Accepted Manuscript

Tree-based models for inductive classification on the Web Of Data

Giuseppe Rizzo, Claudia d'Amato, Nicola Fanizzi, Floriana Esposito

PII: S1570-8268(17)30017-3
DOI: http://dx.doi.org/10.1016/j.websem.2017.05.001
Reference: WEBSEM 433

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received date : 31 July 2016
Revised date : 15 March 2017
Accepted date : 1 May 2017

Please cite this article as: G. Rizzo, et al., Tree-based models for inductive classification on the Web Of Data, *Web Semantics: Science, Services and Agents on the World Wide Web* (2017), http://dx.doi.org/10.1016/j.websem.2017.05.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

*Manuscript
Click here to view linked References

Tree-based Models for Inductive Classification on the *Web Of Data*

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

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

Abstract

The Web of Data, which is one of the dimensions of the Semantic Web (SW), represents a tremendous source of information, which motivates the increasing attention to the formalization and application of machine learning methods for solving tasks such as concept learning, link prediction, inductive instance retrieval in this context. However, the Web of Data is also characterized by various forms of uncertainty, owing to its inherent incompleteness (missing information, uneven data distributions) and noise, which may affect open and distributed architectures. In this paper, we focus on the inductive instance retrieval task regarded as a classification problem. The proposed solution is a framework for learning *Terminological Decision Trees* from examples described in an ontological knowledge base, to be used for performing instance classifications. For the purpose, suitable pruning strategies and a new prediction procedure are proposed. Furthermore, in order to tackle the class-imbalance distribution problem, the framework is extended to ensembles of Terminological Decision Trees called *Terminological Random Forest*s. The proposed framework has been evaluated, in comparative experiments, with the main state of the art solutions grounded on a similar approach, showing that: 1) the employment of the formalized pruning strategies can improve the model predictiveness; 2) *Terminological Random Forest*s outperform the usage of a single *Terminological Decision Tree*, particularly when the knowledge base is endowed with a large number of concepts and roles; 3) the framework can be exploited for solving related problems, such as predicting the values of given properties with finite ranges.

Keywords: inductive query answering, membership prediction, Web ontologies, decision tree, random forest, concept learning, imbalance learning

1. Introduction

In the perspective of the Semantic Web (SW) as a Web Of Data, the *Linked Data* initiative plays a crucial role. It federates a large number of interlinked datasets in a standard format whose semantics is encoded through formal ontologies for plenty of domains, which are accessible through the Web infrastructure, thus ensuring a convenient form for publishing or accessing open data and a superior (semantic) manageability through suitable tools [1].

In this context, a fundamental service, similarly to database management systems, is relational query answering, as a means for assessing properties of interest of individual resources through suitable endpoints. For example, in an academic scenario, given linked data sources that adopt publicly available Web ontologies for such a domain as their vocabularies, a typical query may require determining a person's research interests, his/her co-authorship with other researchers or his/her affiliation to a specific research group. To this purpose, standard technologies such as SPARQL[1] are generally exploited. These services are essentially grounded on pattern matching capabilities that are often inadequate to cope with the issues posed by forms of uncertainty that are inherently related to the distributed architecture of the data sources. This weakness may affect the quality of the answers in terms of *precision* and *completeness*. Especially when also deductive reasoning capabilities are called into play, cases of conflicting information may originate from the diverse quality of the ontologies involved, depending on the axioms in the terminologies and/or by specific assertions made available upon federated data sources. Furthermore, data are expressed in terms of formal vocabularies defined as Web ontologies whose representation and semantics is established on *Description Logics* (DL) [2], a family of languages characterized by the adoption of the *Open World Assumption* (OWA) in the related reasoning services. As a result, the membership of an individual resource w.r.t. a given class or its value for a given property cannot always be ascertained even with the support of a reasoner. A solution borrowed from other multi-relational contexts, such as (deductive) databases and logic programming, amounts to making the *Closed World Assumption* (CWA), i.e. presuming that the current state of knowledge is complete hence, for example, allowing to deem a fact as false

*Principal corresponding author
**Corresponding author
Email addresses: giuseppe.rizzo1@uniba.it (Giuseppe Rizzo), claudia.damato@uniba.it (Claudia d'Amato), nicola.fanizzi@uniba.it (Nicola Fanizzi), floriana.esposito@uniba.it (Floriana Esposito)

[1] http://www.w3.org/TR/rdf-sparql-query/

Download English Version:

https://daneshyari.com/en/article/6950485

Download Persian Version:

https://daneshyari.com/article/6950485

Daneshyari.com