International Journal of Approximate Reasoning 92 (2018) 340–362

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Approximate classification with web ontologies through evidential terminological trees and forests [☆]

Giuseppe Rizzo [*], Nicola Fanizzi [**], Claudia d'Amato [**], Floriana Esposito

LACAM – Dipartimento di Informatica — Università degli Studi di Bari "Aldo Moro", Campus Universitario, Via Orabona 4, 70125 Bari, Italy

ARTICLE INFO

Article history:
Received 15 February 2017
Received in revised form 12 October 2017
Accepted 23 October 2017
Available online 31 October 2017

Keywords:
Ontologies
Logic decision trees
Dempster–Shafer theory
Instance classification

ABSTRACT

In the context of the Semantic Web, assigning individuals to their respective classes is a fundamental reasoning service. It has been shown that, when purely deductive reasoning falls short, this problem can be solved as a prediction task to be accomplished through inductive classification models built upon the statistical evidence elicited from ontological knowledge bases. However also these data-driven alternative classification models may turn out to be inadequate when instances are unevenly distributed over the various targeted classes To cope with this issue, a framework based on logic decision trees and ensemble learning is proposed. The new models integrate the *Dempster–Shafer theory* with learning methods for *terminological decision trees* and *forests*. These enhanced classification models allow to explicitly take into account the underlying uncertainty due to the variety of branches to be followed up to classification leaves (in the context of a single tree) and/or to the different trees within the ensemble model (the forest). In this extended paper, we propose revised versions of the algorithms for learning *Evidential Terminological Decision Trees* and *Random Forests* considering alternative heuristics and additional evidence combination rules with respect to our former preliminary works. A comprehensive and comparative empirical evaluation proves the effectiveness and stability of the classification models, especially in the form of ensembles.

1. Introduction

Sharing knowledge that is encoded along formal ontologies, thus enabling rich reasoning capabilities, plays a key role in the context of the *Semantic Web* (SW). However, standard deductive inference mechanisms sometimes show their limitations because of the inherent incompleteness of the ontological knowledge bases combined with the adoption of an open-world semantics, which is natural in such a Web-scale heterogeneous and distributed context.

In order to tackle the consequences of these distinctive aspects, alternative forms of reasoning, based on statistical models that can be induced through data-driven methods, have been introduced for performing various tasks such as *concept retrieval* and *query answering* [1] more effectively. It has been shown that these tasks have been cast as *classification*

[☆] This paper is part of the Virtual special issue on Uncertainty Reasoning for the Web, Edited by Fernando Bobillo, Kenneth J. Laskey, Trevor Martin, Matthias Nickles.

[*] Principal corresponding author.

[**] Corresponding author.

E-mail addresses: giuseppe.rizzo1@uniba.it (G. Rizzo), nicola.fanizzi@uniba.it (N. Fanizzi), claudia.damato@uniba.it (C. d'Amato), floriana.esposito@uniba.it (F. Esposito).

https://doi.org/10.1016/j.ijar.2017.10.019

problems, which amount to deciding the membership of an individual with respect to a target concept, and they have been solved through inductive learning methods exploiting statistical regularities in the underlying knowledge base. Specifically, the resulting models have been used by approximate classification procedures applied to the knowledge bases also in combination with deductive inference services [2]. The application of these methods has shown interesting results such as the ability to synthesize new concepts and/or produce inductive classification models inspired by *Inductive Logic Programming* (ILP) like *terminological decision trees* [3], i.e. *logic decision trees* [4,5] whose inner node tests are expressed in terminological languages (that is Description Logics [6]). Additionally, exploiting such statistical models, non logically-derivable yet still consistent assertional knowledge may be suggested.

However, such alternative methods and models have also revealed some shortcomings. One of the issues is that they do not allow an explicit representation of uncertainty to be specifically exploited for managing those cases when the classification procedure assigns an uncertain membership. To better tackle these cases, an enhanced model, called *evidential terminological decision tree* has been devised, by integrating primitives of the *Dempster–Shafer Theory* [7]. The main advance with respect to terminological decision trees regards the heuristic used to select the concept installed into inner nodes (based on the *non-specificity measure* [8] rather than the classic measures stemming from *information gain*) and the classification procedure (that explores all the possible paths departing from a node with an uncertain test result).

Another issue concerns the distribution of the training data. In general, the individuals that are known (or can be logically assessed as) positive and negative instances for a given target concept (that is those that are instances of a target concept or of the negated target concept) may not be equally distributed. This skewness may be noticeably larger when considering individuals whose membership cannot be assessed by reasoning under an open-world semantics. This class-imbalanced setting may affect the model, resulting in poor performances. Various methods have been devised to tackle the general unbalance learning problem (see [9] for a survey of the various approaches). As regards the specific task of learning instance classification models for inductive query answering on SW knowledge bases, we investigated the adoption of methods for *ensemble models* [10] that are made up of a certain number of classifiers, trained by the so-called *weak learners*, and whose final prediction results from the combination of the predictions made by each classifier. Specifically, the combination is given by a specific rule playing the role of the *meta-learner*. Particularly, we proposed an algorithm for inducing *terminological random forests* [10] that extends *(First Order) random forests* [11,12] with the use of Description Logics: the model is an ensemble of terminological decision trees [3].

Employing these models, the membership of a test individual w.r.t. a target concept is decided according to a majority vote rule (although various other strategies for combining predictions have been proposed [13–15]): each classifier equally contributes to the final decision returning a vote in favor of a single membership. In this way, some other aspects are not considered explicitly, such as the uncertainty about the single membership-label assignments and the disagreement that may intervene among weak learners. Particularly, the latter issue is crucial for the performance of ensemble models [16]: using the aforementioned type of forests, we noted that most misclassification cases were related to situations in which votes are evenly distributed with respect to the admissible labels. A weighted voting procedure may be an alternative strategy to mitigate the problem, but it requires a criterion for setting the weights.

In this sense, introducing a meta-learner which can manipulate the *soft* predictions made by each classifier (i.e. a prediction with a confidence measure for each membership value) rather than *hard predictions* (where only the predicted label is returned) may be a solution. Adopting the random forests as ensembles, this can be accomplished by considering evidential terminological decision trees [7] as base models. Dempster–Shafer theory has already been used in combination with ensemble learning procedures (e.g. see [17]). However, most of the methods apply to problems that involve simpler knowledge representations. Additionally, none of them has been employed for predicting assertions on ontological knowledge bases.

Therefore, we further extended the model proposing a framework for the induction of *Evidential Terminological Random Forests* for ontological knowledge bases [18]. Employing evidential terminological decision trees, the approach does not require the computation of decision templates. After the induction of the forest, new individuals are classified by combining the evidence on the membership prediction made by each tree through Dempster's rule [19].

However, we noted that the proposed framework had some limitations [7,18]. Firstly, the heuristic to select the most promising label adopted by evidential terminological decision tree learning algorithm did not consider the presence of conflicting evidence. Secondly, the combination rule represented a bottleneck of the classification step: therefore it is important to investigate alternative solutions for improving the efficiency of the classification. Thirdly, the size of evidential terminological random forests seemed not to affect the predictiveness of the ensemble model (due to a weak diversification of the ensemble) but represented a source of complexity during the classification step.

Consequently, in this paper we extended the framework for learning evidential terminological decision trees and random forests along the following directions:

- we used different heuristics based on other total uncertainty measures (than the sole non-specificity measure) to drive the selection of the concepts to be installed into the nodes of evidential terminological decision trees;
- we used further combination rules to pool the evidence obtained by traversing each tree;
- we used further combination rules as meta-learner for evidential terminological random forests;
- we set up a comprehensive and comparative experimental evaluation showing the effectiveness of the proposed extensions when performing inductive instance retrieval.