Chapter 104
Pattern Based Feature Construction in Semantic Data Mining

Agnieszka Ławrynowicz
Poznan University of Technology, Poland

Jędrzej Potoniec
Poznan University of Technology, Poland

ABSTRACT

The authors propose a new method for mining sets of patterns for classification, where patterns are represented as SPARQL queries over RDFS. The method contributes to so-called semantic data mining, a data mining approach where domain ontologies are used as background knowledge, and where the new challenge is to mine knowledge encoded in domain ontologies, rather than only purely empirical data. The authors have developed a tool that implements this approach. Using this the authors have conducted an experimental evaluation including comparison of our method to state-of-the-art approaches to classification of semantic data and an experimental study within emerging subfield of meta-learning called semantic meta-mining. The most important research contributions of the paper to the state-of-art are as follows. For pattern mining research or relational learning in general, the paper contributes a new algorithm for discovery of new type of patterns. For Semantic Web research, it theoretically and empirically illustrates how semantic, structured data can be used in traditional machine learning methods through a pattern-based approach for constructing semantic features.

INTRODUCTION

Pattern discovery is a fundamental data mining task. It deals with the automatic detection of patterns in data. *Pattern* is any regularity, relation or structure inherent in some source of data (Shawe-Taylor & Cristianini, 2004). Various methods have been proposed for finding patterns in a variety of forms such as item sets, association rules, correlations, sequences, episodes etc. From the point of view of this paper, we are interested in structured domains, where data is represented

DOI: 10.4018/978-1-4666-9562-7.ch104

in complex forms like relational databases, logic programs, and in particular semantic data such as ontology-based knowledge bases or *Linked Open Data (LOD)*[1].

Relational pattern discovery has been investigated since the development of WARMR (Dehaspe & Toivonen, 1999), an algorithm for mining patterns using the Datalog subset of first-order logic as the representation language for data and patterns. This has been followed by subsequently proposed relational pattern mining algorithms such as FARMER (Nijssen & Kok, 2001) or c-armr (De Raedt & Ramon, 2004). They can all be classified under *Inductive Logic Programming (ILP)* (Nienhuys-Cheng & Wolf, 1997) methods since they use subsets of logic programs as the representation language.

With the rise of the *Semantic Web* (Berners-Lee, Hendler, & Lassila, 2001), also called *Web of Data*, an interest has grown in employing languages, and knowledge representation formalisms underpinning the Semantic Web in data mining. This interest is motivated by increase of popularity, number and size of such semantic data sources as LOD (containing billions of pieces of data linked together[2]) that require statistical approaches able to handle Semantic Web knowledge representation formalisms. These formalisms include logic-based ontology languages such as *description logics (DLs)* (Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2003) that constitute the formalism underlying the standard ontology language for the Web, the *Web Ontology Language (OWL)* (McGuinness & van Harmelen, 2004). In this line, in (Lisi & Esposito, 2008) the foundations have been laid of an extension of relational learning, called *onto-relational learning*, to account for ontologies. Fanizzi, d'Amato, and Esposito (2010) propose the term *ontology mining* for all such activities that allow to discover hidden knowledge from ontological knowledge bases, by possibly using only a sample of data. Finally, Kralj-Novak, Vavpetic, Trajkovski, and Lavrac (2009) coined the term *semantic data mining*[3] to denote a data mining approach where domain ontologies are used as background knowledge, and where the new challenge is to mine knowledge encoded in domain ontologies, rather than to mine purely empirical data. The above-mentioned interest has been reflected in the development of relevant pattern mining algorithms, firstly onto-relational ones like SPADA (Lisi & Malerba, 2004), SEMINTEC (Józefowska, Ławrynowicz, & Łukaszewski, 2010) or AL-QuIn (Lisi F.A., 2011), and subsequently fully based on a description logic based ontology language like the algorithm Fr-ONT (Ławrynowicz & Potoniec, 2011).

In recent years, a topic of using patterns in predictive models has drawn a lot of attention (Bringmann, Nijssen, & Zimmermann, 2009). Especially in complex, structured domains, such as graphs and sequences, pattern mining can be helpful to obtain models. The main idea is that patterns can be used as features to build a predictive model. For instance, *pattern-based classification* is a process of learning a classification model where patterns are used as features. According to recent studies, classification models making use of pattern-based features may be more accurate or simpler to understand than the original feature set (Cheng, Yan, Han, & Hsu, 2007). In structured domains, pattern mining may work as a *propositionalisation* approach that enables using classical propositional data mining/machine learning methods by decoupling the data representation from the learning task.

This paper describes a method for pattern-based classification based on a novel algorithm for pattern mining. The proposed algorithm discovers patterns represented as *SPARQL*(Prud'hommeaux & Seaborne, 2008) queries over a subset of *RDF Schema (RDFS)*(Brickley & Guha, 2004) suitable to represent lightweight ontologies. The algorithm takes the semantics of RDFS vocabulary into account, which enables it to exploit knowledge encoded in ontologies. Through a propositionali-