Predicting the Possibilistic Score of OWL Axioms through Support Vector Regression

Dario Malchiodi[1], Célia da Costa Pereira[2], and Andrea G. B. Tettamanzi[2]

[1] Università degli Studi di Milano, Dipartimento di Informatica, Italy
dario.malchiodi@unimi.it
[2] Université Côte d'Azur, CNRS, I3S, France
{celia.pereira,andrea.tettamanzi}@unice.fr

Abstract. Within the context of ontology learning, we consider the problem of selecting candidate axioms through a suitable score. Focusing on subsumption axioms, this score is learned coupling support vector regression with a special similarity measure inspired by the Jaccard index and justified by semantic considerations. We show preliminary results obtained when the proposed methodology is applied to pairs of candidate OWL axioms, and compare them with an analogous inference procedure based on fuzzy membership induction.

1 Introduction and Related Works

Schema enrichment is one important ingredient of ontology learning [9]. In particular, in the context of the semantic Web, the increasing amount of Linked Data causes schema enrichment to be an emerging field of research. Its goal is that of automatizing the work of knowledge engineers by leveraging existing ontologies (typically expressed in OWL) and instance data (typically represented in RDF) [8]. The final aim is detecting meaningful patterns and learn schema axioms from existing instance data (facts) and their metadata, if available, using induction-based methods like the ones developed in inductive logic programming and data mining.

Many other researchers are convinced of the benefit of using enriched schemas to improve the quality of the reasoning process in the semantic Web. Fleischhacker and colleagues, for example, use statistical schema induction in the approach they proposed to enrich the schema of any RDF dataset with property axioms [6]. In the same line, Töpper and colleagues propose an approach in which they focus on the enrichment of the DBpedia ontology by using statistical methods [16]. Huitzil et al. explored the possibility of learning dataypes within ontologies [7]. Bühmann and colleagues proposed a light-weight method to enrich knowledge bases accessible via SPARQL endpoints with almost all types of OWL 2 axioms. The aim of their approach was to allow to create a schema in a semi-automatic way [4].

All these approaches to schema enrichment critically rely on (candidate) axiom scoring. 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 [14]) and some related heuristics have been proposed to score concept definitions in concept learning algorithms [13]. The most popular candidate axiom scoring heuristics proposed in the literature are based on statistical inference (see, e.g., [4]). 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, the heavy computational cost of the heuristics makes it hard to apply in practice, unless some implementation tricks are devised (e.g., time capping).

A promising alternative to the direct computation of the possibilistic score consists in training a surrogate model on a sample of candidate axioms for which the score had already been computed or is otherwise available, in order to be capable of *predicting* the score of a novel, unseen candidate axiom. This idea was recently proposed in [10], using an adaptation of support vector clustering for learning the membership functions for fuzzy sets.

In this work, we follow the same general scheme, but we apply support vector regression to obtain a simpler surrogate model and we compare our preliminary results to those obtained with the modified support vector clustering on the same 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 [15].

The paper is structured as follows: Sect. 2 gives some background on the possibilistic axiom scoring heuristics and on how the similarity between axioms is computed. In Sect. 3 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. 4. Some concluding remarks end the paper.

2   Background on Axiom Scoring and Similarity

The possibilistic axiom score we wish to be able to predict was 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, a degree of *possibility* $\Pi(\phi)$ and of *necessity* $N(\phi)$ for ϕ are computed based the *evidence* available contained in an RDF dataset $\mathcal{K}$.

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

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

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

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.

	$\psi = C \sqsubseteq D$	$\psi = C \not\sqsubseteq D$
$\phi = A \sqsubseteq B$	$\mathrm{sim}(\phi,\psi) = \dfrac{\|[A] \cap [B] \cup [C] \cap [D]\|}{\|[A] \cup [C]\|}$	$\mathrm{sim}(\phi,\psi) = \dfrac{\|[A] \cap [B] \cup [C] \cap \overline{[D]}\|}{\|[A] \cup [C]\|}$
$\phi = A \not\sqsubseteq B$	$\mathrm{sim}(\phi,\psi) = \dfrac{\|[A] \cap \overline{[B]} \cup [C] \cap [D]\|}{\|[A] \cup [C]\|}$	$\mathrm{sim}(\phi,\psi) = \dfrac{\|[A] \cap \overline{[B]} \cup [C] \cap \overline{[D]}\|}{\|[A] \cup [C]\|}$

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 proposed in [10] is that, if we can train a model to predict $\Pi(\phi)$ and $\Pi(\neg\phi)$, we have enough information to estimate $\mathrm{ARI}(\phi)$ without having to directly compute it.

Support vector regression, which we use to train a predictor of the possibility of OWL axioms, requires a kernel function which, for our purposes, may be viewed as a similarity measure between candidate axioms. To allow the comparison of the results, we adopt the semantic similarity measure, somehow reminiscent of the Jaccard index, proposed and justified in [10], to which the interested reader is referred. Given two subsumption axioms $A \sqsubseteq B$ and $C \sqsubseteq D$ and their negations $A \not\sqsubseteq B$ and $C \not\sqsubseteq D$, where A, B, C, and D are OWL class expressions, such similarity can be written as shown in Table 1, where $[E] = \{a : E(a)\}$ denotes the extension of class expression E in the RDF dataset at hand.

3 Support Vector Regression

In its simplest formulation, the extension of support-vector framework [5] to regression problems consists in considering a multiobjective optimization in which lines are scored according to their flatness and to their distance from a set of points. The tradeoff between these components is ruled via a parameter $C > 0$. Different flavors of this technique are defined in function of how the above mentioned distance is defined and penalized [12, 3]. In particular:

- ϵ-*insensitive* regression fixes a parameter $\epsilon > 0$ representing the width of a tube centered around the line: the loss is null for all points falling within this tube and equal to the distance w.r.t. its frontier otherwise;

− *ridge regression* considers a quadratic loss in terms of the distance between line and points.

This basic scheme is extended to nonlinear regression using kernel methods, that is mapping the original points via a nonlinear transformation onto a higher-dimensional space, and trying to find a linear regression therein. In particular, as the original points occur in the problem formalization only in form of dot product computations, the only change amounts to replacing all such occurrences with invocations of a suitable kernel function amounting for computing the dot product of the images of its arguments. In the next section we detail how the heuristic of Sect. 2 can be used in order to build a kernel function whose arguments are axioms.

4 Experiments

We adopted the approaches to support vector regression resumed in Sect. 3 to the problem of building a predictor for the ARI value for candidate OWL axioms on the basis of its measurements on a limited set of formulas, using the methodology described in Sect. 2. In particular, we used as a reference the same settings as in [10], briefly summarized hereafter:[1]

− we considered $m = 722$ `SubClassOf` axioms involving atomic classes which were exactly scored against DBpedia,[2] as well as their negations, in a set A (thus a total of $n = 2m = 1444$ formulas), computing sim for each pair;
− we took the possibility of each formula $\phi_i \in A$ previously computed using the heuristic described in [15], henceforth identified as a value $\mu_i = \Pi(\phi_i)$ of the "acceptability" of ϕ_i as an axiom.

We used the sim values as kernel computations and the possibility values as target to be predicted using the two variants of support vector regression described in previous section. Given the small size of available data, we resorted to iterating ten times the following holdout scheme in order to assess the generalization ability of the inferred predictors. Training, validation, and test sets containing the 80%, 10%, and 10% of original data, respectively, have been obtained after shuffling the dataset.[3] Validation involved in both cases the tradeoff parameter C, and the tube width when considering ϵ-insensitive regression.[4] Model selection was guided by RMSE accuracy. Table 2 summarizes the results on test sets in terms of RMSE, median, and standard deviation of the corresponding errors, both considering the ability of the inferred model to predict: (i) the acceptability $\mu_i = \Pi(\phi_i)$ of a formula, and (ii) its ARI according to (1). The ridge regression variant highlights better results w.r.t. both performance metrics with essentially comparable variability, although only ARI was deemed in [10] as a key indicator.

---

[1] Code and data to replicate all experiments is available at `https://github.com/dariomalchiodi/SUM2018`.
[2] The computation of the exact score took 290 CPU days on a 12 6-core CPU machine.
[3] Each pair $(\phi, \neg\phi)$ was assigned to a same set in order to be able to compute its ARI.
[4] After some experimentation these parameters were selected within a grid considering all magnitudes between 10^{-3} and 10^4 and between 10^{-2} and 10^5, respectively.

Table 2. Results of acceptability and ARI learning using ϵ-insensitive regression, ridge regression, and fuzzy inference in 10 repeated holdout experiments measuring root mean square (RMSE), median (Median), and standard deviation (STDEV) of errors.

Method	Acceptability			ARI			Time (mins.)
	RMSE	Median	STDEV	RMSE	Median	STDEV	
ϵ-insensitive	4.83e-01	2.50e-01	3.94e-02	8.46e-01	9.55e-01	4.11e-01	4
ridge	3.89e-01	8.58e-02	2.67e-01	6.28e-01	1.91e-01	6.37e-01	2
fuzzy	3.08e-01	0.00e+00	1.54e-01	4.86e-01	7.56e-04	3.34e-01	500

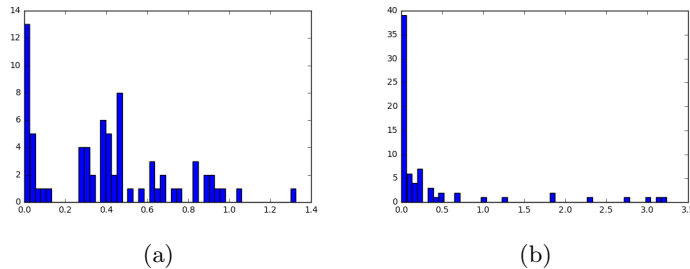(a)   (b)

Fig. 1. Histograms of median errors for two iterations of the holdout scheme when using ridge regression, respectively highliting the presence and the absence of a mix of two distributions.

Table 2 also compares the results with those of an analogous experiment based on a tailored procedure interpreting μ_is as membership values to a fuzzy set to be learned [10].[5] It is clear that the original results outperform the presented ones in terms of the proposed metrics. Things are radically different if we take into account also the time dimension: the eight hours needed to train and tune the fuzzy-based system — already incomparable to the 290 days required to produce the initial ARI labels — are further reduced to two minutes when using ridge regression. This amounts to a speed factor of roughly 0.004 with an equal factor in performance degradation. The ridge regression results also share with the original approach a better performance of the median w.r.t. RMSE, although with a smaller intensity. This led to conjecturing the presence of *easy* and *hard* to learn formulas, but this conjecture is less evident now: indeed, only in around half of the holdout iterations the histogram of median errors highlights a mixture of two error distributions, and anyhow they appear not strongly separated (see Fig. 1 for an example of both cases). Puzzlingly, a very strong separation between two median error distribution is always obtained when considering ϵ-insensitive regression, despite its lower performances. Thus we repeated the experiment proposed in [10] with the aim of testing this conjecture, computing the average median error in all iterations and finding its best clusterization in terms of silhouette index [11]. More precisely, we considered:

---

[5] This procedure is parametrized on the choice of different shapes for the fuzzy set membership function: the table reports for each column the best obtained result.

(i) both regression methods, (ii) only ϵ-insensitive regression, and (iii) only ridge regression. The best clusterization consisted of two groups only in (ii) and (iii). This led us to further inspect the two classes of candidate axioms only in terms of ridge regression. Specifically, we found 30 hard axioms with an overlap of around 40% with those found in the original paper. Such axioms are listed in Table 3; those that were also in [10] are marked with an asterisk.

Table 3. Positive members of the detected "hard" axiom pairs.

```
SubClassOf(schema:Product dbo:MeanOfTransportation) *
SubClassOf(dbo:Chancellor dbo:Person)
SubClassOf(schema:School gml:_Feature) *
SubClassOf(dbo:BeautyQueen dbo:Person)
SubClassOf(dbo:InformationAppliance schema:Person)
SubClassOf(dbo:Racecourse gml:_Feature) *
SubClassOf(dbo:WomensTennisAssociationTournament skos:Concept)
SubClassOf(dbo:VolleyballCoach dbo:Person)
SubClassOf(dbo:VolleyballCoach owl:Thing)
SubClassOf(dbo:VolleyballCoach foaf:Person)
SubClassOf(dbo:Presenter dbo:RadioHost)
SubClassOf(dbo:Venue gml:_Feature) *
SubClassOf(dbo:YearInSpaceflight skos:Concept) *
SubClassOf(dbo:ComedyGroup schema:Organization)
SubClassOf(dbo:ComedyGroup foaf:Person) *
```

5   Conclusions

We have applied support vector regression to the task of predicting the possibilistic score of candidate OWL subsumption axioms. Ridge regression gives better results than ϵ-insensitive regression. A comparison with a previous proposal using a modified support vector clustering for learning fuzzy sets shows that the regression approach allows for a faster training time fairly scaling with performance degradation. Our results also confirm the existence of a small subset of axioms that are much harder to score than the rest; however, this subset appears to depend, at least to some extent, on the method used to predict the score. Future work includes trying other prediction methods [2, 1], but also reformulating the scoring problem as a binary classification problem, which would suit the needs of schema enrichment equally well.

References

1. Apolloni, B., Bassis, S., Malchiodi, D., Pedrycz, W.: Interpolating support information granules. Neurocomputing 71, 2433–2445 (2008)

2. Apolloni, B., Iannizzi, D., Malchiodi, D., Pedrycz, W.: Granular regression. In: Apolloni, B., Marinaro, M., Nicosia, G., Tagliaferri, R. (eds.) Neural Nets. 16th Italian Workshop on Neural Nets, WIRN 2005 and International Workshop on Natural and Artificial Immune Systems, NAIS 2005. Vietri sul Mare, Italy, June 2005. Lecture Notes in Computer Science 3931, Springer (2005)

3. Apolloni, B., Malchiodi, D., Valerio, L.: Relevance regression learning with support vector machines. Nonlinear Analysis 73, 2855–2867 (2010)

4. Bühmann, L., Lehmann, J.: Universal OWL axiom enrichment for large knowledge bases. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Aquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7603, pp. 57–71. Springer (2012)

5. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)

6. Fleischhacker, D., Völker, J., Stuckenschmidt, H.: Mining RDF data for property axioms. In: Meersman, R., Panetto, H., Dillon, T.S., Rinderle-Ma, S., Dadam, P., Zhou, X., Pearson, S., Ferscha, A., Bergamaschi, S., Cruz, I.F. (eds.) On the Move to Meaningful Internet Systems: OTM 2012, Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2012, Rome, Italy, September 10-14, 2012. Proceedings, Part II. Lecture Notes in Computer Science, vol. 7566, pp. 718–735. Springer, Berlin (2012)

7. Huitzil, I., Straccia, U., Díaz-Rodríguez, N., Bobillo, F.: Datil: Learning fuzzy ontology datatypes. In: Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 100–112. Springer (2018)

8. Lehmann, J., Völker, J. (eds.): Perspectives on Ontology Learning, Studies on the Semantic Web, vol. 18. IOS Press, Amsterdam (2014)

9. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems 16(2), 72–79 (2001)

10. Malchiodi, D., Tettamanzi, A.G.B.: 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 (2018)

11. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987), `http://www.sciencedirect.com/science/article/pii/0377042787901257`

12. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and computing 14(3), 199–222 (2004)

13. Straccia, U., Mucci, M.: 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. pp. 345–352. ACM, New York, NY, USA (2015)

14. Stuckenschmidt, H.: Partial matchmaking using approximate subsumption. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada. pp. 1459–1464. AAAI Press (2007)

15. Tettamanzi, A.G.B., Faron-Zucker, C., Gandon, F.: Possibilistic testing of OWL axioms against RDF data. International Journal of Approximate Reasoning 91, 114–130 (December 2017)

16. Töpper, G., Knuth, M., Sack, H.: Dbpedia ontology enrichment for inconsistency detection. In: I-SEMANTICS. pp. 33–40. ACM (2012)