A Note on the Evaluation of Inductive Concept Classification Procedures

Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito

LACAM, Dipartimento di Informatica, Università degli studi di Bari
Campus Universitario, Via Orabona 4, 70125 Bari, Italy
{claudia.damato, fanizzi, esposito}@di.uniba.it

Abstract. The limitations of deductive logic-based approaches at deriving operational knowledge from ontologies may be overcome by inductive (instance-based) methods, which are usually efficient and noise-tolerant. However the evaluation of such methods is made particularly difficult by the open-world semantics which may often cause individuals not to be deductively classified by the reasoner. In this paper an evaluation method is proposed that is suitable for comparing inductive classification methods to standard reasoners. Experimentally we show that the behavior of a nearest neighbor classifier is comparable with the one of a standard reasoner in terms of the proposed indices.

1 Motivation

Classification for retrieving resources from a knowledge base (KB) in the context of the *Semantic Web* (SW) is an important task that is performed by means of logical methods. These may fail due to the inherent incompleteness and incoherence of the KBs caused by their distributed nature. This has given rise to alternative methods for *approximate* reasoning (see the discussion in [1]) or inductive methods [2, 3] which are known to be both efficient and more noise-tolerant. Extending the inductive methods to the SW representations ultimately founded in *Description Logics* (DL) [4], was not straightforward. In particular, a theoretical problem is posed by the *open-world* semantics of the ontologies, as opposed to the typical *closed-world* assumed in the database applications.

The evaluation of an inductive classification procedure would essentially require the comparison of the inductive answers provided by the inductive method to the correct ones which would derive from the intended semantics of the considered KBs. However this setting is often infeasible as it would require querying the experts and knowledge engineers that built the KB.

Alternatively one may want to compare the inductive answers to those provided by a deductive reasoner, often more efficiently and in a more robust way w.r.t. noise. In previous works [2, 3] we have adopted four indices (*match*, *induction*, *omission* error, *commission* error rates) to evaluate inductive methods compared to deductive ones as *zero-one* loss functions [5] where misclassification is charged a single unit. We extend this notion to the case of standard indices employed in Information Retrieval (IR), namely precision, recall and F-measure, originally defined in terms of a notion of relevance. Even more so, we extend the mentioned indices to take into account the likelihood of the inductively derived answer.

In order to perform experiments with evaluating an inductive classification method, an extension of the *Nearest Neighbor* classification procedure (henceforth, *NN*) [5] was applied to the standard SW representations. The procedure can classify individuals w.r.t. query concepts, by analogy with the classification of the nearest (w.r.t. some similarity criterion) training individuals. This method is quite efficient because it requires checking class-membership for a limited set of training instances. Although a number of dissimilarity measures *for concepts* expressed in various concept languages have been proposed [6, 2], we will resort to language-independent pseudo-metrics *for individuals* [3]. The dissimilarity of two individuals is measured by comparing them w.r.t. a given context, i.e. a committee of features (concepts), namely those defined in the KB or that can be generated to this purpose.

The paper is organized as follows. The basics of the NN procedure applied to SW representations and the similarity measures adopted are recalled in §2. The new indices for measuring the performance of inductive classifiers is presented in §3, and §4 reports the outcomes of experiments measuring the performance of the inductive procedure in terms of the new indices. Concluding remarks are reported in §5.

2 Classification by Analogy

In the following, OWL-DL knowledge bases will be considered with their standard semantics borrowed from DL languages [4]. Specifically, a *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ is assumed to be made up of a *TBox* $\mathcal{T}$ and an *ABox* $\mathcal{A}$ containing, resp., a set of axioms that define concepts and a set of assertions concerning the individuals.

2.1 The Nearest Neighbor Classification Procedure

Classification boils down to determining whether an individual belongs to a concept extension (*instance checking*). An inductive classification method should be able to provide an answer even when this may not be logically inferred. Moreover, it may also provide a measure of the likelihood of its answer.

In *instance-based learning* [5] the basic idea is to find the most similar object(s) to the one that is to be classified w.r.t. a dissimilarity measure. The objective is to induce a classifier as an approximation of a discrete-valued function (a *hypothesis*) $h_C : \textit{IS} \mapsto V$ from a space of instances *IS* to a set of values $V = \{v_1, \ldots, v_s\}$ standing for the classifications that have to be predicted. Normally $|\textit{IS}| \ll |\mathsf{Ind}(\mathcal{A})|$ i.e. only a limited number of training instances is needed especially if they are prototypical for the regions of the search space. Let x be the instance whose classification is to be determined. Using a dissimilarity measure, the set of the k nearest (pre-classified) training instances w.r.t. x is selected: $N_k(x) = \{x_i\}_{i=1}^{k}$.

The k-NN algorithm approximates h_C for classifying x on the grounds of the value that h_C is known to assume for the training instances in $N_k(x)$, i.e. the k closest instances to x in terms of a dissimilarity measure. The value is decided by a weighted majority voting procedure: it is simply the most *voted* value by the instances in $N_k(x)$ weighted by the similarity of the neighbor individual.

The estimate of the hypothesis function for the query individual is:

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

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

Note that $\hat{h}_C$ is defined extensionally: the method needs not to provide an analytically defined function, as other inductive methods do [5]. Being based on a majority vote among the individuals in the neighborhood, this procedure is less error-prone compared to a purely logic deductive one in case of noise caused by incorrect assertions: it may be able to give a correct classification even in case of (partially) inconsistent KBs.

To deal with the open-world semantics, the absence of information on whether a training instance x belongs to the extension of the query concept C should count as neutral (uncertain) information. Thus, assuming the alternate viewpoint, the multi-class problem is transformed into a ternary one. Hence another value set has to be adopted, namely $V = \{+1, -1, 0\}$, where the values denote, respectively, membership, non-membership, and uncertainty, respectively.

The task can be cast as follows: given a query concept C, determine the membership of an instance x through the NN procedure (see Eq. 1) where $V = \{-1, 0, +1\}$ and the hypothesis function values for the training instances are determined as follows: $h_C(x) = +1$ if $\mathcal{K} \models C(x)$, $h_C(x) = -1$ if $\mathcal{K} \models \neg C(x)$ and $h_C(x) = 0$ otherwise.

It should be noted that the inductive inference made by the procedure shown above 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 (x has a classification corresponding to the value v maximizing the argmax argument in Eq. 1), given the nearest training individuals in $N_k(x) = \{x_1, \ldots, x_k\}$, the quantity that determined the decision should be normalized by dividing it by the sum of such arguments over the (three) possible values:

$$\ell[\mathit{class}(x) = v | N_k(x)] = \frac{\sum_{i=1}^{k} w_i \cdot \delta(v, h_C(x_i))}{\sum_{u \in V} \sum_{i=1}^{k} w_i \cdot \delta(u, h_C(x_i))} \tag{2}$$

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

2.2 Semantic Pseudo-Metrics for Individuals

Various definitions of semantic similarity (or dissimilarity) measures for concept languages have been proposed [6, 2]. For our purposes, we need a function for measuring the similarity of individuals rather than concepts.

The new dissimilarity measures are based on the idea of comparing the semantics of the input individuals along a number of dimensions represented by a committee of concept descriptions. Indeed, on a semantic level, similar individuals should behave similarly with respect to the same concepts. Totally semantic distance measures for individuals can be defined in the context of a knowledge base. More formally, 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 sublanguage taken into account.

In its simple formulation, a family of distance functions for individuals inspired to Minkowski's norms L_p can be defined as follows [3]:

Definition 2.1 (family of measures). *Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a knowledge base. Given a set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$ and a weight vector $\boldsymbol{w}$, a family of dissimilarity functions $d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0,1]$ with $p > 0$ is defined as follows:*

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \quad d_p^{\mathsf{F}}(a,b) := \frac{1}{|\mathsf{F}|} \left[\sum_{i=1}^{|\mathsf{F}|} w_i \mid \delta_i(a,b) \mid^p \right]^{\frac{1}{p}}$$

where $\forall i \in \{1, \ldots, m\}$ the dissimilarity function δ_i is defined by:

$$\delta_i(a,b) = \begin{cases} 0 & \textit{if } (\mathcal{K} \models F_i(a) \wedge \mathcal{K} \models F_i(b)) \vee (\mathcal{K} \models \neg F_i(a) \wedge \mathcal{K} \models \neg F_i(b)) \\ 1 & \textit{if } (\mathcal{K} \models F_i(a) \wedge \mathcal{K} \models \neg F_i(b)) \vee (\mathcal{K} \models \neg F_i(a) \wedge \mathcal{K} \models F_i(b)) \\ \frac{1}{2} & \textit{otherwise} \end{cases}$$

The weight vector $\boldsymbol{w}$ may be determined by the amount of information conveyed by each feature, which can be measured as its estimated entropy: $w_i = H(F_i)$ (as in [3]).

3 Performance Indices

Machine learning focuses on binary or multiple classification problems that can be reduced to the binary case, as normally classes are assumed to be mutually disjoint. In a representation that adopts an open-world semantics a ternary response function requires a different treatment. Adopting a response set $V = \{-1, 0, +1\}$, cases of uncertain classification may happen. In the following we assume to evaluate a set of inductive classifications I_C to one of deductive classifications D_C on concept C, where $C(a) \in I_C$ iff $\hat{h}_C(a) = +1$ and $\neg C(a) \in I_C$ iff $\hat{h}_C(a) = -1$; the same applies for D_C where $\hat{h}_C = h_C$, i.e. it should be deductively computed by the reasoner.

3.1 Generalized IR Measures

Since classification can be employed to retrieving the individuals that are related to a given query concept, it is quite straightforward to adopt the standard measures used in IR: *precision* (P), *recall* (R), *F-measure*. However this suits settings with binary responses determined by a notion of *relevance* w.r.t. the query. In the SW context, a different purpose is pursued by the *semantic precision* and *recall* measures [7], introduced for assessing the quality of alignments in an ontology matching task.

As a first step, the definition of precision and recall may be generalized adopting different measures of the overlap [8]:

$$P_\omega(S_1, S_2) = \frac{\omega(S_1, S_2)}{|S_1|} \qquad\qquad R_\omega(S_1, S_2) = \frac{\omega(S_1, S_2)}{|S_2|} \tag{3}$$

where originally these sets are made up of alignments, but we may consider them as membership axioms.

These measures should be chosen so that some properties are fulfilled : $\forall S_1, S_2$

– $\omega(S_1, S_2) \geq 0$ *positiveness*
– $\omega(S_1, S_2) \leq \min(|S_1|, |S_2|)$ *maximality*
– $\omega(S_1, S_2) \geq |S_1 \cap S_2|$ *boundedness*

Now considering $S_1 = I_C$ and $S_2 = D_C$, the basic definition of the measures corresponds to $\omega_0 := |I_C \cap D_C|$ which trivially fulfills all properties above.

In our specific setting, since the answers of the inductive classifier are to be compared to those of the reasoner, one may also check the precision and recall of the single responses $v \in V$ separately and then consider the (weighted) average of these precision (or recall) measures as an overall index.

$$\overline{P}(I_C, D_C) := \sum_{v \in V} w_v \frac{|I_C^v \cap D_C^v|}{|I_C^v|} \qquad \textit{average precision} \tag{4}$$

$$\overline{R}(I_C, D_C) := \sum_{v \in V} w_v \frac{|I_C^v \cap D_C^v|}{|D_C^v|} \qquad \textit{average recall} \tag{5}$$

where I_C^v (resp. D_C^v) denotes the subset of the individuals with same classification: $\{a \in I_C \mid \hat{h}_C(a) = v\}$ (resp. $\{a \in D_C \mid \hat{h}_C(a) = v$). The case of uniform weights $w_v = 1/|V|$, $\forall v \in V$ corresponds to macro-averaging over the possible values in V. Alternatively, one may consider the choice $w_v = |D_C^v|/|\mathit{TS}|$, $\forall v \in V$, where TS represents an independent set of individuals employed for testing.

In [7] some properties are introduced for semantic precision and recall measures:

1. $D_C \models I_C \Rightarrow P(I_C, D_C) = 1$ *max-correctness*
2. $I_C \models D_C \Rightarrow R(I_C, D_C) = 1$ *max-completeness*
3. $Cn(I_C) = Cn(D_C)$ iff $P(I_C, D_C) = 1$ and $R(I_C, D_C) = 1$ *definiteness*
4. $P(I_C, D_C) \geq 0$ and $R(I_C, D_C) \geq 0$ *positiveness*
5. $P(I_C, D_C) \leq 1$ and $R(I_C, D_C) \leq 1$ *maximality*
6. $P'(I_C, D_C) \geq P(I_C, D_C)$ and $R'(I_C, D_C) \geq R(I_C, D_C)$ *boundedness*
   for all alternative precision and recall measures P' and R'.

where $\models$ is a shortcut for the entailment relation between the assertions in each set and $Cn(\cdot)$ returns the set of assertions entailed by the input set of assertions. Of course entailment w.r.t. the models of the underlying KB is considered.

Proposition 3.1. *Measures $\overline{P}$ and $\overline{R}$ fulfill properties 1–6.*

Proof. We will consider the uniform weight case $w_v = 1/|V|$:

1. $\overline{P}(I_C, D_C) = \frac{1}{|V|} \sum_{v \in V} |I_C^v \cap D_C^v|/|I_C^v| = \frac{1}{|V|} \sum_{v \in V} |I_C^v|/|I_C^v| = 1$;
2. analogously;
3. $Cn(I_C) = Cn(D_C)$ iff $\forall v \in V\colon\ (I_C^v \cap D_C^v) = I_C^v = D_C^v$ iff
   $\overline{P}(I_C, D_C) = 1$ and $\overline{R}(I_C, D_C) = 1$;
4. trivial;
5. trivial;
6. $\overline{P}(I_C, D_C) = \frac{1}{|V|} \sum_{v \in V} |I_C^v \cap D_C^v|/|I_C^v| \geq \frac{1}{|V|} \sum_{v \in V} |I_C^v \cap D_C^v|/|I_C| =$
   $= |I_C \cap D_C|/|I_C| = P(I_C, D_C)$; analogously $\overline{R}(I_C, D_C) \geq R(I_C, D_C)$.

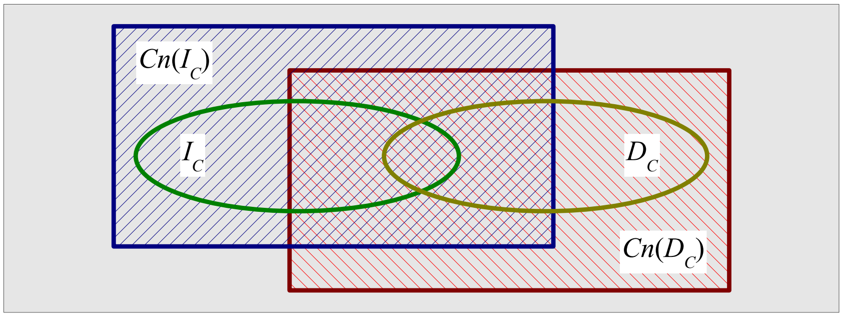

Fig. 1. Representation of the sets involved in the computation of precision and recall.

An ideal semantic generalization of the precision and recall measures in Eq.3 exploits the derivation closures (see Fig. 1) in the computation of the overlaps:

$$P_{ideal}(I_C, D_C) := P_{\omega_0}(Cn(I_C), Cn(D_C)) = \frac{|Cn(I_C) \cap Cn(D_C)|}{|Cn(I_C)|} \tag{6}$$

$$R_{ideal}(I_C, D_C) := R_{\omega_0}(Cn(I_C), Cn(D_C)) = \frac{|Cn(I_C) \cap Cn(D_C)|}{|Cn(D_C)|} \tag{7}$$

However, as noted in [7], these measures would be undefined when the number of consequences is infinite. Better semantic definitions are then:

$$P_{sem}(I_C, D_C) := P_{\omega_0}(I_C, Cn(D_C)) = \frac{|I_C \cap Cn(D_C)|}{|I_C|} \tag{8}$$

$$R_{sem}(I_C, D_C) := R_{\omega_0}(Cn(I_C), D_C) = \frac{|Cn(I_C) \cap D_C|}{|D_C|} \tag{9}$$

where the problem is solved since $|I_C| \leq |TS|$ and $|D_C| \leq |TS|$.

It is easy to prove that:

Proposition 3.2. *Measures P_{sem} and R_{sem} fulfill properties 1–6.*

Proof.

1. $D_C \models I_C \Rightarrow Cn(D_C) \subseteq I_C \Rightarrow P_{sem}(I_C, D_C) = |I_C \cap Cn(D_C)|/|I_C| = |I_C|/|I_C| = 1$*;*
2. *analogously;*
3. $Cn(I_C) = Cn(D_C)$ *iff* $I_C \subseteq Cn(I_C) = Cn(D_C) \supseteq D_C$ *iff*
   $P_{sem}(I_C, D_C) = |I_C|/|I_C| = 1$ *and* $R_{sem}(I_C, D_C) = |D_C|/|D_C| = 1$*;*
4. *trivial;*
5. *trivial;*
6. *trivial since* $I_C \subseteq Cn(I_C)$ *and* $D_C \subseteq Cn(D_C)$*.*

3.2 Other Measures

Alternative evaluation indices have been used [2, 3] that, unlike the previous ones, do not have a direct mapping to the sets of true/false positives/negatives:

- *match*: case of an individual that got the same classification by the reasoner and the inductive classifier;
- *omission error*: case of an individual for which the inductive method could not determine whether it was relevant to the query or not (response 0) while it was found relevant by the reasoner (response ± 1);
- *commission error*: case of an individual found to be relevant to the query concept (response -1 or $+1$) by the inductive classifier, while it logically belongs to its negation or vice-versa (response $+1$ or -1, respectively);
- *induction rate*: case of an individuals found to be relevant to the query concept or to its negation (response ± 1), while either case is not logically derivable from the knowledge base (response 0).

Each case increases an index with a single unit (zero-one loss). This can be generalized by exploiting the likelihood measure provided by the inductive procedure in order to assign parts of the unit to each of the three possible responses. This, in turn, has an impact on the measure of the indices normally presented in terms of rates.

Specifically, comparing the inductive answers to those of a reasoner, for each inductive classification, instead of incrementing a single count, one may selectively increase more indices at the same time, using the estimated likelihood measures as in Eq. 2 ($\forall v \in V\colon\ \ell_v = \ell[class(x) = v]$):

-1: increase the match count with the likelihood ℓ_{-1}, increase the omission error count with ℓ_0, increase the commission error count with ℓ_{+1};
0: increase the match count with the likelihood ℓ_0, increase the induction count with $\ell_{-1} + \ell_{+1}$;
$+1$: increase the match count with the likelihood ℓ_{+1}, increase the omission error count with ℓ_0, increase the commission error count with ℓ_{-1}.

The counts above can be represented in a contingency matrix $M = (m_{uv})_{u,v \in V}$ which gives an idea of the performance in multi-class problems. An even better evaluation can be performed by comparing M to another matrix $R = (r_{uv})_{u,v \in V}$ representing the outcomes with the random classifier exploiting the row and column totals: $r_{uv} = (\sum_{w \in V} m_{uw} \cdot \sum_{w \in V} m_{wv})/N$, with $N = \sum_{w,t \in V} m_{wt}$. The kappa statistic

$$\kappa := \frac{\left(\sum_{v \in V} m_{vv} - \sum_{v \in V} r_{vv}\right)}{\left(N - \sum_{v \in V} r_{vv}\right)}$$

can be employed to measure the relative improvement over the random classifier [5].

4 Some Experiments

In order to test the NN procedure integrated with the pseudo-metric proposed in the previous sections, it was applied to a number of classification problems. To this pur-

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

Table 2. Experimental results in terms of the new IR measures.

ontology	precision	recall	F-measure
SWM	97.46±03.27	96.87±04.23	97.16±03.69
BioPax	92.21±13.00	80.23±14.55	85.80±13.73
LUBM	99.14±04.64	93.85±12.09	96.42±06.71
NTN	82.57±17.22	63.13±17.74	71.55±17.48
SWSD	80.82±08.88	75.66±04.24	78.15±05.74
Financial	97.72±07.36	57.86±13.45	72.68±09.52

pose, we selected some OWL ontologies from different domains, namely: Surface-Water-Model (SWM), NewTestamentNames (NTN) from the Protégé library[1], our Semantic Web Service Discovery dataset[2] (SWSD), one generated by the Lehigh University Benchmark (LUBM), the BioPax glycolysis ontology[3] (BioPax) and Financial ontology[4]. Tab. 1 summarizes important details concerning these ontologies.

A 10-fold cross validation was performed. The simplest version of the distance (d_1^{F}) was employed using all the concepts in the knowledge base for determining the set F. The parameter k was set to $\sqrt{|\mathsf{Ind}(\mathcal{A})|}$ depending on the number of individuals in the ontology. The performance was evaluated comparing the unductive responses to those returned by a standard reasoner[5] as a baseline.

4.1 Generalized IR Measures

The outcomes are reported in Fig.2. For each knowledge base, we report the average values (and the standard deviation) obtained classifying each individual against each concept in the KB using both the reasoner and the NN classifier.

It is possible to note that generally results are good especially for the smaller ontologies (in terms of number of individuals). In particular precision was good ($> 90\%$) for all but for the SWSD and NTN ontologies for which it drops to around 80%. Namely,

[1] `http://protege.stanford.edu/plugins/owl/owl-library`

[2] `https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Projects/xmedia/dl-tree.htm`

[3] `http://www.biopax.org/Downloads/Level1v1.4/`

[4] `http://www.cs.put.poznan.pl/alawrynowicz/`

[5] We employed Pellet v. 1.5.2. See `http://pellet.owldl.com`

Table 3. Results with alternative indices: set-theoretic (s.) and likelihood version (l.).

ontology	type	match	commission	omission	induction
SWM	*s.*	97.89±03.04	00.00±00.00	00.92±01.25	01.20±01.82
	l.	96.96±04.16	00.00±00.00	01.31±01.61	01.73±02.68
BioPax	*s.*	90.90±13.28	08.56±13.37	00.00±00.00	00.54±02.64
	l.	87.75±15.38	11.38±15.35	02.29±01.25	00.48±02.28
LUBM	*s.*	98.48±06.30	00.00±00.00	00.83±02.64	00.69±05.19
	l.	97.69±07.51	00.00±00.00	01.15±04.10	01.15±03.68
NTN	*s.*	88.39±16.69	00.24±01.08	05.99±07.27	05.37±09.33
	l.	86.53±17.51	00.38±01.66	06.34±08.05	06.55±08.48
SWSD	*s.*	98.12±05.01	00.00±00.00	01.15±02.86	00.74±02.17
	l.	97.80±05.47	00.00±00.00	01.12±02.73	01.11±02.73
Financial	*s.*	97.02±07.76	02.64±07.77	00.02±00.07	00.32±00.15
	l.	93.86±09.21	03.55±09.26	02.17±01.83	00.42±00.36

SWSD turned out to be more difficult (also in terms of recall) for two reasons: a very limited number of individuals per concept was available and the number of different concepts is larger than in other knowledge bases. For the other ontologies scores are quite high, as testified also by the F-measure values. The results in terms of recall are also more stable than those for recall as proved by the limited variance observed, whereas some concepts turned out to be quite difficult.

The reason for precision being less than recall are probably due to the open-world assumption. Indeed, in a many cases it was observed that the NN procedure deemed some individuals as relevant for the target concept while the DL 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 expert. Thus different indices would be needed in this case that may make explicit both the rate of inductively classified individuals and the nature of the mistakes.

4.2 Other Measures

Tab. 3 reports the outcomes in terms of the set-theoretic loss-function and the new indices exploiting the likelihood value provided by the NN classifier. Preliminarily, it is important to note that, in each experiment, the commission error was low or absent (except for the BioPax ontology). This means that the search procedure is generally quite accurate: it did not make critical mistakes i.e. cases when an individual is deemed as an instance of a concept while it really is an instance of a disjoint one. Also omission error and induction rates are quite low, yet they were more typically observed in the experiments with the considered ontologies.

The usage of all concepts for the set F of d_1^{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 IR 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 [3], e.g. after being validated by an ontology engineer.

5 Concluding Remarks

In this paper new evaluation measures were proposed that are suitable for comparing inductive classification methods to standard reasoners. Experimentally, we showed that the behavior of a NN classifier is comparable with the one of a standard reasoner in terms of the proposed indices.

Other extensions of the current measures may be made exploiting the probabilistic output and different loss-functions. Futher measures such as specificity, sparsity, fallout could also be generalized. Moreover, the same criteria may be adopted also in the evaluation of approximate reasoning methods [1]. Similar measures may be also employed for evaluating other learning algorithms, such as (un)supervised conceptual clustering [9]. We are currently investigating the possibility of devising classifiers that provide binary responses [10], that would require suitable performance indices.

References

[1] Hitzler, P., Vrandečić, D.: Resolution-based approximate reasoning for OWL DL. In Gil, Y., Motta, E. Benjamins, V., Musen, M.A., eds.: Proceedings of the 4th International Semantic Web Conference, ISWC2005. Number 3279 in LNCS, Galway, Ireland, Springer (2005) 383–397

[2] d'Amato, C., Fanizzi, N., Esposito, F.: Reasoning by analogy in description logics through instance-based learning. In Tummarello, G., Bouquet, P., Signore, O., eds.: Proceedings of Semantic Web Applications and Perspectives, 3rd Italian Semantic Web Workshop, SWAP2006. Volume 201 of CEUR Workshop Proceedings., Pisa, Italy, CEUR-WS.org (2006)

[3] d'Amato, C., Fanizzi, N., Esposito, F.: Query answering and ontology population: An inductive approach. In Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M., eds.: Proceedings of the 5th European Semantic Web Conference, ESWC2008. Volume 5021 of LNCS., Springer (2008) 288–302

[4] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook. Cambridge University Press (2003)

[5] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann (2005)

[6] Borgida, A., Walsh, T., Hirsh, H.: Towards measuring similarity in description logics. In Horrocks, I., Sattler, U., Wolter, F., eds.: Working Notes of the International Description Logics Workshop. Volume 147 of CEUR Workshop Proceedings., Edinburgh, UK (2005)

[7] Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In Veloso, M., ed.: Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI2007, Hyderabad, India (2007) 348–353

[8] Eherig, M., Euzenat, J.: Relaxed precision and recall for ontology matching. In Ashpole, B., Euzenat, J., Eherig, M., Stuckenschmidt, H., eds.: Proceedings of the K-Cap2005 Workshop on Integrating Ontology, Banff, Canada (2005) 25–32

[9] Fanizzi, N., d'Amato, C., Esposito, F.: Conceptual clustering for concept drift and novelty detection. In Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M., eds.: Proceedings of the 5th European Semantic Web Conference, ESWC2008. Volume 5021 of LNAI., Springer (2008) 318–332

[10] d'Amato, C., Fanizzi, N., Esposito, F.: Distance-based classification in OWL ontologies. In Lovrek, I., Howlett, R., Jain, L., eds.: Proceedings of the 12th International Conference, KES2008. Volume 5178 of LNAI., Zagreb, Croatia, Springer (2008) 656–661