Predicting the Possibilistic Score of OWL Axioms through Modified Support Vector Clustering

Dario Malchiodi, Andrea G. B. Tettamanzi

Predicting the Possibilistic Score of OWL Axioms through Modified Support Vector Clustering

Dario Malchiodi
Università degli Studi di Milano, Dipartimento di Informatica
Milano, Italy
dario.malchiodi@unimi.it

Andrea G. B. Tettamanzi
Université Côte d'Azur, I3S, Inria
Sophia Antipolis, France
andrea.tettamanzi@unice.fr

ABSTRACT

We address the problem of predicting a score for candidate axioms within the context of ontology learning. The prediction is based on a learning procedure based on support vector clustering originally developed for inferring the membership functions of fuzzy sets, and on a similarity measure for subsumption axioms based on semantic considerations and reminiscent of the Jaccard index. We show that the proposed method successfully learns the possibilistic score in a knowledge base consisting of pairs of candidate OWL axioms, meanwhile highlighting that a small subset of the considered axioms turns out harder to learn than the remainder.

CCS CONCEPTS

• **Computing methodologies** → **Vagueness and fuzzy logic**; **Kernel methods**; *Ontology engineering*; • **Information systems** → *Web Ontology Language (OWL)*;

KEYWORDS

Support Vector Clustering, Possibilistic OWL Axiom Scoring

ACM Reference Format:
Dario Malchiodi and Andrea G. B. Tettamanzi. 2018. Predicting the Possibilistic Score of OWL Axioms through Modified Support Vector Clustering. In *SAC 2018: Symposium on Applied Computing , April 9–13, 2018, Pau, France*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3167132.3167345

1 INTRODUCTION

Ontology learning [8] is an emerging field of research, whose goal is to overcome the knowledge acquisition bottleneck through the automatic generation of ontologies, mainly within the context of the semantic Web. The input for ontology learning can be text in natural language or existing ontologies (typically expressed in OWL) and instance data (typically represented in RDF) [7]. In the latter case, induction-based methods like the ones developed in inductive logic programming and data mining are developed to detect meaningful patterns and learn schema axioms from existing instance data (facts) and their metadata, if available.

SAC 2018, April 9–13, 2018, Pau, France

Ontology learning relies critically on (candidate) axiom scoring. To see why, let us consider the following example. While constructing an ontology for a given domain (say, politics), based on the description of instances in a given dataset, e.g., DBpedia, we might suspect that a mayor is an elected representative. Before we add this piece of knowledge to the ontology, we should test the corresponding axiom `SubClassOf(Mayor ElectedRepresentative)` against the statements in the dataset. In practice, testing an axiom boils down to computing an acceptability score, measuring the extent to which the axiom is compatible with the recorded facts.

Methods to approximate the semantics of given types of axioms have been throughly investigated in the last decade (e.g., approximate subsumption [13]) and some related heuristics have been proposed to score concept definitions in concept learning algorithms [12]. The most popular candidate axiom scoring heuristics proposed in the literature are based on statistical inference (see, e.g., [4]). Because such a probability-based framework is not always completely satisfactory (see Sect. 4 of [15] for a detailed critique), an alternative axiom scoring heuristics based on a formalization in possibility theory of the notions of logical content of a theory and of falsification and complying with an open-world semantics has recently been proposed [15]. While empirical evidence has been found that such a possibilistic scoring heuristics may lead to more accurate ontologies [16], the heavy computational cost of the heuristics makes it hard to apply in practice, unless some implementation tricks are devised (e.g., time capping [14]).

This work considers a promising alternative to the direct computation of the possibilistic score, consisting in training a surrogate model on a sample of candidate axioms for which the score has already been computed or is otherwise available, in order to be capable of *predicting* the score of a novel, unseen candidate axiom. The major weakness of this approach is related to the fact that training such a model may consume a significant amount of resources; On the other hand, that can be done once and for all: once trained, the model can be used to score new candidate axioms at a comparatively negligible additional cost.

We apply a support vector clustering method, originally developed for learning the membership functions of fuzzy sets [9], to this task, i.e., predicting the possibilistic score of candidate OWL axioms. To prove the feasibility of this concept, we perform an experiment on a dataset of `SubClassOf` (i.e., subsumption) axioms, whose possibilistic score has been previously determined by direct application of the heuristics on the DBpedia RDF dataset [14].

The paper is structured as follows: Sect. 2 depicts the heuristics estimating the possibilistic score of the axioms used in order to train and test the predictors, while Sect. 3 explains how the similarity between axioms has been computed. In Sect. 4 we illustrate the learning procedure having as input the above mentioned scores and similarities, and producing predictors as output. The performed numerical experiments are described and discussed in Sect. 5. Some concluding remarks end the paper.

2 POSSIBILISTIC AXIOM SCORING

We briefly recall the possibilistic scoring heuristics for OWL axioms proposed in [15], to which the reader is referred to for the details.

Given a candidate OWL 2 axiom ϕ, expressing a *hypothesis* about the relations holding among some entities of a domain, we wish to evaluate its credibility, in terms of possibility and necessity, based on the *evidence* available in the form of a set of facts contained in an RDF dataset $\mathcal{K}$.

The content of ϕ is defined as a (finite) set of *basic statements* ψ which are logical consequences of ϕ, i.e., $\phi \models \psi$. The open-world hypothesis (OWA) is fully taken into account. Therefore, given $\mathcal{K}$, for each such ψ, there are three cases:

(1) $\mathcal{K} \models \psi$: in this case, we will call ψ a *confirmation* of ϕ;
(2) $\mathcal{K} \models \neg\psi$: if so, we will call ψ a *counterexample* of ϕ;
(3) $\mathcal{K} \not\models \psi$ and $\mathcal{K} \not\models \neg\psi$: in this case, ψ is neither a confirmation nor a counterexample of ϕ.

Possibility theory [5] is a mathematical theory of epistemic uncertainty. Its central notion is that of a possibility distribution which assigns to each elementary event a degree of possibility ranging from 0 (impossible, excluded) to 1 (completely possible, normal). A possibility distribution π induces a *possibility measure* Π, corresponding to the greatest of the possibilities associated to an event and the dual *necessity measure* N, equivalent to the impossibility of the negation of an event.

If we denote by u_ϕ the *support* of ϕ, which is the cardinality of its content, by u_ϕ^+ the number of confirmations of ϕ and by u_ϕ^- the number counterexamples of ϕ, the possibility and the necessity of candidate axiom ϕ may be defined as follows:

- if $u_\phi > 0$,

$$\Pi(\phi) = 1 - \sqrt{1 - \left(\frac{u_\phi - u_\phi^-}{u_\phi}\right)^2}; \qquad (1)$$

$$N(\phi) = \begin{cases} \sqrt{1 - \left(\frac{u_\phi - u_\phi^+}{u_\phi}\right)^2}, & \text{if } u_\phi^- = 0, \\ 0, & \text{if } u_\phi^- > 0; \end{cases} \qquad (2)$$

- if $u_\phi = 0$, $\Pi(\phi) = 1$ and $N(\phi) = 0$, i.e., we are in a state of maximum ignorance, given that no evidence is available in the RDF dataset to assess the credibility of ϕ.

The possibility and necessity of an axiom can be combined into a single handy acceptance/rejection index

$$\begin{aligned} \mathrm{ARI}(\phi) &= N(\phi) + \Pi(\phi) - 1 = N(\phi) - N(\neg\phi) \\ &= \Pi(\phi) - \Pi(\neg\phi) \in [-1, 1], \end{aligned} \qquad (3)$$

because $N(\phi) = 1 - \Pi(\neg\phi)$ and $\Pi(\phi) = 1 - N(\neg\phi)$ (duality of possibility and necessity). A negative $\mathrm{ARI}(\phi)$ suggests rejection of ϕ ($\Pi(\phi) < 1$), whilst a positive $\mathrm{ARI}(\phi)$ suggests its acceptance ($N(\phi) > 0$), with a strength proportional to its absolute value. A value close to zero reflects ignorance about the status of ϕ.

One nice property of this acceptance/rejection index, which stems from the duality of possibility and necessity, is that, for all ϕ,

$$\mathrm{ARI}(\neg\phi) = -\mathrm{ARI}(\phi).$$

The idea, then, is that, if we can train a model to predict the possibility of a candidate axiom ϕ, $\Pi(\phi)$, and of its negation, $\Pi(\neg\phi)$, we have enough information to estimate $\mathrm{ARI}(\phi)$ without having to compute u_ϕ, u_ϕ^+, and u_ϕ^-. To this aim, we can use a set of candidate axioms whose ARI (i.e., possibility and necessity) is known to construct a training set consisting of axioms and their negations, labeled with their possibility, which may be regarded formally as a degree of membership (in the fuzzy set of possible formulas).

3 AXIOM SIMILARITY

The support vector clustering method we train to predict the possibilistic score of candidate axioms requires a kernel function which, for our purposes, may be assumed to return the similarity between two candidate axioms.

It is clear that such similarity should be based on the semantics of axioms and not on their syntax. Furthermore, since we operate in a possibilistic framework, it makes sense to define a similarity measure with values in $[0, 1]$, satisfying the following desirable properties: for all axioms ϕ and ψ,

(1) $0 \le \mathrm{sim}(\phi, \psi) \le 1$;
(2) $\mathrm{sim}(\phi, \psi) = 1$ if and only if $\phi \equiv \psi$;
(3) $\mathrm{sim}(\phi, \psi) = \mathrm{sim}(\psi, \phi)$.

Our basic intuition is to base the definition of such a similarity measure on an underlying fuzzy implication operator [10]. If such a suitable operator Impl is given, then the similarity between two axioms ϕ and ψ may be defined as

$$\mathrm{sim}(\phi, \psi) = \min\{\mathrm{Impl}(\phi, \psi), \mathrm{Impl}(\psi, \phi)\}, \qquad (4)$$

which may be paraphrased as saying that ϕ and ψ are similar to the extent that $\phi \Rightarrow \psi$ and $\psi \Rightarrow \phi$ (i.e., the min in the definition translates a logical conjunction).

This moves the problem one step away, to finding a suitable definition for the Impl operator. Classical (material) implication would be defined as

$$\mathrm{Impl}(\phi, \psi) = \begin{cases} 1, & \text{if } \models \neg\phi \vee \psi; \\ 0, & \text{otherwise.} \end{cases}$$

A tentative fuzzy definition, based on the Herbrand semantics of the axioms, might be the following:

$$\begin{aligned}\mathrm{Impl}(\phi,\psi) &= \tfrac{\|\{\mathcal{I}:\mathcal{I}\models\neg\phi\vee\psi\}\|}{\|\Omega\|} \\ &= \tfrac{\|[\neg\phi]\cup[\psi]\|}{\|\Omega\|} = \tfrac{\|\overline{[\phi]}\cup[\psi]\|}{\|\Omega\|},\end{aligned} \quad (5)$$

$[\phi]$ and Ω denoting the set of the models of ϕ and the universe set.

One problem with Equation 5 is that exactly computing the numerator would require to count the models and countermodels of both axioms being compared. This can be very inefficient, as the Herbrand universe of a real-world RDF dataset may be huge and Ω is its power set. However, a rough approximation of it can be obtained by replacing interpretations by individuals occurring in the RDF dataset that confirm or contradict the two axioms.

We now restrict our attention to subsumption axioms, of the form SubClassOf$(C\ D)$, or, to use the more compact DL syntax, $C \sqsubseteq D$, where C and D are two OWL class expressions, and their negation, which one may write as $\neg(C \sqsubseteq D)$ or $C \not\sqsubseteq D$. Given an individual a occurring in an RDF dataset, we may say that

- a confirms axiom $C \sqsubseteq D$ (and contradicts $C \not\sqsubseteq D$) iff $C(a) \wedge D(a)$;
- a contradicts axiom $C \sqsubseteq D$ (and confirms $C \not\sqsubseteq D$) iff $C(a) \wedge \neg D(a)$.

Instead of counting the models of an axiom, we will be counting the individuals that confirm it; instead of counting its countermodels, we will count the individuals that contradict it. Given an OWL class expression C, let us denote by

$$[C] = \{a : C(a)\}$$

the extension of C in the RDF dataset at hand. Given four OWL class expressions A, B, C, and D, we can now compute the degree to which axiom $A \sqsubseteq B$ implies axiom $C \sqsubseteq D$ as

$$\begin{aligned}\mathrm{Impl}(A \sqsubseteq B, C \sqsubseteq D) &= \tfrac{\|\{a:(A(a)\wedge\neg B(a))\vee(C(a)\wedge D(a))\}\|}{\|\{a:(A(a)\vee C(a))\}\|} \\ &= \tfrac{\|[A]\cap\overline{[B]}\cup[C]\cap[D]\|}{\|[A]\cup[C]\|}.\end{aligned} \quad (6)$$

Substituting in Equation 4 then would yield

$$\begin{aligned}&\mathrm{sim}(A \sqsubseteq B, C \sqsubseteq D) = \\ &= \min\{\mathrm{Impl}(A \sqsubseteq B, C \sqsubseteq D), \mathrm{Impl}(C \sqsubseteq D, A \sqsubseteq B\} \\ &= \min\left\{\tfrac{\|[A]\cap\overline{[B]}\cup[C]\cap[D]\|}{\|[A]\cup[C]\|}, \tfrac{\|[C]\cap\overline{[D]}\cup[A]\cap[B]\|}{\|[A]\cup[C]\|}\right\} \\ &= \tfrac{\min\{\|[A]\cap\overline{[B]}\cup[C]\cap[D]\|, \|[C]\cap\overline{[D]}\cup[A]\cap[B]\|\}}{\|[A]\cup[C]\|}.\end{aligned}$$

An alternative, simpler definition, much in the same spirit, might be

$$\mathrm{sim}(A \sqsubseteq B, C \sqsubseteq D) = \frac{\|[A]\cap[B]\cup[C]\cap[D]\|}{\|[A]\cup[C]\|}, \quad (7)$$

which is somehow reminiscent of the Jaccard index. We decide to stick with this latter definition of similarity.

Given that, in order to predict the ARI of subsumption axioms we will need to compute similarities between positive or negated subsumption axioms, we can give the formulas to apply in all four cases as in Table 1.

Table 1: A summary of the formulas to be used to compute the similarity $\mathrm{sim}(\phi,\psi)$ between positive or negated subsumption axioms ϕ and ψ.

$\downarrow\phi \quad \psi\rightarrow$	$C \sqsubseteq D$	$C \not\sqsubseteq D$
$A \sqsubseteq B$	$\frac{\|[A]\cap[B]\cup[C]\cap[D]\|}{\|[A]\cup[C]\|}$	$\frac{\|[A]\cap[B]\cup[C]\cap\overline{[D]}\|}{\|[A]\cup[C]\|}$
$A \not\sqsubseteq B$	$\frac{\|[A]\cap\overline{[B]}\cup[C]\cap[D]\|}{\|[A]\cup[C]\|}$	$\frac{\|[A]\cap\overline{[B]}\cup[C]\cap\overline{[D]}\|}{\|[A]\cup[C]\|}$

The similarity between two candidate OWL axioms of the form $A \sqsubseteq B$ and $C \sqsubseteq D$, as defined in Equation 7 can be easily computed using SPARQL counting queries. For instance, the denominator $\|[A]\cup[C]\|$ may be computed by

```
SELECT (count(DISTINCT ?x) AS ?n)
WHERE { { ?x a A . } UNION { ?x a C . } },        (8)
```

whereas the numerators for the four cases covered in Table 1 may be computed by SPARQL queries of the form

```
SELECT (count(DISTINCT ?x) AS ?n)
WHERE { { Q([A]) . Q([B]) . }
        UNION                                     (9)
        { Q([C]) . Q([D]) . } },
```

where

$$\begin{aligned}Q([X]) &= \texttt{?x a } X, \\ Q(\overline{[X]}) &= \texttt{FILTER NOT EXISTS ?x a } X.\end{aligned}$$

4 FUZZY MEMBERSHIP INFERENCE

Let A be a fuzzy set and denote by μ_A the corresponding membership function. Given a set $\{x_1,\ldots,x_n\}$ of objects and the corresponding set of membership values $\{\mu_1,\ldots,\mu_n\}$, that is $\mu_i = \mu_A(x_i)$ for each i, the procedure described in [9] induces an approximation $\hat{\mu}$ of μ_A, provided that for each $i,j = 1,\ldots,n$ a similarity value between x_i and x_j, denoted $k(x_i,x_j)$, is available. This procedure, relying on a customization of an SVM algorithm (see for instance [1] and [2] for similar approaches used for the inference of regression functions), is based on the minimization of

$$f(\chi_1,\ldots,\chi_n) = \sum_{i,j=1}^{n} \chi_i\chi_j k(x_i,x_j) - \sum_{i=1}^{n} \chi_i k(x_i,x_i) \quad (10)$$

as a function of $\chi_1,\ldots,\chi_n$ under the constraints $\sum_i \chi_i = 1$ and $-C(1-\mu_i) \leq \chi_i \leq C\mu_i$ for each $i = 1,\ldots,m$, where $C > 0$ is a hyperparameter. Consider the optimal values $\chi_1^*,\ldots,\chi_n^*$ of all independent variables, and term *support vector* any x_i such that $-C(1-\mu_i) < \chi_i^* < C\mu_i$. It can be shown that the constrained optimization corresponds to the search of a sphere S containing all points having unitary membership value, and having the property that the distance between its center and the remaining points is consistent w.r.t. their membership values. The role of C is that of ruling a trade-off between the radius of this sphere and the accuracy of the obtained model (see [9] for a discussion about the sensitivity w.r.t. this parameter). As a result, the quantity

$$R(x) = k(x,x) - 2\sum_{i=1}^{n} \chi_i^* k(x_i, x_i) + \sum_{i,j=1}^{n} \chi_i^* \chi_j^* k(x_i, x_j) \quad (11)$$

computes the distance of a generic point x from the center of S, and the latter will have a radius of length $R_1 = R(x_{\mathrm{SV}})$, being x_{SV} any of the support vectors.

The learnt sphere S will identify therefore with the *crisp support* of the fuzzy set, and an approximation $\hat{\mu}$ for μ_A can be easily induced mapping a point x to $\hat{\mu}(x) = 1$ whenever $R(x) \leq R_1$, and letting otherwise $\hat{\mu}(x)$ decrease towards zero as much as x is far from the center of S through application of a suitable *fuzzification function*. For the sake of concision, such a function will be described in terms of the *corrected distances* defined as the differences between R_1 and the distance of a point from the center of S (thus amounting to the distance of a point from the surface of the sphere). The experiments described in Sect. 5 will refer to the following fuzzification functions.

- *CrispFuzzifier*, associating null membership to all points not belonging to the crisp support of the set:

$$\hat{\mu}_{\mathrm{crisp}}(x) = \begin{cases} 1 & \text{if } R(x) \leq R_1, \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

- *LinearFuzzifier*, decreasing from 1 to 0 as the distance from the center of S ranges from R_1 to the maximum observed value (cfr. Figure 1(a)); more precisely, denoted by max the maximum corrected distance, the fuzzification has the followig form:

$$\hat{\mu}_{\mathrm{lin}}(x) = \begin{cases} 1 & \text{if } R(x) \leq R_1, \\ \frac{R_1 - x}{\max} - 1 & \text{if } R_1 < r \leq R_1 + \max, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

- *QuantileConstantPiecewiseFuzzifier*, computing a constant piecewise membership funtion whose steps are identified by the four empirical quartiles of corrected distances, and the height of the steps are respectively 1, 0.75, 0.5, 0.25 and 0 (cfr. Figure 1(b)); namely, denoted by q_1 and q_3 the first and third quantile, and by m the median:

$$\hat{\mu}_{\mathrm{qconst}}(x) = \begin{cases} 1 & \text{if } R(x) \leq R_1, \\ \frac{3}{4} & \text{if } R_1 < r \leq R_1 + q_1, \\ \frac{1}{2} & \text{if } R_1 + q_1 < r \leq R_1 + m, \\ \frac{1}{4} & \text{if } R_1 + m < r \leq R_1 + q_3, \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

- *QuantileLinearPiecewiseFuzzifier*, amounting to a piecewise linearization of *QuantileConstantPiecewiseFuzzifier* (cfr. Figure 1(c))):

$$\hat{\mu}_{\mathrm{qlin}}(x) = \begin{cases} 1 & \text{if } R(x) \leq R_1, \\ -\frac{r-R}{4q_1} + 1 & \text{if } R_1 < r \leq R_1 + q_1, \\ -\frac{r-R-q_1}{4(m-q_1)} + \frac{3}{4} & \text{if } R_1 + q_1 < r \leq R_1 + m, \\ -\frac{r-R-m}{4(q_3-m)} + \frac{1}{2} & \text{if } R_1 + m < r \leq R_1 + q_3, \\ -\frac{r-R-q_3}{4(\max-q_3)} + \frac{1}{4} & \text{if } R_1 + q_3 < r \leq R_1 + \max, \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

- *ExponentialFuzzifier*, computing an exponentially decaying membership function belonging to the family

$$\hat{\mu}_{\exp,\alpha}(x) = \begin{cases} 1 & \text{if } R(x) \leq R_1 \\ \exp\left(\frac{\ln \alpha}{q_\alpha}(r - R_1)\right) & \text{otherwise,} \end{cases} \quad (16)$$

parametrized on the value $\alpha \in [0,1]$ insuring that $\hat{\mu}_{\exp,\alpha}(R_1 + q_\alpha) = \alpha$, where q_α denotes the α-quantile of corrected distances (see Figure 2).

5 EXPERIMENTS

We based the inference of possibilistic scores of candidate OWL axioms on an enhancement of the scheme proposed in Sect. 4[1]. The first step consisted in building a knowledge base as follows:

- starting from a set of $m = 722$ axioms in a set $A_{\mathrm{start}} = \{\phi_1, \ldots, \phi_n\}$, we considered all formulas in A_{start}, as well as their negations, getting $A = \{\phi \ \forall \phi \in A_{\mathrm{start}}\} \cup \{\neg\phi \ \forall \phi \in A_{\mathrm{start}}\}$, including therefore a total of $n = 2m = 1444$ formulas;
- we computed the similarity function sim described in Sect. 3 for all pairs of formulas in A, obtaining a Gram matrix K;
- we considered the possibility of each $\phi_i \in A$, previously computed using the heuristic described in Sect. 2, as a degree of membership in a fuzzy set (to be understood as the set of "valid", "acceptable", or "likely" formulas), henceforth identified as the membership value $\mu_i = \Pi(\phi_i)$.

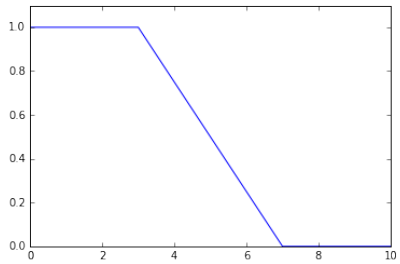
The axioms of A_{start} are `SubClassOf` axioms involving atomic classes which were exactly scored against DBpedia; altogether, computing their scores required a little less than 290 days of CPU time on quite a powerful machine [16].

We subsequently applied the procedure described in Sect. 4, inferring a function $\hat{\mu}$ approximating the membership/possibility of any formula. Evaluation and model selection have been performed through the following repeated holdout scheme:[2]

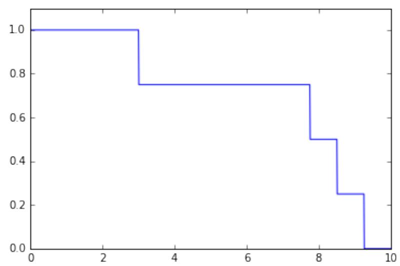
- formulas and memberships in the knowledge base have been shuffled and subsequently divided into three sets

[1] Code and data to replicate all experiments is available at https://github.com/dariomalchiodi/SAC2018

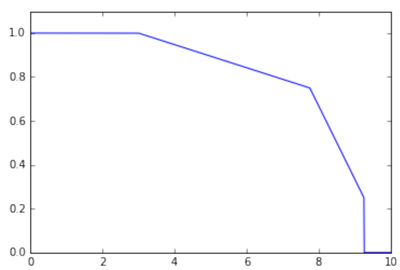
[2] We could not apply more sophisticated schemes such as cross-validation, because the relative small size of the obtained training sets did not suffice in order to infer meaningful models.



(a) $\hat{\mu}_{\mathrm{lin}}$

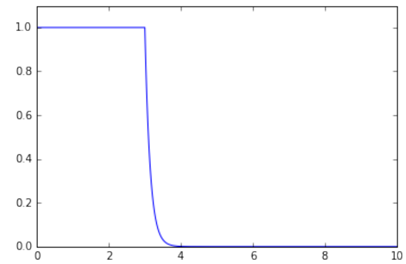


(b) $\hat{\mu}_{\mathrm{qconst}}$

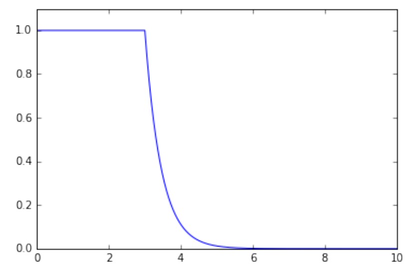


(c) $\hat{\mu}_{\mathrm{qlin}}$

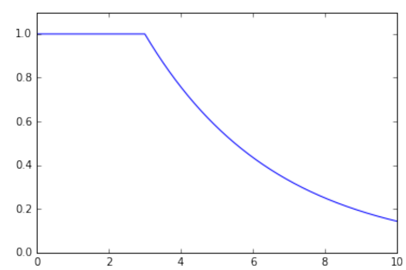
Figure 1: Graph of linear (a), piecewise constant (b), and piecewise linear (c) fuzzification function.



(a) $\hat{\mu}_{\mathrm{exp},0.001}$



(b) $\hat{\mu}_{\mathrm{exp},0.07}$



(c) $\hat{\mu}_{\mathrm{exp},0.5}$

Figure 2: Graph of exponential fuzzifiers with parameters 0.001 (a), 0.07 (b), and 0.5 (c).

respectively devoted to training, model selection,[3] and model validation (these sets contained the 80%, 10%, and 10% of the original data, and any formula was assigned to the same set of its negation, so as to insure the subsequent computability of ARIs);

- we inferred from the training set an approximate membership function $\hat{\mu}$ for each C in the grid $\{0.005, 0.007, 0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1, 10, 100\}$,[4] measuring its accuracy in terms of RMSE over the model selection set;
- we evaluated the function giving rise to the lowest error using RMSE, standard deviation, and median of accuracies computed on membership values and on possibility scores over the model validation set.

[3]The model selection procedure, fine-tuning fuzzification functions, only involved the choice of the tradeoff parameter C, as we used a precomputed kernel based on the values returned by sim.

[4]This grid was selected empirically after some experimentations.

For each fuzzification function proposed in Sect. 4 we iterated the above procedure 10 times. Table 2 summarizes the results in terms of RMSE, standard deviation and median of squared errors observed w.r.t. (i) *membership*, thus referring to the ability of $\hat{\mu}$ in predicting the values μ_i, and (ii) ARI, computed for each pair $(\phi, \neg\phi)$ of axioms as the value $\mu(\phi) - \mu(\neg\phi) = \Pi(\phi) - \Pi(\neg\phi)$ (cf. Equation 3).

As the best results were obtained using a crisp fuzzifier, that is learning a *crisp* rather than a fuzzy set, we decided to repeat all experiments using the original SV one-class classifier [3] as base learner. The results of this new round of experiments, illustrated in Table 3, are in line, although with slightly lower performance, with those previously obtained. This suggests that, although there is a cluster of elements naturally gathering in a crisp set, information about fuzzy membership allows to better single out this set.

Note that performances, rather poor in terms of RMSE, are quite good when considering median errors. The histograms

Table 2: Results of the membership learning procedure, in terms of root mean square error (RMSE), standard deviation (STD), and median (Median) for membership and ARI values inferred in 10 repeated holdout experiments.

	$\hat{\mu}_{\text{crisp}}$	$\hat{\mu}_{\text{lin}}$	$\hat{\mu}_{\text{qconst}}$	$\hat{\mu}_{\text{qlin}}$	$\hat{\mu}_{\text{exp},0.001}$	$\hat{\mu}_{\text{exp},0.005}$
Membership						
RMSE	3.77E-01	3.13E-01	3.14E-01	3.14E-01	3.62E-01	3.92E-01
Median	0.00E+00	4.09E-03	7.85E-03	1.38E-02	5.00E-19	4.19E-14
STDEV	3.33E-01	1.91E-01	1.79E-01	1.64E-01	3.21E-01	3.44E-01
ARI						
RMSE	5.72E-01	4.90E-01	4.91E-01	4.91E-01	5.51E-01	6.04E-01
Median	7.56E-04	6.05E-02	8.40E-02	1.07E-01	8.45E-04	8.94E-04
STDEV	6.11E-01	3.80E-01	3.70E-01	3.34E-01	5.95E-01	7.00E-01

	$\hat{\mu}_{\text{exp},0.070}$	$\hat{\mu}_{\text{exp},0.100}$	$\hat{\mu}_{\text{exp},0.200}$	$\hat{\mu}_{\text{exp},0.300}$	$\hat{\mu}_{\text{exp},0.400}$	$\hat{\mu}_{\text{exp},0.500}$
Membership						
RMSE	3.86E-01	3.92E-01	3.39E-01	3.37E-01	3.08E-01	3.41E-01
Median	1.28E-06	1.53E-05	2.35E-04	8.54E-03	3.22E-02	7.54E-02
STDEV	3.33E-01	3.31E-01	2.57E-01	2.16E-01	1.61E-01	1.54E-01
ARI						
RMSE	5.97E-01	6.09E-01	5.28E-01	5.32E-01	4.86E-01	5.35E-01
Median	1.35E-03	3.03E-03	1.20E-02	4.29E-02	7.56E-02	1.70E-01
STDEV	6.75E-01	6.54E-01	5.20E-01	4.63E-01	3.77E-01	3.52E-01

Table 3: Results of the membership learning procedure using one-class SV classifiers. Same notations of Table 2.

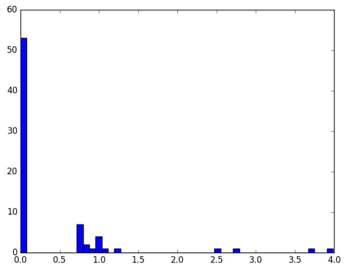
	$\hat{\mu}_{\text{crisp}}$	$\hat{\mu}_{\text{lin}}$	$\hat{\mu}_{\text{qconst}}$	$\hat{\mu}_{\text{qlin}}$	$\hat{\mu}_{\text{exp},0.001}$	$\hat{\mu}_{\text{exp},0.005}$
Membership						
RMSE	3.86E-01	3.46E-01	3.54E-01	3.55E-01	3.84E-01	4.15E-01
Median	0.00E+00	2.52E-04	1.08E-04	3.80E-03	4.61E-18	3.76E-07
STDEV	3.37E-01	2.68E-01	2.54E-01	2.43E-01	3.38E-01	3.60E-01
ARI						
RMSE	5.73E-01	5.27E-01	5.23E-01	5.25E-01	5.77E-01	6.22E-01
Median	1.14E-03	2.80E-02	1.39E-01	1.19E-01	9.44E-04	1.51E-03
STDEV	5.37E-01	5.01E-01	3.95E-01	3.73E-01	5.58E-01	6.17E-01

	$\hat{\mu}_{\text{exp},0.070}$	$\hat{\mu}_{\text{exp},0.100}$	$\hat{\mu}_{\text{exp},0.200}$	$\hat{\mu}_{\text{exp},0.300}$	$\hat{\mu}_{\text{exp},0.400}$	$\hat{\mu}_{\text{exp},0.500}$
Membership						
RMSE	4.02E-01	3.78E-01	3.72E-01	3.45E-01	3.59E-01	3.88E-01
Median	2.74E-06	1.57E-05	1.65E-04	1.16E-02	4.37E-02	1.69E-02
STDEV	3.41E-01	3.21E-01	2.96E-01	2.35E-01	2.13E-01	2.26E-01
ARI						
RMSE	6.07E-01	5.68E-01	5.52E-01	5.08E-01	5.33E-01	5.57E-01
Median	2.58E-03	1.90E-03	1.18E-02	7.75E-02	1.38E-01	2.21E-01
STDEV	6.16E-01	5.42E-01	4.72E-01	3.62E-01	3.41E-01	2.65E-01

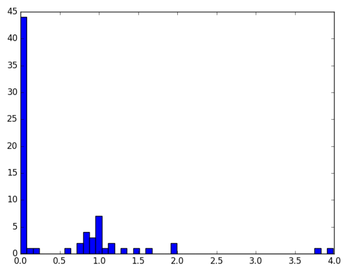
of the latter, in all experiments (see for instance Figure 3) suggest a mixture of two distributions, respectively for *easy* and *hard* to learn pairs of candidate axioms. In order to get furhter insights about this hypothesis, we adopted the following setting:

(1) for each pair of axioms and each fuzzifier we computed the average median error in all experiments selecting the pair,[5] and organized the results in a table having pairs as rows and fuzzifiers as columns;
(2) there is a non-negligible probability that a pair is never selected in the 10 iterations, and indeed for almost each pair of axioms there is one fuzzifier for which the previous operation cannot be carried out: we resolved

[5] Recall that in each experiment pairs are shuffled and only 10% of them are used in the model validation phase.



(a) $\hat{\mu}_{\mathrm{crisp}}$



(b) $\hat{\mu}_{\exp,0.001}$

Figure 3: Histograms of median errors for two of the used fuzzifiers in an iteration of the experiments.

the corresponding missing values in the table throug standard row-wise average imputation;

(3) we considered all rows of the post-processed table as points in a Euclidean space, to be clustered through K-means algorithm [6], letting the number k of clusters range from 2 to 9 and computing each time the silhoutette index [11] (whose values range from -1 and 1, the higher the better); Table 4 shows that indeed the best grouping corresponds to $k = 2$;

(4) finally, we considered the two groups of axioms in the best clustering and computed the histograms of ARI errors within each cluster: the results highlight a strong separation of the distributions, especially when the nonlinearity degree of the fuzzifier increases (Fig. 4 shows some examples of these histograms).

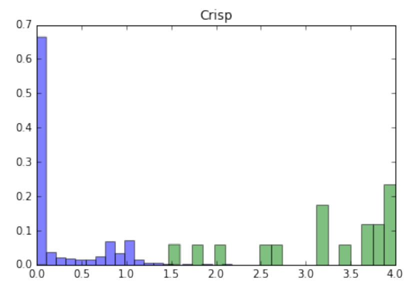
Thus there is strong evidence that axiom pairs belonging to these distinct distributions need to be learned separately. The positive members of the 17 pairs of "hard" axioms are listed in Table 5. Puzzlingly, they are almost impossible to tell apart intuitively from "easy" axioms, like `SubClassOf(dbo:ArchitecturalStructure dbo:Agent)` or `SubClassOf(dbo:Village dbo:Settlement)`.

6 CONCLUSIONS

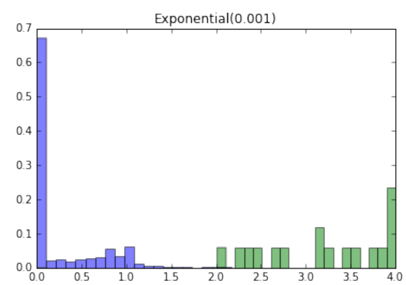
Within the emerging field of ontology learning, a promising task is that of investing resources for estimating the possibilistic score of a relatively small set of formulas and using

k	Silhouette index
2	0.834
3	0.707
4	0.631
5	0.632
6	0.594
7	0.591
8	0.563
9	0.531

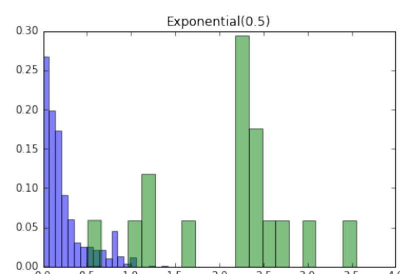
Table 4: Values of the silhouette index for different clusterizations of the points summarinzing the fuzzifiers average median error for each axiom pair.



(a) $\hat{\mu}_{\mathrm{crisp}}$



(b) $\hat{\mu}_{\exp,0.001}$



(c) $\hat{\mu}_{\exp,0.5}$

Figure 4: Histograms of ARI **errors within a grouping of two clusters for (a) a crisp fuzzifier, (b) an exponential fuzzifier of parameter** 0.001**, and (c) an exponential fuzzifier of parameter** 0.5**.**

Table 5: Positive members of the detected "hard" axiom pairs.

```
SubClassOf(dbo:ArchitecturalStructure gml:_Feature)
SubClassOf(schema:Product dbo:MeanOfTransportation)
SubClassOf(dbo:Eukaryote dbo:Animal)
SubClassOf(dbo:Library gml:_Feature)
SubClassOf(schema:School gml:_Feature)
SubClassOf(dbo:Racecourse gml:_Feature)
SubClassOf(dbo:WomensTennisAssociationTournament
                                            gml:_Feature)
SubClassOf(schema:Airport gml:_Feature)
SubClassOf(dbo:GovernmentAgency
                          schema:GovernmentOrganization)
SubClassOf(dbo:GovernmentAgency gml:_Feature)
SubClassOf(dbo:Venue dbo:Building)
SubClassOf(dbo:Venue dbo:Theatre)
SubClassOf(dbo:Venue gml:_Feature)
SubClassOf(dbo:YearInSpaceflight skos:Concept)
SubClassOf(dbo:Village gml:_Feature)
SubClassOf(dbo:ProtectedArea gml:_Feature)
SubClassOf(dbo:ComedyGroup foaf:Person)
```

the result as training data for learning a predictor to be henceforth used to get scores for new formulas with negligible computational costs.

In this paper we proposed a procedure aiming at inferring a predictor for the possibilistic score of candidate OWL subsumption axioms. The corresponding learning algorithm relies on a procedure mapping the inference task to the estimation of the membership function to a fuzzy set. In turn, the sample of axioms used in the training and testing phase should be equipped with their own possibilistic score, as well as with pairwise similarity values. After having proposed two heuristics devoted to compute this additional information, we applied the proposed procedure to a knowledge base and obtained a very good performance in terms of the accuracy of the induced predictors. The obtained results also highlighted the existence of a (relatively small) subset of axioms whose structure results harder to learn w.r.t. the rest of the knowledge base. Further research will be devoted to consider how different choices of the axiom similarity measure and of the estimation of the possibilistic score for training and testing data may impact on performances. Moreover, an extended experimentation phase involving a higher number of candidate axioms will allow to specifically investigate the detected hard axiom class.

REFERENCES

[1] Bruno Apolloni, Simone Bassis, Dario Malchiodi, and Witold Pedrycz. 2008. Interpolating Support Information Granules. *Neurocomputing* 71 (2008), 2433–2445.

[2] Bruno Apolloni, Domenico Iannizzi, Dario Malchiodi, and Witold Pedrycz. 2005. Granular Regression. (2005), 147–156.

[3] Asa Ben-Hur, David Horn, Hava T Siegelmann, and Vladimir Vapnik. 2001. Support vector clustering. *Journal of machine learning research* 2, Dec (2001), 125–137.

[4] Lorenz Bühmann and Jens Lehmann. 2012. Universal OWL Axiom Enrichment for Large Knowledge Bases. In *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings (Lecture Notes in Computer Science)*, Annette ten Teije, Johanna Völker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Aquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez (Eds.), Vol. 7603. Springer, 57–71. https://doi.org/10.1007/978-3-642-33876-2_8

[5] D. Dubois and H. Prade. 1988. *Possibility Theory—An Approach to Computerized Processing of Uncertainty.* Plenum Press, New York.

[6] Edward W Forgy. 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics* 21 (1965), 768–769.

[7] Jens Lehmann and Johanna Völker (Eds.). 2014. *Perspectives on Ontology Learning.* Studies on the Semantic Web, Vol. 18. IOS Press, Amsterdam. https://doi.org/10.3233/978-1-61499-379-7-i

[8] Alexander Maedche and Steffen Staab. 2001. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems* 16, 2 (2001), 72–79. https://doi.org/10.1109/5254.920602

[9] Dario Malchiodi and Witold Pedrycz. 2013. Learning Membership Functions for Fuzzy Sets through Modified Support Vector Clustering. In *Fuzzy Logic and Applications - 10th International Workshop, WILF 2013, Genoa, Italy, November 19-22, 2013. Proceedings (Lecture Notes in Computer Science)*, Francesco Masulli, Gabriella Pasi, and Ronald R. Yager (Eds.), Vol. 8256. Springer, 52–59.

[10] Margarita Mas, Miquel Monserrat, Joan Torrens, and Enric Trillas. 2007. A Survey on Fuzzy Implication Functions. *IEEE Trans. Fuzzy Systems* 15, 6 (2007), 1107–1121.

[11] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[12] Umberto Straccia and Matteo Mucci. 2015. pFOIL-DL: Learning (Fuzzy) EL Concept Descriptions from Crisp OWL Data Using a Probabilistic Ensemble Estimation. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing.* ACM, New York, NY, USA, 345–352.

[13] Heiner Stuckenschmidt. 2007. Partial Matchmaking using Approximate Subsumption. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada.* AAAI Press, 1459–1464.

[14] Andrea G. B. Tettamanzi, Catherine Faron-Zucker, and Fabien Gandon. 2015. Dynamically Time-Capped Possibilistic Testing of SubClassOf Axioms Against RDF Data to Enrich Schemas. In *K-CAP 2015. Proceedings of the 8th International Conference on Knowledge Capture, Palisades, NY, USA, October 07–10, 2015*, Ken Barker and José Manuel Gómez-Pérez (Eds.). ACM, New York, Article No.: 7. https://doi.org/10.1145/2815833.2815835

[15] Andrea G. B. Tettamanzi, Catherine Faron-Zucker, and Fabien Gandon. 2017. Possibilistic Testing of OWL Axioms Against RDF Data. *International Journal of Approximate Reasoning* 91 (December 2017), 114–130.

[16] Andrea G. B. Tettamanzi, Catherine Faron-Zucker, and Fabien L. Gandon. 2014. Testing OWL Axioms Against RDF Facts: A possibilistic approach. In *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings (Lecture Notes in Artificial Intelligence)*, Krzysztof Janowicz, Stefan Schlobach, Patrick Lambrix, and Eero Hyvönen (Eds.), Vol. 8876. Springer, 519–530. https://doi.org/10.1007/978-3-319-13704-9_39