# Distance-based Classification in OWL Ontologies

Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito

Department of Computer Science, University of Bari
{claudia.damato|fanizzi|esposito}@di.uniba.it

**Abstract.** We propose inductive distance-based methods for instance classification and retrieval in ontologies. Casting retrieval as a classification problem with the goal of assessing the individual class-memberships w.r.t. the query concepts, we propose an extension of the *k-Nearest Neighbor* algorithm for OWL ontologies based on an *epistemic* distance measure. The procedure can classify the individuals w.r.t. the known concepts but it can also be used to retrieve individuals belonging to query concepts. Experimentally we show that the behavior of the classifier is comparable with the one of a standard reasoner. Moreover we show that new knowledge (not logically derivable) is induced. It can be suggested to the knowledge engineer for validation, during the ontology population task.

## 1 Introduction

Classification for retrieving resources in a knowledge base (KB) is generally performed through logical approaches that may fail in distributed settings, such as the Semantic Web (SW) context, since they are exposed to inconsistency. Another problem is related to the inherent incompleteness of the KBs in the SW applications, where new resources (web docs or services) are likely to be made available along the time. Statistical methods may be suitable for distributed KBs since they can be often efficient and noise-tolerant. An inductive distance-based method for *concept retrieval* may also suggest new assertions which could not be logically derived, providing also a measure of their likelihood which may help dealing with the uncertainty caused by the incompleteness of the KBs. The time-consuming ontology population task can be facilitated since the knowledge engineer would only have to validate the suggested assertions [1].

Retrieval can be cast as a classification problem, i.e. assessing the class-membership of the individuals in the KB w.r.t. query concepts. Similar individuals should likely belong to similar concepts. Moving from such an idea, an instance-based framework for retrieving resources contained in OWL KBs has been devised. Differently from logic-based approaches to (approximate) instance retrieval [4], we propose an extension of the *Nearest Neighbor* (NN) search to the standard representations for ontologies. Our procedure retrieves individuals belonging to query concepts, by analogy with other training instances, based on the classification of the nearest ones in terms of a dissimilarity measure. Extending the NN search to expressive representations founded in Description Logics (DLs) requires suitable metrics. NN search is generally devised for settings where classes are assumed to be disjoint, which is unlikely in the SW context where an individual can be mapped to a hierarchy of concepts. Furthermore, the DL reasoners make the *Open World Assumption* (OWA), differently from the typical (deductive) database engines working with the *Closed World Assumption* (CWA).

For our purposes, fully semantic metrics [3] are adopted. These language-independent measures assess the dissimilarity of two individuals by comparing them on the grounds of their behavior w.r.t. a committee of features (concepts) that are defined in the KB or that are be generated to this purpose. All the features have the same importance in determining the dissimilarity. However, it may well be that some features have a larger discriminating power w.r.t. the others. In this case, they should be more relevant in determining the dissimilarity value. We propose an extension of former measures, where each feature of the committee is weighted on the grounds of the amount of information that it conveys. This weight is then determined as an *entropic* measure.

The measure has been integrated in the NN procedure [2] and the classification of resources (individuals) w.r.t. a query concept has been performed through a voting procedure weighted by the neighbors' similarity (Sect. 2). The resulting system allowed for an experimentation of the method on performing instance classification with a number ontologies drawn from public repositories (Sect. 4). Its predictions were compared to assertions that were logically derived by a deductive reasoner. The experiments shows that the classification results are comparable (although slightly less complete) and also that the classifier is able to induce new knowledge that is not logically derivable.

## 2 Resource Retrieval as Nearest Neighbor Search

In the following, we assume that concept descriptions are defined in terms of a DL language that can be mapped to OWL-DL. A *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ contains a *TBox* $\mathcal{T}$ and an *ABox* $\mathcal{A}$. $\mathcal{T}$ is a set of (equivalence or also inclusion) axioms that define concepts. $\mathcal{A}$ contains factual assertions concerning the resources, also known as *individuals*. The set of the individuals occurring in $\mathcal{A}$ will be denoted with $\mathsf{Ind}(\mathcal{A})$. As regards the inference services, our procedure may require performing *instance-checking*, namely determining whether an individual, say $a$, belongs to a concept extension, i.e. whether $C(a)$ holds for a certain concept $C$. Because of the OWA, it can happen that a reasoner may be unable to give a positive or negative answer to a class-membership query.

Query answering boils down to determining whether a resource belongs to a concept. Here, an alternative inductive method is proposed. It consists in casting the query answering problem as determining the correct classification for a query individual. The method is grounded on the NN search. It could be able to provide an answer when it may not be inferred by deduction. Moreover, it may also provide a measure of the likelihood of its answer. The basic idea of the NN search is to find the most similar individuals w.r.t. the one that has to be classified. Let $x_q$ be the query instance whose class-membership has to be determined. Using a dissimilarity measure, the set of the $k$ nearest (pre-classified) training instances w.r.t. $x_q$ is selected: $NN(x_q) = \{x_i \mid i = 1, \ldots, k\}$. The objective is to induce an approximation for a discrete-valued target hypothesis function $h : \mathit{IS} \mapsto V$ from a space of instances *IS* to a set of values $V = \{v_1, \ldots, v_s\}$ standing for the classes (concepts) that have to be predicted. In its simplest setting, the $k$-NN algorithm approximates $h$ for classifying $x_q$ on the grounds of the (weighted) value that $h$ is known to assume for the training instances in $NN(x_q)$ as follows:

$$\hat{h}(x_q) := \operatorname*{argmax}_{v \in V} \sum_{i=1}^{k} w_i \delta(v, h(x_i)) \tag{1}$$

where $\delta$ returns 1 in case of matching arguments and 0 otherwise, and, given a dissimilarity measure $d$, the weights are determined by $w_i = 1/d(x_i, x_q)$.

This setting assigns a value to the query instance which stands for one in a set of pairwise disjoint concepts (corresponding to the value set $V$). In a multi-relational setting this assumption cannot be made in general. An individual may be an instance of more than one concept. Moreover, to deal with the OWA, the absence of information on whether a training instance $x$ belongs to the extension of the query concept $Q$ should not be interpreted negatively, as in the standard settings which adopt the CWA. Rather, it should count as neutral (uncertain) information. In order to solve this problems, the multi-class problem is transformed into a ternary one and another value set $V = \{+1, -1, 0\}$ is adopted, where the three values denote, respectively, membership, non-membership, and uncertainty. Specifically, the task can be cast as follows: given a query concept $Q$, determine the membership of an instance $x_q$ through the NN procedure (see Eq. 1) where $V = \{-1, 0, +1\}$ and the hypothesis function values for the training instances are determined by the entailment of the corresponding assertions from the knowledge base, as follows:

$$h_Q(x) = \begin{cases} +1 & \mathcal{K} \models Q(x) \\ -1 & \mathcal{K} \models \neg Q(x) \\ \phantom{+}0 & \textit{otherwise} \end{cases}$$

Note that, being the procedure based on a majority vote of the individuals in the neighborhood, it is less error-prone in case of noise in the data (e.g. incorrect assertions) w.r.t. a purely logic deductive procedure. Therefore, it may be able to give a classification even in case of inconsistent knowledge bases. However, the classification result is not guaranteed to be deductively valid. Indeed, inductive inference naturally yields a certain degree of uncertainty. In order to measure the likelihood of the decision made by the procedure, the quantity that determined the decision should be normalized by dividing it by the sum of such arguments over the (three) possible values:

$$l(\textit{class}(x_q) = v | NN(x_q, k)) = \frac{\sum_{i=1}^{k} w_i \cdot \delta(v, h_Q(x_i))}{\sum_{v' \in V} \sum_{i=1}^{k} w_i \cdot \delta(v', h_Q(x_i))} \qquad (2)$$

Hence the likelihood of the assertion $Q(x_q)$ corresponds to the case when $v = +1$.

## 3 A Family of Epistemic Metrics for Individuals

For the NN procedure, we intend to exploit a family of new measures for DL representations that totally depend on semantic aspects of the individuals in the KB. They are based on the idea that, on a semantic level, similar individuals should behave similarly w.r.t. the same concepts. The rationale is to compare individuals on the grounds of their semantics w.r.t. a collection of concept descriptions, say $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, which stands as a group of discriminating *features* expressed in the OWL-DL sub-language taken into account. They are formally defined as follows [3]:

**Table 1.** Facts concerning the ontologies employed in the experiments.

| Ontology | DL language | #concepts | #object prop. | #data prop. | #individuals |
|---|---|---|---|---|---|
| SWM | $\mathcal{ALCOF}(D)$ | 19 | 9 | 1 | 115 |
| BioPAX | $\mathcal{ALCHF}(D)$ | 28 | 19 | 30 | 323 |
| LUBM | $\mathcal{ALR}^{+}\mathcal{HI}(D)$ | 43 | 7 | 25 | 555 |
| NTN | $\mathcal{SHIF}(D)$ | 47 | 27 | 8 | 676 |
| SWSD | $\mathcal{ALCH}$ | 258 | 25 | 0 | 732 |
| Financial | $\mathcal{ALCIF}$ | 60 | 17 | 0 | 1000 |

**Definition 3.1 (family of measures).** *Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a knowledge base. Given the set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, a family of functions $\{d_p^{\mathsf{F}}\}_{p \in \mathbb{N}}$ with $d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0,1]$ is defined as follows:*

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \qquad d_p^{\mathsf{F}}(a,b) := \frac{L_p(\pi(a), \pi(b))}{m} = \frac{1}{m} \left( \sum_{i=1}^{m} \mid \pi_i(a) - \pi_i(b) \mid^p \right)^{\frac{1}{p}}$$

*where $p > 0$, the weights $w_i \in [0,1]$, $1 \leq i \leq m$, and the* projection function $\pi_i$ *is:*

$$\forall a \in \mathsf{Ind}(\mathcal{A}) \quad \pi_i(a) = \begin{cases} 1 & \mathcal{K} \models F_i(a) \\ 0 & \mathcal{K} \models \neg F_i(a) \\ 1/2 & otherwise \end{cases}$$

Note that in the measure definition, the features have all the same weights. However, each feature could have a different discriminating power. In order to take into account such an aspect, a weight for each feature is introduced. It is determined by exploiting the quantity of information conveyed by the feature, namely by measuring the feature entropy as: $H(F) = -\left(P_F \log(P_F) + P_{\neg F} \log(P_{\neg F}) + P_U \log(P_U)\right)$ where $P_F$ represents the probability of the feature $F$ and is computed as: $P_F = |\mathsf{retrieval}(F)|/|\mathsf{Ind}(\mathcal{A})|$.

## 4 Experiments

The NN procedure integrated with the new distance has been tested for solving a number of retrieval problems. To this purpose, we selected several OWL ontologies, summarized in Tab. 1, available on the web. For each ontology, 20 queries were randomly generated by composition (conjunction and/or disjunction) of (2 through 8) concepts or restrictions of object and data-properties. The performance of the inductive method was evaluated by comparing its responses to those returned by Pellet reasoner. We selected limited training sets (TrSet) that amount to only 4% of the individuals occurring in each ontology. The parameter $k$ was set to $\sqrt{|\mathsf{TrSet}|}$. The simpler distances ($d_1^{\mathsf{F}}$) were employed from the family (weighted on the feature entropy), using all the concepts in the KB for determining the set F. We performed two experiments: one aiming at a three-way classification and the other forcing the response to attribute each test instance to either the target class or to its negation. Initially the standard measures precision, recall, $F_1$-measure were employed to evaluate the system performance. However, due to the OWA, several cases were observed when, it could not be (deductively) ascertained

**Table 2.** Average results (percentages) of the 3-way classification experiment.

| | precision | recall | F-measure | match | commission | omission | induction |
|---|---|---|---|---|---|---|---|
| SWM | 99.0 | 75.8 | 79.5 | 97.5 | 0.0 | 2.2 | 0.3 |
| BIOPAX | 99.9 | 97.3 | 98,2 | 99.9 | 0.1 | 0.0 | 0.0 |
| LUBM | 100.0 | 81.6 | 85.0 | 99.5 | 0.0 | 0.5 | 0.0 |
| NTN | 97.0 | 40.1 | 45.1 | 97.5 | 0.6 | 1.3 | 0.6 |
| SWSD | 94.1 | 38.4 | 46.5 | 98.0 | 0.0 | 1.9 | 0.1 |
| FINANCIAL | 99.8 | 95.0 | 96.6 | 99.7 | 0.0 | 0.0 | 0.2 |

whether a resource was relevant or not for a given query. Hence, we have introduced the following further evaluation indices: 1) *match rate*: number of individuals that got exactly the same classification ($v \in V$) by both the inductive and the deductive classifier w.r.t. the overall number of individuals ($v$ vs. $v$); 2) *omission error rate*: amount of individuals for which inductive method returns $0$ while they were actually relevant according to the reasoner ($0$ vs. $\pm 1$); 3) *commission error rate*: number of individuals found to be relevant to the query concept, while they actually belong to its negation or vice-versa ($+1$ vs. $-1$ or $-1$ vs. $+1$); 4) *induction rate*: amount of individuals found to be relevant to the query concept or to its negation, while either case is not logically derivable from the KB ($\pm 1$ vs. $0$). For each KB, we report the average values (percentages) obtained over the 20 query concepts randomly generated. The outcomes of the three-way classification experiments are reported in Tab. 2. Note that precision and recall are generally quite good for all ontologies but SWSD, where especially recall is significantly lower. SWSD turned out to be more difficult (also in terms of precision) for two reasons: a very limited number of individuals per concept was available and the number of different concepts is larger w.r.t. the other KBs. For the other ontologies values are much higher, as testified also by the F-measure values. Moreover, the results in terms of precision are more stable than those for recall as proved by the limited variance observed, whereas single queries happened to turn out quite difficult as regards the correctness of the answer. The reason for precision being generally higher is probably due to the OWA. Indeed, in many cases it was observed that the NN procedure deemed some individuals as relevant for the query issued while the reasoner was not able to assess this relevance and this was computed as a mistake while it may likely turn out to be a correct inference when judged by a human agent. It is also important to note that, in each experiment, the commission error was quite low or absent. This means that the inductive search procedure did not make critical mistakes. Also the omission error rate was generally quite low, yet more frequent than the previous type of error.

Tab. 3 reports the outcomes of the 2-way experiment where omission errors were ruled out. As expected a dramatic increase of inductive assertions was observed since the system is not trying to compare its inferences to deductive ones but rather it is trying to suggest potential class-memberships. However the observed commission error rates were still quite low which shows that the system was really forcing an answer in the unknown cases and this answer is likely to be correct, at least for those individuals for which the classification can be logically derived from the knowledge base. Conversely, the reported recall rates are higher than with the 3-way setting. The noteworthy low

**Table 3.** Average results (percentages) of the 2-way classification experiment.

| | precision | recall | F-measure | match | commission | omission | induction |
|---|---|---|---|---|---|---|---|
| SWM | 60.3 | 100.0 | 75.2 | 59.9 | 0.0 | 0.0 | 40.1 |
| BIOPAX | 92.2 | 58.4 | 71.5 | 87.1 | 12.9 | 0.0 | 0.0 |
| LUBM | 63.1 | 100.0 | 65.3 | 63.0 | 0.0 | 0.0 | 37.0 |
| NTN | 44.9 | 100.0 | 48.6 | 40.2 | 0.3 | 0.0 | 59.5 |
| SWSD | 6.3 | 100.0 | 7.5 | 6.2 | 0.0 | 0.0 | 93.8 |
| FINANCIAL | 86.8 | 72.6 | 71.8 | 91.6 | 6.2 | 0.0 | 2.2 |

precision and match rate for SWSD (and, partially, for NTN), can be explained by the fact that this ontology has been artificially populated with very few (often 2) individuals per concept, which is harmful for instance-based methods like nearest-neighbor search, especially in a two-classification setting. However less precision is often compensated by high induction rates. The usage of all concepts for the set $\mathsf{F}$ of $d_1^{\mathsf{F}}$ made the measure quite accurate, which is the reason why the procedure resulted quite conservative as regards inducing new assertions. In many cases, it matched rather faithfully the reasoner decisions. From the retrieval point of view, the cases of induction are interesting because they suggest new assertions which cannot be logically derived by using a deductive reasoner yet they might be used to complete a knowledge base [1], e.g. after being validated by an ontology engineer. For each candidate new assertion, Eq. 2 may be employed to assess the likelihood and hence decide on its inclusion. If we compare these outcomes with those reported in other works on instance retrieval and inductive classification [2], where the highest average match rate observed was around 80%, we find a significant increase of the performance due to the accuracy of the new measure.

## 5 Conclusions

This paper investigated the application of a distance-based classification method for KBs represented in OWL. We employed an extended family of dissimilarity measures based on feature committees [3] taking into account the amount of information conveyed by each feature based on an estimate of its entropy. The measures were integrated in a distance-based search procedure that have been exploited for the task of approximate instance retrieval. The experiments made showed that the method is quite effective and can be applied to any domain.

## References


[1] F. Baader, B. Ganter, B. Sertkaya, and U. Sattler. Completing description logic knowledge bases using formal concept analysis. In M. Veloso, editor, *Proc. of IJCAI*, 2007.

[2] C. d'Amato, N. Fanizzi, and F. Esposito. Reasoning by analogy in description logics through instance-based learning. In *Proc. of SWAP 2006 Workshop*, volume 201. CEUR, 2006.

[3] N. Fanizzi, C. d'Amato, and F. Esposito. Induction of optimal semi-distances for individuals based on feature sets. In *Working Notes of DL 2007 Workshop*, volume 250. CEUR, 2007.

[4] R. Möller, V. Haarslev, and M. Wessel. On the scalability of description logic instance retrieval. volume 189. CEUR, 2006.