - \mathbf{w}_j\|^2 \tag{2}$$

where χ assigns each observation $\mathbf{x}_i$ to a single cell in the map $\mathcal{C}$. This cost function can be minimized using both stochastic and batch techniques [11].

4.2 Ontology Based Map Labeling

Once we obtain the SOM's map, we extract the set of referent vectors $\mathcal{W}$ (prototypes), and transform them to OWL instances. In fact, before injecting the prototypes in the

reasoner. They have to be transformed to OWL axioms, where each prototype is presented as an OWL instance and described using the properties present in the TBox of the Ontology O. 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 (formalized expert knowledge), and the prototypes obtained by SOM $\mathcal{W}$. Based on the properties $\mathcal{N}_\mathcal{P}$ of the TBox and the set of variables (features) V describing the data, the process suggests a mapping between the inputs (Algorithm 1, line : 2). Once the mapping is established, our process generates OWL axioms that represent the data. Each prototype $\mathbf{w}_j$ is represented as an OWL instance a_i (Algorithm 1), where a_i is described by the properties available in TBox (Algorithm 1, line : 9-11), and where these properties get their values from the prototypes vector (Algorithm 1, line : 12). At this point, all the required components to perform reasoning are available. Once we transform the prototypes using our algorithm, we obtain an ABox containing

Algorithme 1 : Semi-Automatic Projection of the prototypes in the ontology

Inputs:
Set of prototypes $\mathcal{W} = \{\mathbf{w}_i, \mathbf{w}_i \in \mathbb{R}^d\}_{i=1}^{|\mathcal{W}|}$ 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:

1: **for all** p_k in $\mathcal{N}_\mathcal{P}$ **and** v_j in V **do**
2: **Boolean** Query = Does p_k correspond to v_j
3: **if** Query.isTrue() **then**
4: $map(\mathcal{N}_\mathcal{P}, V)$.add($p_k$,$v_j$)
5: **end if**
6: **end for**
7: **for all** $\mathbf{w}_i$ in $\mathcal{W}$ **do**
8: $a_i := createOWLInstance()$;
9: **for all** p_k in $map(\mathcal{N}_\mathcal{P}, V)$ **do**
10: $a_i.addProperty(p_k)$
11: $a_i.setPropertyType(p_k, \mathcal{T}.getPropertyType(p_k))$
12: $a_i.setPropertyValue(p_k, \mathbf{w}_i.getValueOf(v_k))$
13: **end for**
14: **return** a_i : OWL representation of $\mathbf{w}_i$
$\mathcal{A}.add(a_i)$
15: **end for**

all the OWL instances $a_i \in$ ABox representing the prototypes $w_i \in \mathcal{W}$ of the SOM's map. We inject the ABox with the TBox in a DL reasoner to obtain our knowledge base. We use the Pellet[20] reasoner in our approach because it effectively implements the instance checking task and supports OWL 2 specifications. As mentioned above, instance checking consists in finding the most specific concept which a given instance belongs to. Performing this reasoning task over the constructed knowledge base will label the SOM's map with the concepts formalized in the TBox of the ontology.

5 Experiments

In this section, we present the conducted experimentations and the results we obtained. The purpose of our evaluation is to highlight the effectiveness of our approach to automatically label the SOM's map based on ontology reasoning. We apply our method on two datasets.

The first one is the UCI wine dataset[5], which consists of 178 instances. Each instance is described with 13 variables that represent the quantities of 13 constituents (e.g. alcohol, Mg...) found in each of the wines. The inputs used in our method are the unlabeled dataset, and an ontology about three concepts, those concepts have been constructed following a similar approach to the one proposed by Sheeren et al.[19]. Each concept is defined in OWL 2 using qualified number restrictions over the properties. We apply the different steps described in our approach. First, we applied the SOM algorithm to obtain the map. We fixed the size to 6 x 11. Then we extract the prototypes, transform them using the algorithm 1 and perform an *instance checking* with the Hermit[5] reasoner to label them based on the three concepts. We evaluate the results using purity and the labeling percentage. The labeling percentage is important as it shows the efficiency of our ontology to give automatic interpretation of the results. The purity of our map is 96,62% and 61 (92,42%) prototypes were correctly labeled.

5.1 Satellite images classification

We also apply our approach to a real-world problem of satellite image classification. The image we used is an extract of a Landsat 5 TM image. The Landsat program is a joint NASA/USGS program[6] that freely provides satellite images covering all the earth surface. The image can be downloaded from the USGS Earth Explorer[7]. The Landsat 5 TM have a spatial resolution of 30 meters and seven spectral bands. The size of our image is of 760x680 pixels. The images concern the region of the river Rio Tapajos in the Amazon, Brazil. The input TBox of our ontology contains 2 thematic concepts: Water and Vegetation. 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(Normalized Difference Vegetation Index)[2] and NDWI(Normalized Difference Water Index)[14]. For example, the water concept is defined as follows:
$Water\_Pixel \equiv Pixel \wedge ((\exists TM4. < 0.05 \wedge \exists ndvi. < 0.01) \vee (\exists TM4. < 0.11 \wedge \exists ndvi. < 0.001))$
We applied our approach as described before with a fixed map of 20 x 20. The figure 1 shows visually the the obtained results projected using the satellite image. We evaluated the purity of the results using a reference classification made by a domain expert. We obtained 98,45% for the purity index, with 76% of the prototypes correctly labeled using ontology reasoning. This experiment illustrates how using SOM helped the reasoning to scale as it was performed only on the 400 prototypes instead of the 460.000

[5] Wine dataset : http://archive.ics.uci.edu/ml/datasets/Wine
[6] Landsat Science : http://landsat.gsfc.nasa.gov/
[7] USGS Earth Explorer : http://earthexplorer.usgs.gov/

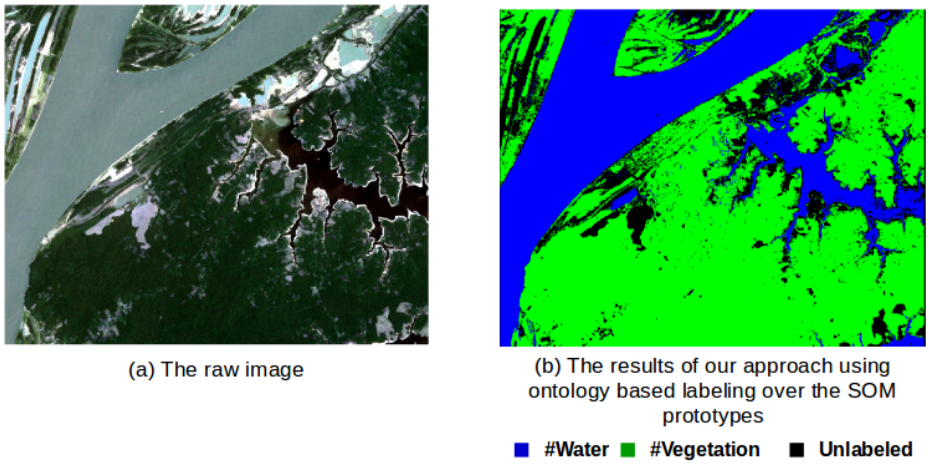

Fig. 1. Application of our approach for the classification of a landsat satellite image

pixels of the image. If we had to reason over all the dataset, the reasoner would not have been able to scale[8].

6 Conclusion and Future Work

We have presented in this paper a new hybrid approach combining the unsupervised topographic learning with ontology reasoning in order to semantically label clustering results. Combining both deductive and inductive reasoning, our method can automate the interpretation of clustering based on the ontology and in the same time scales and speed up the reasoning process by exploiting the proposed prototype label propagation. We have applied our approach to multiple datasets and evaluate the results. We have also shown how our approach can be used in the real-world problem of satellite images classification. As future work, we plan to extend our approach by using a constraints based on the ontology reasoning results to modify the obtained maps and improve its semantic coherence.

Acknowledgment

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

References

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

2. DeFries, R., Townshend, J.: Ndvi-derived land cover classifications at a global scale. International Journal of Remote Sensing 15(17), 3567–3586 (1994)
3. Durand, N., Derivaux, S., Forestier, G., Wemmert, C., Gançarski, P., Boussaid, O., Puissant, A.: Ontology-based object recognition for remote sensing image interpretation. In: Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on. vol. 1, pp. 472–479. IEEE (2007)
4. Forestier, G., Puissant, A., Wemmert, C., Gançarski, P.: Knowledge-based region labeling for remote sensing image interpretation. Computers, Environment and Urban Systems 36(5), 470–480 (2012)
5. Glimm, B., Horrocks, I., Motik, B., Stoilos, G., Wang, Z.: Hermit: an owl 2 reasoner. Journal of Automated Reasoning 53(3), 245–269 (2014)
6. Group, W.O.W., et al.: Owl 2 web ontology language document overview (2009)
7. Hare, J.S., Lewis, P.H., Enser, P.G., Sandom, C.J.: Mind the gap: another look at the problem of the semantic gap in image retrieval. In: Electronic Imaging 2006. pp. 607309–607309. International Society for Optics and Photonics (2006)
8. Horrocks, I., Li, L., Turi, D., Bechhofer, S.: The instance store: Dl reasoning with large numbers of individuals. In: Proc. of the 2004 Description Logic Workshop (DL 2004). pp. 31–40 (2004)
9. Horrocks, I., Sattler, U.: Ontology reasoning in the shoq (d) description logic. In: IJCAI. vol. 1, pp. 199–204 (2001)
10. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM computing surveys (CSUR) 31(3), 264–323 (1999)
11. Kohonen, T.: The self-organizing map. Proceedings of the IEEE 78(9), 1464–1480 (1990)
12. Li, Z., Li, J., Liao, Y., Wen, S., Tang, J.: Labeling clusters from both linguistic and statistical perspectives: A hybrid approach. Knowledge-Based Systems 76, 219–227 (2015)
13. Lutz, C.: Description logics with concrete domains–a survey (2003)
14. McFeeters, S.: The use of the normalized difference water index (ndwi) in the delineation of open water features. International journal of remote sensing 17(7), 1425–1432 (1996)
15. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 490–499. ACM (2007)
16. Popescul, A., Ungar, L.H.: Automatic labeling of document clusters. Unpublished manuscript, available at http://citeseer. nj. nec. com/popescul00automatic. html (2000)
17. Rauber, A., Merkl, D.: Automatic labeling of self-organizing maps: Making a treasure-map reveal its secrets. In: Methodologies for Knowledge Discovery and Data Mining, pp. 228–237. Springer (1999)
18. Shadbolt, N., Berners-Lee, T., Hall, W.: The semantic web revisited. IEEE Intelligent Systems 21(3), 96–101 (May 2006), http://dx.doi.org/10.1109/MIS.2006.62
19. Sheeren, D., Quirin, A., Puissant, A., Gançarski, P., Weber, C.: Discovering rules with genetic algorithms to classify urban remotely sensed data. In: Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS2006). pp. 3919–3922 (2006)
20. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. Web Semantics: science, services and agents on the World Wide Web 5(2), 51–53 (2007)
21. Treeratpituk, P., Callan, J.: Automatically labeling hierarchical clusters. In: Proceedings of the 2006 international conference on Digital government research. pp. 167–176. Digital Government Society of North America (2006)