Transductive Inference for Class-Membership Propagation in Web Ontologies

Pasquale Minervini, Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito

LACAM, Dipartimento di Informatica — Università degli Studi di Bari "Aldo Moro"
via E. Orabona, 4 - 70125 Bari - Italia
firstname.lastname@uniba.it

Abstract. The increasing availability of structured machine-processable knowledge in the context of the Semantic Web, allows for inductive methods to back and complement purely deductive reasoning in tasks where the latter may fall short. This work proposes a new method for similarity-based class-membership prediction in this context. The underlying idea is the *propagation* of class-membership information among similar individuals. The resulting method is essentially non-parametric and it is characterized by interesting complexity properties, that make it a candidate for the application of transductive inference to large-scale contexts. We also show an empirical evaluation of the method with respect to other approaches based on inductive inference in the related literature.

1 Introduction

Standard reasoning services for the Semantic Web (SW) often rely on deductive inference. However, sometimes purely deductive approaches may suffer from limitations owing to the relative complexity of reasoning tasks, the inherent incompleteness of the knowledge bases and the occurrence of logically conflicting (incorrect) pieces of knowledge therein.

Approximate approaches based on both deductive and inductive inference have been proposed as a possible solutions to these limitations. In particular, various methods extend inductive learning techniques to tackle SW representations that are ultimately based on Description Logics (DL): they perform some sort of approximate reasoning efficiently by predicting assertions which were not derivable (or refutable) from the knowledge base and even coping with potential cases of inconsistency, since they are essentially data-driven (see [14], for a recent survey). Approximate data-driven forms of class-membership prediction could be useful for addressing cases such as the one illustrated in Ex. 1:

Example 1 (Academic Citation Network). Let us consider a knowledge base representing a *Bibliographic Citation Network* where papers, venues and authors are linked by relations such as writtenBy, publishedIn and citedBy. Assuming that specializations of paper based on the topics are also given, e.g. by means of disjoint classes such as MachineLearningPaper and DatabasePaper, one may want to ascertain the membership of an instance (a new paper) to either class. Owing to the Open-world assumption which is typically made when reasoning with SW representations, this task may not lead to a definite (positive or negative) conclusion in absence of explicit assertions.

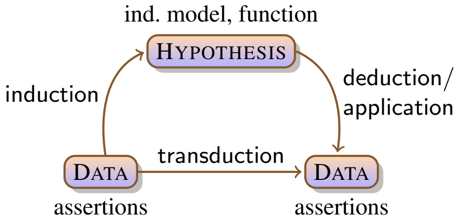


Fig. 1: Transductive and inductive inference

Bridging the gap caused by missing data can be cast as a *statistical learning* problem [18] for which efficient solutions may be found by adapting techniques proposed in the related literature. In principle, they may serve for completions even for large and Web-scale knowledge bases.

A variety of approaches to the class-membership prediction problem have been proposed in the literature. Among the various approaches, *discriminative* methods proposed so far tend to ignore unlabeled instances (individuals for which the target value of such class-membership is unknown); however, accounting for unlabeled instances during learning can provide more accurate results if some conditions are met [3]. *Generative* methods, on the other hand, try to model a joint probability distribution on both instances and labels, thus facing a possibly harder learning problem than only predicting the most probable membership for any given instance.

Several approaches to the class-membership prediction problem belong to the former category. They are often based on a notion of similarity, such as the k-Nearest Neighbors (k-NN) algorithm applied to DL knowledge bases [4]. A variety of similarity (and dissimilarity) measures between either individuals or concepts have been proposed [5]: some are based on *features* and objects are described in terms of a set of them (e.g. see [9]), some on a *semantic-network* structure that provides a form of background information (e.g. see [10]), while some rely on the *information content* (where both the semantic network structure and population are considered). Kernel-based algorithms have been proposed for various learning tasks from DL-based representations. This is made possible by the existence of a variety of kernel functions, either for concepts or individuals (e.g. see [6, 2, 14]). By (implicitly) projecting instances into a high-dimensional feature space, kernel functions allow to adapt a multitude of machine learning algorithms to structured representations. SW literature also includes methods for inducing classifiers from DL knowledge bases using some sort of RBF networks [7].

Also, methods based on a generative approach to learning have been proposed. In [15], 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 [13], which focuses on learning theories in a probabilistic extension of the $\mathcal{ALC}$ DL named CR$\mathcal{ALC}$. Extending our previous work [12], we propose a novel *transductive inference* method to be applied to class-membership prediction problem with knowledge bases expressed in standard SW representations. The nature of transductive inference, as opposed to induction, is illustrated in Fig. 1. Induction essentially generalizes existing data constructing an intermediate hypothesis (e.g. a classification function) that allows for making predictions on arbitrary individuals by deduction from the hypothesis (i.e. applying the induced classifier); transduction aims at propagating information on class-membership from the individuals for which membership is explicitly known towards those for which this information is missing (i.e. predicting new assertions), exploiting some notion of similarity among individuals (with *smooth variations*). Note that no generalization is made in this case.

Example 2 (Academic Citation Network, cont'd). It may be quite expensive to inductively build an inductive classifier that, given an arbitrary previously unseen paper, outputs the class of papers representing its specific topic. If one assumes that the citedBy relation can be associated to an indicator that two papers are likely to deal with the same topics or, similarly, that the same is likely to hold for papers written by the same author, transductive inference may be exploited to find a topic (i.e. a class-membership) assignment which varies smoothly among similar papers, and is consistent with the membership of examples provided by some domain expert.

In this work, we propose a method for spreading class-membership information among individuals for which this information is neither explicitly available nor derivable through deductive reasoning. This is accomplished by first constructing a *semantic similarity graph* encoding similarity relations among individuals. Then, class-membership information is propagated by minimizing a cost function grounded on a graph-based regularization approach. The remainder of the paper is organized as follows. In Sect. 2, transductive inference and the corresponding variant to the classic class-membership prediction problem are defined. In Sect. 3 we describe the proposed method, the assumptions it relies on, and how it can be used for class-membership prediction also on larger knowledge bases. In Sect. 4, we provide empirical evidence for the effectiveness of the proposed transductive class-membership propagation method in comparison with other methods in literature. In Sect. 5 we provide a brief summary of this work and about further developments of the proposed method.

2 Preliminaries

In the following, instances are described by features ranging in a certain space X and their classification with respect a given concept is indicated by labels in Y. In a probabilistic setting, instances are assumed to be sampled i.i.d. from an unknown joint probability distribution P ranging over $X \times Y$; *generative* methods are characterized by building an estimate $\hat{P}$ of $P(X, Y)$ from a given sample of instances, that is used to infer $\hat{P}(Y \mid x) = \hat{P}(Y, x)/\hat{P}(x)$ for some instance $x \in X$ whose unknown label is to be predicted. On the other hand, *discriminative* methods focus on conditional distributions to identify $\arg\max_y P(y \mid x)$, for any given $(x, y) \in X \times Y$, that is an easier problem than estimating the joint probability distribution.

2.1 Semi-Supervised Learning and Transductive Inference

Classic learning methods tend to ignore unlabeled instances. However, real-life scenarios are usually characterized by an abundance of unlabeled instances and a few labeled ones. This is also the case of class-membership prediction problem from formal ontologies: explicit class-membership assertions may be difficult to obtain during ontology engineering tasks (e.g. due to availability of domain experts) and inference (e.g. since deciding instance-membership may have an intractable time complexity with knowledge bases described by expressive Web-ontology languages).

Making use of unlabeled instances during learning is commonly referred to in literature as *Semi-Supervised Learning* [3] (SSL). A variant of this setting known as *Transductive Learning* [18] refers to finding a labeling only to unlabeled instances provided in the training phase, without necessarily generalizing to further unseen instances, resulting in a possibly *simpler* learning problem [18]. If the marginal distribution of instances P_X is informative w.r.t. the conditional distribution $P(Y \mid x)$, accounting for unlabeled instances during learning can provide more accurate results [3]. A possible approach is including terms dependent on P_X into the objective function.

The method proposed in this work relies on the so-called *semi-supervised smoothness assumption* [3]: *if two instances $x_i, x_j \in X$ in a high-density region are close then so should be the corresponding labels $y_i, y_j \in Y$*. Learning smooth labeling functions, this can be exploited by transitivity along paths of high density.

We will face a slightly different version of the classic class-membership prediction problem, namely *transductive class-membership prediction*. It is inspired by the *Main Principle* [18]: “If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem”. In this setting, the learning algorithm only aims at estimating the class-membership relation of interest for a given training set of individuals, without necessarily being able to generalize to instances outside this sample. In this work, transduction and induction differ in the target of the regularization: the latter would target the hypothesis (i.e. the inductive model), while the former targets directly the results of predictions.

2.2 Transductive Class-Membership Learning Problem in DL

Transductive class-membership learning with DL knowledge bases can be formalized as a cost minimization problem: given a set of training individuals $\mathrm{Ind}_C(\mathcal{K})$ whose class-membership w.r.t. a target concept C is either known or unknown, find a function $f^* : \mathrm{Ind}_C(\mathcal{K}) \rightarrow \{+1, -1\}$ 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. More formally:

Definition 1 (Transductive Class-Membership Learning).

- **Given:**
  - *a* target *concept C in a knowledge base $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$;*
  - *a set of training individuals* $\mathrm{Ind}_C(\mathcal{K}) \subseteq \mathrm{Ind}(\mathcal{A})$ *in $\mathcal{K}$ partitioned, according to their membership w.r.t. C, into the following sets:*
    - ∗ $\mathrm{Ind}^+_C(\mathcal{K}) = \{a \in \mathrm{Ind}_C(\mathcal{K}) \mid \mathcal{K} \models C(a)\}$ *positive examples,*
    - ∗ $\mathrm{Ind}^-_C(\mathcal{K}) = \{a \in \mathrm{Ind}_C(\mathcal{K}) \mid \mathcal{K} \models \neg C(a)\}$ *negative examples,*

    - $*$ $\mathrm{Ind}^0_C(\mathcal{K}) = \{a \in \mathrm{Ind}_C(\mathcal{K}) \mid \mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a)\}$ *unlabeled examples (i.e. whose concept-membership relation w.r.t. C is unknown);*
  - *A* cost function *cost*$(\cdot) : \mathcal{F} \mapsto \mathbb{R}$*, specifying the* cost *associated to labeling functions $f \in \mathcal{F}$ of the form* $\mathrm{Ind}_C(\mathcal{K}) \mapsto \{+1, -1\}$;
- **Find** $f^* \in \mathcal{F}$ *minimizing cost*$(\cdot)$ *w.r.t. the training individuals in* $\mathrm{Ind}_C(\mathcal{K})$*:*

$$f^* \leftarrow \arg\min_{f \in \mathcal{F}} \mathit{cost}(f).$$

The function f^* determined by a proper transductive class-membership learning method can then be used to predict class-membership relations w.r.t. the target concept C for all training individuals (including those in $\mathrm{Ind}^0_C(\mathcal{K})$): it will return $+1$ (resp. -1) if an individual is likely to be a member of C (resp. $\neg C$). Note that the function is defined on the whole set of training individuals but it is not a generalization stemming from them; therefore, possibly, it may contradict class-membership assertions that are already available (thus being able to handle noisy knowledge). Since $\mathrm{Ind}_C(\mathcal{K})$ is finite, the space of labeling functions $\mathcal{F}$ is also finite, and each function $f \in \mathcal{F}$ can be equivalently expressed as a vector in $\{-1, +1\}^n$, where $n = |\mathrm{Ind}_C(\mathcal{K})|$.

In order to solve this problem, we propose a similarity-based, non-parametric and computationally efficient method for predicting missing class-membership relations. This method is essentially discriminative, and may account for unknown class-membership relations during learning.

3 Propagating Class-Membership Information Among Individuals

A transductive method based on *graph-regularization*[1] [3] is presented allowing for class-membership prediction with knowledge bases expressed in DL. The method relies on a weighted *semantic similarity graph*, where nodes represent positive, negative and unlabeled examples of the transductive class-membership prediction problem, and weighted edges define similarity relations among such individuals.

Given an instance of the *transductive class-membership learning problem* (see Def. 1), the approach proposed in this work is outlined in Alg. 1 and summarized by the following basic steps:

1. Given a class-membership prediction task and a set of training individuals (either labeled and unlabeled), create an undirected *semantic similarity graph* (SSG) where two individuals are linked iff they are considered *similar* (that is, their class-membership is not likely to change from one individual to another).
2. Propagate class-membership information among similar individuals (transduction step), by minimizing a cost function based on a graph regularization approach (where the graph is given by the SSG) and defined over possible class-membership relations for training individuals.

[1] In brief, *regularization* consists in introducing additional terms to an objective function to be optimized to prevent overfitting. These terms add usually some penalty for complexity and have the form of restrictions for smoothness, bounds on the vector space norm or number of model parameters.

Algorithm 1 Transductive Class-Membership Prediction via Graph-Based Regularization with the Semantic Similarity Graph

```
Input: Initial class-membership relations $\mathrm{Ind}_C^{+}(\mathcal{K})$, $\mathrm{Ind}_C^{-}(\mathcal{K})$ and $\mathrm{Ind}_C^{0}(\mathcal{K})$ w.r.t. a concept $C$
    and a knowledge base $\mathcal{K}$;
Output: $f^* \in \mathcal{F}$
    {Compute the Semantic Similarity Graph (SSG) $G$, encoding neighborhood relations among
    individuals in $\mathrm{Ind}_C(\mathcal{K})$.}
    $G \leftarrow \mathit{semanticSimilarityGraph}(\mathrm{Ind}_C(\mathcal{K}))$;
    {Minimize a cost function $\mathit{cost}$ defined over a set of labeling functions $\mathcal{F}$. The cost function
    is based on the SSG $G$ and enforces smoothness in class-membership relations among similar
    individuals as well as consistency with initial class-membership relations.}
    $f^* \leftarrow \arg\min_{f \in \mathcal{F}} \mathit{cost}(f, G, \mathit{Ind}_C(\mathcal{K}))$;
    return $f^*$;
```

This method can be seen as inducing a new metric, in which neighborhood relations among training individuals are preserved; and then, performing classic supervised learning using the new distance.

Example 3 (Academic Citation Network (cont.d)). Assuming that papers written by the same authors or cited by the same articles (where such information is encoded by the *writtenBy* and *citedBy* roles respectively) have a tendency to have similar domain-memberships, we can construct a SSG in which each paper is linked to its k most similar papers, and rely on this structure to propagate domain-membership information.

In the following, the procedure for building a SSG among individuals in the training set $\mathrm{Ind}_C(\mathcal{K})$ is illustrated. As regards the labeling process of unlabeled training examples, namely the transductive step, a optimal labeling function f^* has to be found by minimizing a given cost function. For defining a cost over the space of the labeling functions $f \in \mathcal{F}$, the proposed method (see Sect. 3.2) aims at finding a labeling function that is both consistent with the given labels, and changes smoothly between similar instances (where similarity relations are encoded in the SSG). This is formalized through a *regularization by graph framework*, using the loss function as a measure of consistency to the given labels, and a measure of smoothness among the similarity graph as a regularizer.

3.1 Semantic Similarity Graph

A similarity graph for a set of training examples is a graph where the set of nodes is given by the training examples and edges between nodes connect similar training examples with respect to a given similarity measure. Edges are labeled with the corresponding computed similarity values.

A similarity graph can be modeled as a weighted adjacency matrix $\mathbf{W}$ (or, briefly, weight matrix), where $\mathbf{W}_{ij}$ represents the similarity value of x_i and x_j. Specifically, $\mathbf{W}$ is often obtained as a k-Nearest Neighbor (NN) graph [3] where each instance is connected to the k most similar instances in the graph, or to those with a similarity value above a given threshold ϵ, while the remaining similarity values are set to 0.

For building such a similarity graph given the individuals in $\mathrm{Ind}_C(\mathcal{K})$, a solution is relying on the family of dissimilarity[2] measures defined in [14], since they do not constrain to any particular family of DLs. Since this measure is a *semantic similarity measures*, following the formalization in [5], we call the resulting similarity graph as the *semantic similarity graph* (SSG).

The adopted dissimilarity measure is briefly recalled in the following. Given a set of concept descriptions $F = \{F_1, \ldots, F_n\}$ in $\mathcal{K}$ and a weight vector $\mathbf{w} = (w_1, \ldots, w_n)$, the family of dissimilarity measures $d_p^F : Ind(\mathcal{K}) \times Ind(\mathcal{K}) \mapsto [0, 1]$ is defined as:

$$d_p^F(x_i, x_j) = \left[\sum_{k=1}^{|F|} w_k |\delta_k(x_i, x_j)|^p \right]^{\frac{1}{p}} \tag{1}$$

where $p > 0$, $Ind(\mathcal{K})$ is the set of all individuals in the knowledge base $\mathcal{K}$, $x_i, x_j \in Ind(\mathcal{K})$ and $\forall k \in \{1, \ldots, n\}$ results:

$$\delta_k(x_i, x_j) = \begin{cases} \;0 \text{ if } (\mathcal{K} \models F_i(x) \land \mathcal{K} \models F_i(y)) \lor (\mathcal{K} \models \neg F_i(x) \land \mathcal{K} \models \neg F_i(y)) \\ \;1 \text{ if } (\mathcal{K} \models F_i(x) \land \mathcal{K} \models \neg F_i(y)) \lor (\mathcal{K} \models \neg F_i(x) \land \mathcal{K} \models F_i(y)) \\ u_k \text{ otherwise} \end{cases}$$

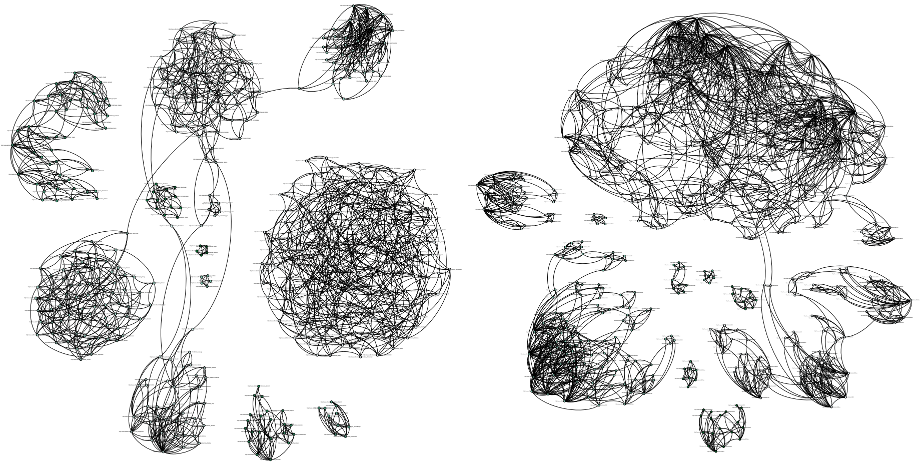
where u_k can reflect the degree of uncertainty on the membership w.r.t the k-th feature in the concept committee [14]. We proposed such a measure in our previous work [12] for building the SSG among a set of individuals in a knowledge base. Such a dissimilarity measure can be used to obtain a kernel function among individuals by simply turning it into a similarity measure [14].

An alternative approach for obtaining the SSG among a set of individuals in a knowledge base, by relying more on the corresponding network structure, is by means of *graph* and RDF kernels: a kernel provides an (implicitly) mapping for individuals into an embedding space, by calculating their inner product. A recently proposed kernel for RDF data is the Full SubTree (FST) kernel [11].

Let $k : Ind(\mathcal{K}) \times Ind(\mathcal{K}) \to \mathbb{R}$ be a kernel function defined over individuals in a knowledge base $\mathcal{K}$. Since k corresponds to an embedding function ϕ mapping individuals to points in an embedding space, that is $\forall x_i, x_j \in Ind(\mathcal{K}) : k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, it is immediate to derive the Euclidean distance in the embedding space among two individuals [16]: $||\phi(x_i) - \phi(x_j)|| = \sqrt{k(x_i, x_i) - 2k(x_i, x_j) + k(x_j, x_j)}$.

Two examples of k-NN SSGs among individuals in the AIFB Affiliations ontology (representing instances of the concepts Person and Article), which is also used in empirical evaluations in Sect.4, are shown in Fig. 2. In both cases, a clustered structure emerges from the graphs. In the case of the SSG modeling the similarity relations among instances of the Person concept, an highly connected subgraph groups persons working in the EOrg research group; another connected component (composed by two highly connected subgraphs) groups persons in the BIK research group; two connected components group persons affiliated to the CoM research group; and three single connected components group respectively persons with no available affiliation (the larger

[2] A dissimilarity measure $d \in [0, 1]$ can be transformed in a similarity measure $s = 1 - d$ [5].



(a) AIFB Affiliations (Persons)

(b) AIFB Affiliations (Articles)

Fig. 2: Semantic Similarity Graphs for individuals representing persons and articles in the AIFB Affiliations ontology (5-NN graphs obtained using the Full SubTree kernel [11] with parameters $d = 1$ and $\lambda = 0.9$)

component) and affiliated to the WBS and EffAlg research groups. Also instances of the Article concept tend to be grouped into different components of their SSG. Similarly to the previous example, articles tend to be grouped according to their research group affiliation, such as CoM or EffAlg. However, some articles affiliated to different research group share one or more authors, causing the presence of a few connections among the different clusters.

In this work, we propose to leverage such emerging structures in class-membership prediction tasks. The underlying idea is to *propagate* class-membership information among similar individuals, assuming that such information tends not to vary within regions of the instance space with an high density of instances (due to the semi-supervised smoothness assumption discussed in Sect. 2).

3.2 Transductive Inference via Quadratic Cost Criteria

In this section the transductive step is illustrated. It basically consists in labeling the unlabeled training examples. For doing this, a optimal labeling function f^* has to be found by minimizing a given cost function (see Def. 1). For determining a cost over the space of the labeling functions $f \in \mathcal{F}$, the method finds a function that is: 1) consistent with the given labels; 2) changes smoothly between similar instances (encoded in the semantic similarity graph). The first issue is addressed by adopting the loss function as a measure of consistency with respect to the given labels. The second issue is addressed by regularizing the labeling of the function with respect to the structure of the semantic similarity graph.

For addressing the consistency issue, the quadratic cost criteria [3, ch. 11] are considered where the adopted label space $\{-1, +1\}$ is the one for the binary classification case. We relax this label space to the interval $[-1, +1]$ that allows to express the confidence associated to a labeling. Consequently, also the labeling functions space $\mathcal{F}$ is relaxed to functions of the form $f : \mathrm{Ind}_C(\mathcal{K}) \mapsto [-1, +1]$. Labeling functions can be equivalently represented as vectors $\mathbf{y} \in [-1, +1]^n$ where n is the number of the training examples. Let $\hat{\mathbf{y}} \in [-1, +1]^n$ be a possible labeling for n instances. $\hat{\mathbf{y}}$ can be seen as a $(l + u) = n$ dimensional vector, where the first l indices refer to already labeled instances, and the last u to unlabeled instances: $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_l, \hat{\mathbf{y}}_u]$. The consistency of $\hat{\mathbf{y}}$ with respect to the original labels is then formulated in the form of a quadratic cost: $\sum_{i=1}^{l} (\hat{y}_i - y_i)^2 = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2$.

To regularize the labellings with respect to the graph structure, the *graph Laplacian* [3] can be exploited. Let $\mathbf{W}$ be the weight matrix corresponding to the similarity graph G, and let $\mathbf{D}$ be the diagonal matrix obtained from $\mathbf{W}$ as $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$ namely by summing the elements in each column of $\mathbf{W}$. Hence, two alternative definitions for the graph Laplacian can be considered [3]:

– Unnormalized graph Laplacian: $\mathbf{L} = \mathbf{D} - \mathbf{W}$;
– Normalized graph Laplacian: $\mathcal{L} = \mathbf{D}^{-0.5}\mathbf{L}\mathbf{D}^{-0.5} = \mathbf{I} - \mathbf{D}^{-0.5}\mathbf{W}\mathbf{D}^{-0.5}$.

Following [1], a possible graph-based regularization factor is $0.5 \sum_{i,j=1}^{n} \mathbf{W}_{ij} (\hat{y}_i - \hat{y}_j)^2 = \hat{\mathbf{y}}^T \mathbf{L} \hat{\mathbf{y}}$; in alternative it is possible to resort to the normalized graph Laplacian [19, 20], using the slightly different regularization factor $\hat{\mathbf{y}}^T \mathcal{L} \hat{\mathbf{y}}$.

For preventing overfitting, an additional regularization term, in the form of $||\hat{\mathbf{y}}||^2$ (or $||\hat{\mathbf{y}}_u||^2$, as in [19]), can be added. This additional low norm regularizer on $\hat{\mathbf{y}}$ helps avoiding overfitting and preventing arbitrary labellings in connected components of the semantic similarity graphs containing only unlabeled instances.

Putting the pieces together, two quadratic cost criteria in the form proposed in the literature are obtained, namely Regularization on Graph [1] (RG) and Consistency Method [19] (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:** $cost(\hat{\mathbf{y}}) = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2 + \mu \hat{\mathbf{y}}^T \mathcal{L} \hat{\mathbf{y}} + ||\hat{\mathbf{y}}_u||^2$.

Once the form of the cost function is determined, the minimum for the function has to be found. As a title of example, a closed form solution for the problem of finding a (global) minimum for the quadratic cost criterion in RG is showed.

Let $\mathbf{S}$ be the diagonal matrix $\mathbf{S} = diag(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ obtained by setting $\mathbf{s}_i = 1$ iff $i \leq l$ and 0 otherwise. The first order derivative for the case of the cost function in RG can be written as:

$$\frac{1}{2} \frac{\partial cost(\hat{\mathbf{y}})}{\partial \hat{\mathbf{y}}} = (\mathbf{S} + \mu \mathbf{L} + \mu\epsilon \mathbf{I})\hat{\mathbf{y}} - \mathbf{S}\mathbf{y}.$$

The second order derivative is a positive definite matrix if $\epsilon > 0$, since $\mathbf{L}$ is positive semi-definite. Hence, setting the first order derivative to 0 leads to a global minimum:

$$\hat{\mathbf{y}} = (\mathbf{S} + \mu \mathbf{L} + \mu\epsilon \mathbf{I})^{-1} \mathbf{S}\mathbf{y},$$

showing that $\hat{\mathbf{y}}$ can be obtained either by matrix inversion or by solving a (possibly sparse) linear system.

In this way, this work leverages quadratic cost criteria to efficiently solve the transductive class-membership prediction problem. An advantage of quadratic cost criteria is that their minimization ultimately reduces to solving a large sparse linear system [19, 3], a well-known problem in the literature whose time complexity is nearly linear in the number of non-zero entries in the coefficient matrix [17]. For large-scale datasets, a subset selection method is described in [3, ch. 18], which allows to greatly reduce the size of the original linear system.

4 Empirical Evaluation

In this section, we evaluate several (inductive and transductive) methods for class-membership prediction, with the aim of comparing the methods discussed in Sect. 3 with respect to other methods in the SW literature.

Specifically, we empirically compared a set of different methods for the class-membership prediction task. Those can be partitioned in transductive (Regularization on Graph [1] (RG), Consistency Method [19] (CM) and Label Propagation [21] (LP)) and inductive (Soft-Margin Support Vector Machines with L_1 norm (SM-SVM) and $\sqrt{l}$-Nearest Neighbors). Such inductive approaches have also been discussed in the task of class-membership prediction in [14], and previously in the context of inducing robust classifiers from ontological knowledge bases [8]. Implementations for the evaluated methods, as well as the dataset used in this work, are available online [3].

4.1 Evaluated Methods

LP is a graph-based transductive inference algorithm relying on the idea of propagating labeling information among similar instances through an iterative process involving matrix operations. It can be equivalently formulated under the quadratic criterion framework [3, ch. 11]. More formally it associates, to each unlabeled instance in the graph, the probability of performing a random walk until a positively (resp. negatively) example is found. Support Vector Machine classifiers, on the other hand, come in different flavors: the classic (Hard-Margin) SVM binary classifier aims at finding the hyperplane in the feature space separating the instances belonging to different classes, which maximizes the *geometric margin* between the hyperplane and nearest training points. The SM-SVM relaxes this method, by allowing for some misclassification in training instances (by relaxing the need of having perfectly linearly separable training instances in the feature space). We adopted this latter solution to handle the lack of perfect linear separability of the instances belonging to different classes. Note that the aforementioned methods can be seen as relying on a *change of representation*: instances of the prediction problem are represented as points in an embedding space, and implicitly described by means of their pairwise Euclidean distances, inner products (as in the case of kernel-based methods, such as SVM) or neighborhood relations. We evaluated dif-

[3] At the address `http://lacam.di.uniba.it/~nico/research/ontologymining.html`

Table 1: Results for a 10-fold cross validation obtained when predicting the affiliations of AIFB staff members to research groups, using the **Atomics** kernel (and the corresponding dissimilarity measure)

EffAlg	Match	Omission	Commission	F1
LP+Atomics	0.53 ± 0.189	0 ± 0	0.47 ± 0.189	0.488 ± 0.217
RG+Atomics	0.458 ± 0.166	0.01 ± 0.032	0.532 ± 0.158	0.405 ± 0.194
SM-SVM+Atomics	0.6 ± 0.125	0 ± 0	0.4 ± 0.125	0.555 ± 0.198
$\sqrt{l}$-NN+Atomics	0.5 ± 0	0 ± 0	0.5 ± 0	0.667 ± 0

CoM	Match	Omission	Commission	F1
LP+Atomics	0.533 ± 0.317	0 ± 0	0.467 ± 0.317	0.419 ± 0.39
RG+Atomics	0.475 ± 0.294	0 ± 0	0.525 ± 0.294	0.36 ± 0.333
SM-SVM+Atomics	0.517 ± 0.207	0 ± 0	0.483 ± 0.207	0.403 ± 0.31
$\sqrt{l}$-NN+Atomics	0.5 ± 0.167	0 ± 0	0.5 ± 0.167	0.517 ± 0.277

BIK	Match	Omission	Commission	F1
LP+Atomics	0.502 ± 0.116	0.037 ± 0.064	0.46 ± 0.117	0.451 ± 0.176
RG+Atomics	0.531 ± 0.089	0.005 ± 0.014	0.464 ± 0.083	0.488 ± 0.147
SM-SVM+Atomics	0.514 ± 0.068	0 ± 0	0.486 ± 0.068	0.337 ± 0.214
$\sqrt{l}$-NN+Atomics	0.522 ± 0.072	0 ± 0	0.478 ± 0.072	0.404 ± 0.125

EOrg	Match	Omission	Commission	F1
LP+Atomics	0.667 ± 0.167	0 ± 0	0.333 ± 0.167	0.65 ± 0.146
RG+Atomics	0.692 ± 0.157	0 ± 0	0.308 ± 0.157	0.667 ± 0.136
SM-SVM+Atomics	0.692 ± 0.197	0 ± 0	0.308 ± 0.197	0.647 ± 0.286
$\sqrt{l}$-NN+Atomics	0.717 ± 0.185	0 ± 0	0.283 ± 0.185	0.713 ± 0.174

WBS	Match	Omission	Commission	F1
LP+Atomics	0.504 ± 0.069	0.012 ± 0.028	0.484 ± 0.072	0.489 ± 0.081
RG+Atomics	0.512 ± 0.09	0 ± 0	0.488 ± 0.09	0.512 ± 0.101
SM-SVM+Atomics	0.603 ± 0.084	0 ± 0	0.397 ± 0.084	0.503 ± 0.131
$\sqrt{l}$-NN+Atomics	0.513 ± 0.097	0 ± 0	0.487 ± 0.097	0.522 ± 0.152

ferent choices for such change of representation, consisting in different choices for the (dis-)similarity measure used to construct the k-Nearest Neighborhood graph, and the kernel function. Specifically, we evaluated the following choices:

Atomics – a dissimilarity measure defined in [14] (outlined in Eq. 1) was used to construct the k-Nearest Neighborhood graph (with $p = 2$, using all atomic concepts in the ontology as features and weighting each concept with its associated entropy [14]). The corresponding kernel function was obtained as discussed in Sect. 3.

Full SubTree kernel (FST) – a kernel for RDF data proposed in [11]; it was used to construct a k-NN SSG as shown in Sect. 3. The optimal kernel parameters $(depth, \lambda)$ were found within the training set using a k-fold cross validation procedure (with $k = 10$), and varied in $\{1, 2\}$ and $\{0.1, 0.5, 0.9\}$ respectively.

4.2 Evaluation Procedure

Extending our previous results in [12], we are evaluating the proposed approach on a knowledge base in which a quantity of information is stored in the network structure rather than in the concept hierarchy. The empirical evaluation involved the metadata available in the Semantic Portal of the institute AIFB [4]. The ontology models key concepts within a research community: it comprises 44351 individuals and the Person,

[4] `http://www.aifb.kit.edu/web/Wissensmanagement/Portal`, as of 21 Feb. 2012

Table 2: Results for a 10-fold cross validation obtained when predicting the affiliations of AIFB staff members to research groups, using the **Full SubTree** kernel (and corresponding dissimilarity measure)

EffAlg	Match	Omission	Commission	F1
LP+FST	0.565 ± 0.167	0.09 ± 0.099	0.345 ± 0.201	0.611 ± 0.218
RG+FST	0.548 ± 0.154	0.08 ± 0.103	0.372 ± 0.187	0.58 ± 0.2
SM-SVM+FST	0.6 ± 0.125	0 ± 0	0.4 ± 0.125	0.587 ± 0.246
$\sqrt{l}$-NN+FST	0.57 ± 0.134	0 ± 0	0.43 ± 0.134	0.65 ± 0.129

CoM	Match	Omission	Commission	F1
LP+FST	0.617 ± 0.261	0.083 ± 0.136	0.3 ± 0.201	0.563 ± 0.35
RG+FST	0.583 ± 0.157	0.083 ± 0.136	0.333 ± 0.124	0.613 ± 0.106
SM-SVM+FST	0.55 ± 0.201	0 ± 0	0.45 ± 0.201	0.393 ± 0.298
$\sqrt{l}$-NN+FST	0.542 ± 0.148	0 ± 0	0.458 ± 0.148	0.575 ± 0.217

BIK	Match	Omission	Commission	F1
LP+FST	0.536 ± 0.107	0.077 ± 0.08	0.386 ± 0.114	0.556 ± 0.146
RG+FST	0.534 ± 0.13	0.06 ± 0.053	0.406 ± 0.13	0.53 ± 0.206
SM-SVM+FST	0.609 ± 0.075	0 ± 0	0.391 ± 0.075	0.443 ± 0.162
$\sqrt{l}$-NN+FST	0.559 ± 0.074	0 ± 0	0.441 ± 0.074	0.423 ± 0.132

EOrg	Match	Omission	Commission	F1
LP+FST	0.692 ± 0.258	0.075 ± 0.121	0.233 ± 0.222	0.65 ± 0.388
RG+FST	0.725 ± 0.249	0.067 ± 0.11	0.208 ± 0.201	0.69 ± 0.33
SM-SVM+FST	0.792 ± 0.163	0 ± 0	0.208 ± 0.163	0.793 ± 0.152
$\sqrt{l}$-NN+FST	0.717 ± 0.185	0 ± 0	0.283 ± 0.185	0.713 ± 0.174

WBS	Match	Omission	Commission	F1
LP+FST	0.583 ± 0.09	0.07 ± 0.044	0.347 ± 0.09	0.591 ± 0.101
RG+FST	0.64 ± 0.064	0.033 ± 0.043	0.327 ± 0.065	0.606 ± 0.058
SM-SVM+FST	0.632 ± 0.091	0 ± 0	0.368 ± 0.091	0.629 ± 0.108
$\sqrt{l}$-NN+FST	0.467 ± 0.094	0 ± 0	0.533 ± 0.094	0.314 ± 0.189

Document and Project FOAF concepts (among others) are associated to respectively 509, 4731 and 128 individuals, and roles include affiliation relationships between persons and research groups, authorship relations between persons and documents, and other knowledge inherent to the academic domain. The knowledge base consists also in 312738 axioms, 49 classes, 96 object properties and 184 data properties, resulting in a $\mathcal{ALEHO}(\mathcal{D})$ knowledge base (encoded in a OWL 2 RL fragment). The learning task, as defined in [2], consisted in predicting affiliations of AIFB staff members to research groups, which we denoted as class-membership relations. All knowledge inherent affiliation relations to research group was removed from the ontology before the experiment. As in [11], negative examples were artificially created (in the same number as positive examples) to mend the lack of training data (due to the Open World Assumption).

A DL reasoner [5] was employed to decide on the concept-membership of individuals to query concepts to be used as a baseline. Performance is measured employing the evaluation indexes proposed in [4], which take into account the specificity deriving from the presence of missing knowledge in the assertions considered as the baseline:

Match Case of an individual that got the same label by the reasoner and the inductive classifier.

[5] Pellet v2.3.0 – `http://clarkparsia.com/pellet/`

Omission Error Case of an individual for which the inductive method could not determine whether it was relevant to the query concept or not while it was found relevant by the reasoner.

Commission Error Case of an individual found to be relevant to the query concept while it logically belongs to its negation or vice-versa.

To provide a term of comparison with methods and results in [2] and [11], we also provide results obtained by the F1-score metric (defined as the harmonic mean of precision and recall). Before evaluating on the test set, parameter tuning was performed for each of the methods via a k-fold cross validation ($k = 10$) within the training set, for finding the parameters with lower classification error in cross-validation. SM-SVM follows the implementation in [16, pg. 223]: the C parameter was allowed to vary in $\{10^{-4}, 10^{-3}, \ldots, 1\}$. The (μ, ϵ) parameters in RG and CM were respectively allowed to vary in $\{10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}\}$ and fixed to 10^{-4}. The number of neighbors for each node, needed for the construction of the SSG, was allowed to vary in $\{2, 3, 5, 7\}$.

RG, CM and LP give an indication of the uncertainty associated to a specific labeling by associating values in the set $[-1, +1]$ to each node; when such values are ≈ 0 (specifically, when the label was in the set $[-10^{-4}, 10^{-4}]$) we decided to leave the node unlabeled, so to try to provide more robust estimates (and thus a possibly lower commission error and match rates and higher omission error rates). This may happen e.g. when there are no labeled examples within a connected component of the SSG.

4.3 Discussion

From this empirical evaluation, it emerged that the Consistency Method (CM) discussed in Subsect. 3.2 (which we do not report in Tables 1,2 for brevity) may be too conservative: this was suggested by its low Match rate (always reported lower than 0.1) and high Omission rate (always reported higher than 0.9). This may be justified by the fact that its regularizer $||\hat{y}_u||$ is not weighted by any term, unlike Regularization on Graph (RG) (which weights the regularizer $||\hat{y}||$ by means of the term $\mu\epsilon$). The presence of such a regularization term influences the results of transductive methods. Inductive classification methods such as SVM and k-NN define straight decision boundaries in the instance space: a classification result may happen by chance.On the other hand, relaxing binary labels to continuous ones and pulling to 0 labels of unlabeled examples allows to provide more robust labellings: they will be less likely to be determined by chance, and more likely to be statistically justified.

Also from our previous work [12], the choice of the SSG strongly affects final results, and it is likely to be task-dependent: in this case, results obtained by using the Atomics kernel/dissimilarity measure were significantly worse than those obtained with the FST kernel. An explaination is that, in this knowledge base, (atomic) concept-membership relations tend not to carry much information w.r.t. the affiliation prediction task, while the network structure (exploited by the FST kernel) tends to be informative. For example, object properties encoding competence fields tend to encode homophily relations – persons sharing competence fields have a tendency to also have the same research group affiliation. A significant part of the classification error is caused by the fact that persons with not much available information other than their research group affiliation, are now clustered together with nodes where even such information is not available: this is of course non necessarily correct, since lack of information (given by the Open World Assumption) on both individuals does not necessarily imply the presence of a similarity relation between them. A graph kernel might capture similarity relations in case of full information (such as in the SSGs discussed in Sect. 3) but might have problems in case of missing information (such as in this case).

Co-authorship relations to articles, as discussed in Sect. 3, can also encode useful information; however, analysing the results, it emerges that such information is only available from the analysis of inverse roles, which have not been considered in our implementation of the FST kernel. It also emerges that potentially unuseful relations (such as shared first or last names) have concurred in estabilishing similarity relations among individuals. This suggests that simple graph or RDF kernel can fail exploiting the informativeness of potentially useful paths in the ontology's relational graph.

5 Conclusion and Future Work

This work proposes a method for transductive inference for class-membership prediction in Description Logic knowledge bases. It leverages unlabeled examples by propagating class-membership information among similar individuals in the knowledge base. The proposed method relies on graph regularization using quadratic cost criteria, whose optimization can be reduced to solving a (possibly sparse) linear system. In this work, we assumed information propagates homogeneously within the similarity graph defined over a set of individuals in the knowledge base. However, real world ontologies describe domains characterized by *heterogeneity*, either on individuals or on relations among them. For example, persons in the AIFB Affiliations ontology (see Sect. 4) can belong to different categories (e.g. according to their contract type) and be linked by multiple types of relations (for example, given by co-authored articles or shared competence fields), which can have a variable level of informativeness w.r.t. a specific prediction task. Considering multiple similarity measures boils down to defining a cost function with multiple graph-based regularizers, with the side effect of an increased number of parameters. In future work we aim at extending our approach to include multiple similarity relations among different types of instances, and working on methods to efficiently learn the regularization parameters.

References

[1] Belkin, M., Matveeva, I., Niyogi, P.: Regularization and semi-supervised learning on large graphs. In: Shawe-Taylor, J., et al. (eds.) Proceedings of the 17th Annual Conference on Learning Theory, COLT2004. LNCS, vol. 3120, pp. 624–638. Springer (2004)

[2] Bloehdorn, S., Sure, Y.: Kernel methods for mining instance data in ontologies. In: Proceedings of the 6th International Semantic Web Conference, ISWC'07. pp. 58–71. Springer (2007)

[3] Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press (2006)

[4] d'Amato, C., Fanizzi, N., Esposito, F.: Query answering and ontology population: an inductive approach. In: Hauswirth, M., et al. (eds.) Proceedings of the 5th European Semantic Web Conference, ESWC'08. Springer (2008)

[5] d'Amato, C., Staab, S., Fanizzi, N.: On the influence of description logics ontologies on conceptual similarity. In: Proceedings of the 16th international conference on Knowledge Engineering: Practice and Patterns, EKAW'08. pp. 48–63. Springer (2008)

[6] Fanizzi, N., d'Amato, C.: Inductive concept retrieval and query answering with semantic knowledge bases through kernel methods. In: Proceedings of the 11th international conferenceon Knowledge-based intelligent information and engineering systems, KES'07: Part I. pp. 148–155. Springer (2007)

[7] Fanizzi, N., d'Amato, C., Esposito, F.: Reduce: A reduced coulomb energy network method for approximate classification. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E.P.B. (eds.) ESWC. Lecture Notes in Computer Science, vol. 5554, pp. 323–337. Springer (2009)

[8] Fanizzi, N., d'Amato, C., Esposito, F.: Induction of robust classifiers for web ontologies through kernel machines. J. Web Sem. 11, 1–13 (2012)

[9] Hu, B., Dasmahapatra, S., Lewis, P.: Semantic metrics. Int. J. Metadata Semant. Ontologies 2(4), 242–258 (Jul 2007)

[10] Janowicz, K., Wilkes, M.: Sim-dla: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity. In: Aroyo, L., et al. (eds.) Proceedings of the 6th European Semantic Web Conference, ESWC2009. LNCS, vol. 5554, pp. 353–367. Springer (2009)

[11] Lösch, U., Bloehdorn, S., Rettinger, A.: Graph kernels for rdf data. In: Simperl, E., et al. (eds.) ESWC. Lecture Notes in Computer Science, vol. 7295, pp. 134–148. Springer (2012)

[12] Minervini, P., d'Amato, C., Fanizzi, N.: A graph regularization based approach to transductive class-membership prediction. In: Bobillo, F., et al. (eds.) Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web, URSW2012. CEUR Workshop Proceedings, vol. 900, pp. 39–50. CEUR-WS.org (2012)

[13] Ochoa-Luna, J.E., Cozman, F.G.: An algorithm for learning with probabilistic description logics. In: Bobillo, F., et al. (eds.) Proceedings of the 5th International Workshop on Uncertainty Reasoning for the Semantic Web, URSW09. CEUR Workshop Proceedings, vol. 654, pp. 63–74. CEUR-WS.org (2009)

[14] Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., Fanizzi, N.: Mining the Semantic Web: Statistical learning for next generation knowledge bases. Data Min. Knowl. Discov. 24(3), 613–662 (2012)

[15] Rettinger, A., Nickles, M., Tresp, V.: Statistical relational learning with formal ontologies. In: Buntine, W.L., et al. (eds.) Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML/PKDD'09. LNCS, vol. 5782, pp. 286–301. Springer (2009)

[16] Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)

[17] Spielman, D.A., Teng, S.H.: Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In: Proceedings of the 36th ACM Symposium on Theory of Computing, STOC'04. pp. 81–90. ACM (2004)

[18] Vapnik, V.N.: Statistical learning theory. Wiley, 1 edn. (Sep 1998)

[19] Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Thrun, S., Saul, L.K., Schölkopf, B. (eds.) NIPS. MIT Press (2003)

[20] Zhou, D., Huang, J., Schölkopf, B.: Learning from labeled and unlabeled data on a directed graph. In: Raedt, L.D., Wrobel, S. (eds.) ICML. ACM International Conference Proceeding Series, vol. 119, pp. 1036–1043. ACM (2005)

[21] Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Tech. rep., CMU CALD tech report CMU-CALD-02 (2002)