Class-Membership Propagation in Web Ontologies

Pasquale Minervini

LACAM Laboratory – Dipartimento di Informatica – Università degli Studi di Bari "Aldo Moro"

pasquale.minervini@uniba.it

Abstract

Considering the increasing availability of structured machine processable knowledge in the context of the Semantic Web, relying only on purely deductive inference may be limiting [11]. This work proposes a new method for similarity-based class-membership prediction in Description Logic knowledge bases. The underlying idea is based on the concept of propagating *class-membership information among similar individuals; it is non-parametric in nature, and characterised by a promising time complexity (making it a potential candidate for transductive and inductive reasoning on large and Web-scale knowledge bases).*

1. Introduction and Motivation

Standard Semantic Web (SW) reasoning services rely on purely deductive inference; however, this kind of inference may be infeasible on large-scale and Web-scale knowledge bases. Also, it does not exploit statistical regularities in data; knowledge is inherently incomplete and knowledge bases suitable for deductive inference may be expensive to engineer. For a *reasonable SW* [8], approximate deductive and inductive inference are being discussed as possible alternatives to purely deductive inference [11]. Various approaches to extend inductive inference methods towards SW formalisms have been proposed in SW literature: inductive (and transductive) methods can perform some sort of approximate and uncertain reasoning and derive conclusions which are not derivable or refutable from the knowledge base [11]. This work proposes an approach to transductive inference in Description Logic (DL) representations: the underlying idea is to spread class-membership information among similar individuals.

2. Related Work and Preliminaries

A variety of approaches have been proposed in literature for class-membership prediction, either *discriminative* or *generative* [9]. Informally speaking, generative methods aim at modelling the probability distribution $P(X, Y)$ underlying instances X and their labels Y, while discriminative methods aim at predicting, for a generic instance $x \in X$, whether $P(y \mid x) \geq 0.5$ (binary classification case).

2.1 Discriminative Methods

Some of the approaches proposed for solving the class-membership problem are similarity-based. For instance, methods relying on the k-Nearest Neighbours (k-NN) algorithm are discussed in [4]. Kernel-based algorithms have been proposed for various learning tasks from DL-based representations. This is possible thanks to the existence of a variety of kernel functions, either for concepts or individuals (such as [2, 6]); by (implicitly) projecting instances into an high-dimensional feature space, kernel functions allow to adapt a multitude of machine learning methods to structured representations. SW literature includes methods for inducing robust classifiers [5] or learning to rank [7] from DL knowledge bases.

2.2 Generative Methods

For learning from formal ontologies, a generative approach has been discussed in [12]. In this work, each individual is associated to a *latent variable* which influences its attributes and the relations it participates in. A quite different approach is discussed in [10]: this work focuses on learning theories in a probabilistic extension of the $\mathcal{ALC}$ DL named CR$\mathcal{ALC}$, using DL refinement operators to efficiently explore the space of concepts.

2.3 Semi-Supervised/Transductive Learning

Classic discriminative learning methods ignore unlabelled instances. However, real life scenarios are usually characterized by an abundance of unlabelled instances and a few labelled ones [15]. This may also be the case for class-membership prediction from formal ontologies: class-membership relations may be difficult to obtain during ontology engineering tasks (due to availability of domain experts) and inference (deciding instance-membership may have an intractable time complexity in some languages). Using unlabelled instances during learning is generally known in the machine learning community as *Semi-Supervised Learning* [3, 15] (SSL). If the marginal distribution of instances P_X is informative with respect to the conditional probability distribution $P(Y \mid x)$, accounting for unlabelled instances during learning can provide more accurate results [3, 15]. A possible approach is including terms dependent from P_X into the objective function. This results in the two fundamental assumptions [3]:

- **Cluster assumption** – The joint probability distribution $P(X, Y)$ is structured in such a way that points in the same *cluster* are likely to have the same label.
- **Manifold assumption** – Assume that P_X is supported on a low-dimensional manifold: then, $P(Y \mid x)$ *varies smoothly*, as a function of x, with respect to the underlying structure of the manifold.

We will discuss a similarity-based, non-parametric method for estimating missing class-membership relations, with potentially interesting time complexity characteristics. This method is discriminative in nature, but also accounts for unknown class-membership during learning.

We will face a slightly different version of the classic class-membership prediction problem, namely *transductive class-membership prediction*. We formalise the transductive class-membership prediction problem as a cost minimisation problem: given a set of training individuals $\text{Ind}_C(\mathcal{K})$ whose class-membership relation to a target concept C is either known or unknown, find a function $f^* : \text{Ind}_C(\mathcal{K}) \rightarrow \{+1, -1\}$ defined over training individuals and returning a value $+1$ (resp. -1) if the individual likely to be a member of C (resp. $\neg C$), minimizing a given cost function.

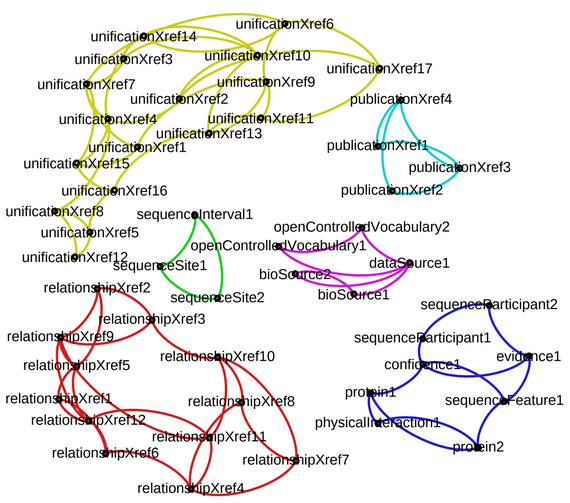
3. Propagating Class-Membership Information Among Individuals

This section discusses a *graph-based semi-supervised* [15] method for class-membership prediction from DL representations. The proposed method relies on a weighted *semantic similarity graph*, where nodes represent positive, negative and neutral examples of the transductive class-membership prediction problem, and weighted edges define similarity relations among such individuals.

More formally, let $\mathcal{K}$ be a knowledge base, $\text{Ind}_C(\mathcal{K})$ a set of training individuals with respect to a target concept C in $\mathcal{K}$, and $Y = \{-1, +1\}$ a space of labels each corresponding to a type of class-membership relation with respect to C. Each training individual $a \in \text{Ind}_C(\mathcal{K})$ is associated to a label, which will be $+1$ (resp. -1) if $\mathcal{K} \models C(a)$ (resp. $\mathcal{K} \models \neg C(a)$), and will be unknown otherwise, thus representing an unlabelled instance. For defining a cost over functions $f \in \mathcal{F}$, the proposed method relies on *regularization by graph*: the learning process aims at finding a labelling function that is both consistent with given labels, and changes smoothly between similar instances (where similarity relations are encoded in the semantic similarity graph). This can be formalised through a *regularization framework*, using a measure of the consistency to the given labels as a loss function, and a measure of smoothness among the similarity graph as a regulariser. Several cost functions have been proposed in SSL literature. An appealing class of functions, from the side of efficiency, relies on the *quadratic cost criterion* framework [3, ch. 11]: for this class of functions, a closed form solution to the cost minimisation problem can be found efficiently (subsection 3.2).

3.1 Constructing a Semantic Similarity Graph

A *semantic similarity graph* encodes similarity relations between individuals in a formal ontology. It can be represented as a weight matrix $\mathbf{W}$, where the value of $\mathbf{W}_{ij}$ encodes the strength of the similarity relation between two training instances x_i and x_j. $\mathbf{W}$ can be obtained either as a Nearest Neighbour (NN) graph (where each instance is connected to the k most similar instances in the graph, or to those with a distance under a radius ϵ). When empirically evaluating the proposed method, we used the dissimilarity relation among individuals within a DL knowledge base described in [11], since it does no constrain to any particular class of DLs.



3.2 Quadratic Cost Criteria

In quadratic cost criteria [3, ch. 11], the original label space $\{-1, +1\}$ (binary classification case) is relaxed to $[-1, +1]$. This allows to express the confidence associated to a labelling (and may give an indication about $P(Y \mid x)$). For such a reason, in the proposed method, the labelling functions space $\mathcal{F}$ will be relaxed to functions of the form $f : \text{Ind}_C(\mathcal{K}) \mapsto [-1, +1]$. Labelling functions can be equivalently represented as vectors $\mathbf{y} \in [-1, +1]^n$. Let $\hat{\mathbf{y}} \in [-1, +1]^n$ be a possible labelling for n instances. We can see $\hat{\mathbf{y}}$ as a $(l + u) = n$ dimensional vector, where the first l indices refer to already labelled instances, and the last u to unlabelled instances: $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_l, \hat{\mathbf{y}}_u]$.

Consistency of $\hat{\mathbf{y}}$ with respect to original labels can be formulated in the form of a quadratic cost: $\sum_{i=1}^{l} (\hat{y}_i - y_i)^2 = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2$. Similarly, labellings can be regularised with respect to the graph structure: as in [1], such consistency with respect to the geometry of instances can be estimated as $0.5 \sum_{i,j=1} \mathbf{W}_{ij} (\hat{y}_i - \hat{y}_j)^2 = \hat{\mathbf{y}}^T \mathbf{L} \hat{\mathbf{y}}$, where $\mathbf{W}$ is the semantic similarity graph and $\mathbf{L} = \mathbf{D} - \mathbf{W}$, $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$ ad 0 otherwise, is the unnormalized graph Laplacian. A different criterion, discussed in [13, 14], measures it as $(\mathbf{D}^{-0.5} \hat{\mathbf{y}})^T \mathbf{L} (\mathbf{D}^{-0.5} \hat{\mathbf{y}})$. Another regularization term in the form of $||\hat{\mathbf{y}}||^2$ (or $||\hat{\mathbf{y}}_u||^2$, as in [13]) can be added to the final cost function to prefer smaller values in $\hat{\mathbf{y}}$.

Putting the pieces together, we obtain two quadratic cost criteria discussed in the literature, namely Regression on Graph [1] (RG) and the Consistency Method [13] (CM):

RG: $cost(\hat{\mathbf{y}}) = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2 + \mu \hat{\mathbf{y}}^T \mathbf{L} \hat{\mathbf{y}} + \mu \epsilon ||\hat{\mathbf{y}}||^2$;

CM: $cost(\hat{\mathbf{y}}) = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2 + \mu (\mathbf{D}^{-0.5} \hat{\mathbf{y}})^T \mathbf{L} (\mathbf{D}^{-0.5} \hat{\mathbf{y}}) + ||\hat{\mathbf{y}}_u||^2$.

This work proposes using quadratic cost criteria as a solution to the transductive class-membership prediction problem. Finding a minimum $\hat{\mathbf{y}}$ for a predefined cost criterion is equivalent to finding a labelling function f^* in the form $f^* : \text{Ind}_C(\mathcal{K}) \mapsto [-1, +1]$, where the labelling returned for a generic training individual $a \in Ind_C(\mathcal{K})$ correspond to the value in $\hat{\mathbf{y}}$ in the position mapped to a. An advantage of quadratic cost criteria is that their minimization reduces to solving a large sparse linear system [13, 3], a problem in whose time complexity is nearly linear in the number of non-zero entries in the coefficient matrix [3, ch. 11].

4. Preliminary Empirical Evaluations

Following the procedure in [11], we evaluated the proposed approaches based on graph regularization and quadratic criteria with Soft-Margin SVM (discussed in [11] to induce robust classifiers from formal ontologies) and its SSL extension Laplacian SVM [3, ch. 12].

Leo	Match	Omission	Commission	Induction
RG	1 ± 0	0 ± 0	0 ± 0	0 ± 0
CM	1 ± 0	0 ± 0	0 ± 0	0 ± 0
SM-SVM	0.963 ± 0.1	0 ± 0	0.037 ± 0.1	0 ± 0
LapSVM	0.978 ± 0.068	0 ± 0	0.022 ± 0.068	0 ± 0
BioPAX Proteomics	Match	Omission	Commission	Induction
RG	0.986 ± 0.051	0.004 ± 0.028	0.008 ± 0.039	0.002 ± 0.02
CM	0.986 ± 0.051	0.002 ± 0.02	0.01 ± 0.044	0.002 ± 0.02
SM-SVM	0.972 ± 0.075	0 ± 0	0.026 ± 0.068	0.002 ± 0.02
LapSVM	0.972 ± 0.075	0 ± 0	0.026 ± 0.068	0.002 ± 0.02
MDM0.73	Match	Omission	Commission	Induction
RG	0.953 ± 0.063	0.003 ± 0.016	0.011 ± 0.032	0.015 ± 0.039
CM	0.953 ± 0.063	0.001 ± 0.009	0.013 ± 0.036	0.018 ± 0.04
SM-SVM	0.793 ± 0.252	0 ± 0	0.174 ± 0.255	0.033 ± 0.054
LapSVM	0.915 ± 0.086	0 ± 0	0.052 ± 0.065	0.033 ± 0.054
Wine	Match	Omission	Commission	Induction
RG	0.24 ± 0.03	0 ± 0.005	0.007 ± 0.017	0.5 ± 0.176
CM	0.242 ± 0.028	0 ± 0.005	0.005 ± 0.015	0.326 ± 0.121
SM-SVM	0.235 ± 0.036	0 ± 0	0.012 ± 0.024	0.753 ± 0.024
LapSVM	0.238 ± 0.033	0 ± 0	0.009 ± 0.021	0.753 ± 0.024

References

[1] Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In John Shawe-Taylor and Yoram Singer, editors, *COLT*, volume 3120 of *Lecture Notes in Computer Science*, pages 624–638. Springer, 2004.

[2] Stephan Bloehdorn and York Sure. Kernel methods for mining instance data in ontologies. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 58–71, Berlin, Heidelberg, 2007. Springer.

[3] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[4] Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito. Query answering and ontology population: an inductive approach. In Manfred Hauswirth, Manolis Koubarakis, and Sean Bechhofer, editors, *Proceedings of the 5th European Semantic Web Conference (ESWC'08)*, Tenerife, Spain, 2008. Springer.

[5] Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. Reduce: A reduced coulomb energy network method for approximate classification. In *Proceedings of the 6th European Semantic Web Conference (ESWC'09)*, pages 323–337. Springer.

[6] Nicola Fanizzi, Claudia D'Amato, and Floriana Esposito. Statistical learning for inductive query answering on owl ontologies. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, pages 195–212, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. Towards learning to rank in description logics. In Helder Coelho, Rudi Studer, and Michael Wooldridge, editors, *ECAI*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 985–986. IOS Press, 2010.

[8] Pascal Hitzler and Frank van Harmelen. A reasonable semantic web. *Semantic Web*, 1(1-2):39–44, 2010.

[9] Julia Lasserre and Christopher M. Bishop. Generative or discriminative? getting the best of both worlds. *BAYESIAN STATISTICS*, 8:3–24, 2007.

[10] José Eduardo Ochoa-Luna and Fabio Gagliardi Cozman. An algorithm for learning with probabilistic description logics. In Fernando Bobillo, Paulo Cesar G. da Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Trevor Martin, Matthias Nickles, Michael Pool, and Pavel Smrz, editors, *URSW*, pages 63–74, 2009.

[11] Achim Rettinger, Uta Lösch, Volker Tresp, Claudia d'Amato, and Nicola Fanizzi. Mining the semantic web - statistical learning for next generation knowledge bases. *Data Mining and Knowledge Discovery - Special Issue on Web Mining*, 2012.

[12] Achim Rettinger, Matthias Nickles, and Volker Tresp. Statistical relational learning with formal ontologies. In Wray L. Buntine, Marko Grobelnik, Dunja Mladenic, and John Shawe-Taylor, editors, *ECML/PKDD (2)*, volume 5782 of *LNCS*, pages 286–301. Springer, 2009.

[13] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, 2004.

[14] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 1036–1043, New York, NY, USA, 2005. ACM.

[15] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.