Using Prior Knowledge in Data Mining

Francesca A. Lisi
Università degli Studi di Bari, Italy

INTRODUCTION

One of the most important and challenging problems in current Data Mining research is the definition of the *prior knowledge* that can be originated from the process or the domain. This contextual information may help select the appropriate information, features or techniques, decrease the space of hypotheses, represent the output in a most comprehensible way and improve the process. Ontological foundation is a precondition for efficient automated usage of such information (Chandrasekaran *et al.*, 1999). An *ontology* is a formal explicit specification of a shared conceptualization for a domain of interest (Gruber, 1993). Among other things, this definition emphasizes the fact that an ontology has to be specified in a language that comes with a *formal semantics*. Due to this formalization ontologies provide the machine interpretable meaning of concepts and relations that is expected when using a semantic-based approach (Staab & Studer, 2004). In its most prevalent use in Artificial Intelligence (AI), an ontology refers to an engineering artifact (more precisely, produced according to the principles of *Ontological Engineering* (Gómez-Pérez *et al.*, 2004)), constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words. This set of assumptions has usually the form of a *First-Order Logic* (FOL) theory, where vocabulary words appear as unary or binary predicate names, respectively called concepts and relations. In the simplest case, an ontology describes a hierarchy of concepts related by subsumption relationships; in more sophisticated cases, suitable axioms are added in order to express other relationships between concepts and to constrain their intended interpretation.

Ontologies can play several roles in Data Mining (Nigro *et al.*, 2007). In this chapter we investigate the use of ontologies as prior knowledge in Data Mining. As an illustrative case throughout the chapter, we choose the task of Frequent Pattern Discovery, it being the most representative product of the cross-fertilization among Databases, Machine Learning and Statistics that has given rise to Data Mining. Indeed it is central to an entire class of descriptive tasks in Data Mining among which Association Rule Mining (Agrawal *et al*., 1993; Agrawal & Srikant, 1994) is the most popular. A pattern is considered as an intensional description (expressed in a given language $\mathcal{L}$) of a subset of a data set **r**. The support of a pattern is the relative frequency of the pattern within **r** and is computed with the evaluation function *supp*. The task of Frequent Pattern Discovery aims at the extraction of all frequent patterns, i.e. all patterns whose support exceeds a user-defined threshold of minimum support. The blueprint of most algorithms for Frequent Pattern Discovery is the *levelwise search* (Mannila & Toivonen, 1997). It is based on the following assumption: If a generality order $\geq$ for the language $\mathcal{L}$ of patterns can be found such that $\geq$ is monotonic w.r.t. *supp*, then the resulting space $(\mathcal{L}, \geq)$ can be searched breadth-first by starting from the most general pattern in $\mathcal{L}$ and alternating candidate generation and candidate evaluation phases.

BACKGROUND

The use of prior knowledge is already certified in Data Mining. Proposals for taking concept hierarchies into account during the discovery process are relevant to our survey because they can be considered a less expressive predecessor of ontologies, e.g. concept hierarchies are exploited to mine multiple-level association rules (Han & Fu, 1995; Han & Fu, 1999) or generalized association rules (Srikant & Agrawal, 1995). Both extend the levelwise search method so that patterns can refer to multiple levels of description granularity. They differ in the strategy used in visiting the concept hierarchy: the former visits the hierarchy top-down, the latter bottom-up.

The use of prior knowledge is also the distinguishing feature of Inductive Logic Programming (ILP) which was born at the intersection of Machine Learning (more precisely, Inductive Learning) and Logic Programming (Nienhuys-Cheng & de Wolf, 1997). Due to the com-

mon roots between Logic Programming and relational databases (Ceri et al., 1990), ILP has been more recently proposed as a logic-based approach to *Relational Data Mining* (Džeroski, 1996; Džeroski & Lavrač, 2001; Džeroski, 2002). Relational Data Mining is intended to overcome some limits of traditional Data Mining, e.g., in Association Rule Mining, by representing patterns and rules either as Datalog conjunctive queries (Dehaspe & De Raedt, 1997; Dehaspe & Toivonen, 1999) or as tree data structures (Nijssen & Kok, 2001; Nijssen & Kok, 2003). Note that none of these proposals for Association Rule Mining can exploit the semantic information conveyed by concept hierarchies because both adopt a syntactic generality relation for patterns and rules. More generally, prior knowledge in ILP is often not organized around a well-formed conceptual model such as ontologies.

MAIN FOCUS

In this section we consider the task of mining multiple-level association rules extended to the more complex case of having an ontology as prior knowledge and tackled with an ILP approach (Lisi & Malerba, 2004). We focus on the phase of Frequent Pattern Discovery.

The data set **r** must encompass both a database and an ontology, loosely or tightly integrated, so that the semantics can flow from the ontology to the database. To represent one such data set, a logical language that treats relational and structural knowledge in a unified way is necessary. Among the logical languages proposed by Ontological Engineering, *Description Logics* (DLs) are the most widely used (Baader *et al.*, 2007). The relationship between DLs and databases is rather strong. Several investigations have been carried out on the usage of DLs to formalize semantic data models. In these proposals concept descriptions are used to present the schema of a database. Unfortunately, DLs offer a weaker than usual query language. This makes also pure DLs inadequate as a Knowledge Representation (KR) framework in Data Mining problems that exploit ontological prior knowledge. Hybrid languages that integrate DLs and Datalog appear more promising. In (Lisi & Malerba, 2004), the data set **r** is a knowledge base represented according to the KR framework of $\mathcal{AL}$-log (Donini *et al.*, 1998), thus composed of a relational database in Datalog (Ceri *et al.*, 1990) and an ontology in the DL $\mathcal{ALC}$ (Schmidt-Schauss & Smolka, 1991).

The language $\mathcal{L}$ of patterns must be able to capture the semantics expressed in the background ontology. In (Lisi & Malerba, 2004), $\mathcal{L}$ is a language of unary conjunctive queries in $\mathcal{AL}$-log where the distinguished variable is constrained by the reference concept and the other variables are constrained by task-relevant concepts. All these concepts are taken from the underlying ontology, thus they convey semantics. Furthermore, the language is multi-grained in the sense that patterns that can be generated describe data at multiple levels of granularity. These levels refer to levels in the background ontology.

The generality order $\geq$ for the language $\mathcal{L}$ of patterns must be based on a semantic generality relation, i.e. a relation that checks whether a pattern is more general than another with respect to the prior knowledge. Up to now, most algorithms have focussed on a syntactical approach. However, the use of background knowledge would greatly improve the quality of the results. First, patterns and rules which are not equivalent from a syntactical point of view, may be semantically equivalent. Taking into account the semantical relationships between patterns improves the comprehensibility while decreasing the size of the discovered set of patterns. Second, while the use of prior knowledge increases the expressivity and therefore comes with a cost, it also allows to better exploit the benefits of some optimizations. In (Lisi & Malerba, 2004), the space of patterns is structured according to the semantic generality relation of $\mathcal{B}$-subsumption (Lisi & Malerba, 2003a) and searched by means of a downward refinement operator (Lisi & Malerba, 2003b). It has been proved that $\mathcal{B}$-subsumption fulfills the monotonicity requirement of the levelwise search (Lisi & Malerba, 2004). Note that the support of patterns is computed with respect to the background ontology.

This ILP approach to Frequent Pattern Discovery within the KR framework of $\mathcal{AL}$-log has been very recently extended to Cluster Analysis (Lisi & Esposito, 2007).

FUTURE TRENDS

Using ontologies as prior knowledge in Data Mining will become central to any application area where ontologies are playing a key role and Data Mining can be of help to users, notably the *Semantic Web* (Berners-Lee *et al.*, 2001). In particular, the main focus of