Knowledge-Based Systems 23 (2010) 743–756

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Concept discovery on relational databases: New techniques for search space pruning and rule quality improvement

Y. Kavurucu, P. Senkul *, I.H. Toroslu

Middle East Technical University, Department of Computer Engineering, 06531 Ankara, Turkey

ARTICLE INFO

Article history:
Received 4 February 2010
Received in revised form 20 April 2010
Accepted 21 April 2010
Available online 28 April 2010

Keywords:
ILP
Data mining
MRDM
Concept discovery
Transitive rules
Support
Confidence

ABSTRACT

Multi-relational data mining has become popular due to the limitations of propositional problem definition in structured domains and the tendency of storing data in relational databases. Several relational knowledge discovery systems have been developed employing various search strategies, heuristics, language pattern limitations and hypothesis evaluation criteria, in order to cope with intractably large search space and to be able to generate high-quality patterns. In this work, we introduce an ILP-based concept discovery framework named Concept Rule Induction System (CRIS) which includes new approaches for search space pruning and new features, such as defining aggregate predicates and handling numeric attributes, for rule quality improvement. In CRIS, all target instances are considered together, which leads to construction of more descriptive rules for the concept. This property also makes it possible to use aggregate predicates more accurately in concept rule construction. Moreover, it facilitates construction of transitive rules. A set of experiments is conducted in order to evaluate the performance of proposed method in terms of accuracy and coverage.

1. Introduction

The amount of data collected on relational databases has been increasing due to increase in the use of complex data for real life applications. This motivated the development of multi-relational learning algorithms that can be applied to directly multi-relational data on the databases [19,17]. For such learning systems, generally the first-order predicate logic is employed as the representation language. The learning systems, which induce logical patterns valid for given background knowledge, have been investigated under a research area, called Inductive Logic Programming (ILP) [46]. In general, using logic in data mining is a common technique in the literature [52,47,57,12,63,67,39,23,49,50,5,40,35,10,36,64,16,55,18,51,65,66].

Concept is a set of patterns to be discovered by using the hidden relationships in the database. Concept discovery in relational databases is a predictive learning task. In predictive learning, there is a specific target concept to be learned in the light of the past experiences [45]. The problem setting of the predictive learning task introduced by Muggleton in [45] can be stated as follows: given target class/concept C (target relation), a set E of positive and negative examples of the class/concept C, a finite set of background facts/clauses B (background relations), concept description language L (language bias); find a finite set of clauses H, expressed in concept description language L, such that H together with the background knowledge B entail all positive instances $E(+)$ and none of the negative instances $E(-)$. In other words, H is complete and consistent with respect to B and E, respectively.

Association rule mining in relational databases is a descriptive learning task. In descriptive learning, the task is to identify frequent patterns, associations or correlations among sets of items or objects in databases [45]. Relational association rules are expressed as query extensions in the first-order logic [11,13]. In the proposed work, there is a specific target concept and association rule mining techniques are employed to induce association rules which have only the target concept as the only head relation.

In this paper, we present Concept Rule Induction System (CRIS), which is a concept learning ILP system that employs relational association rule mining concepts and techniques to find frequent and strong concept definitions according to given target relation and background knowledge [31]. CRIS utilizes absorption operator of inverse resolution for generalization of concept instances in the presence of background knowledge and refines these general patterns into frequent and strong concept definitions with an APRIORI-based specialization operator based on confidence.

1.1. Contributions

Major contributions and the main features of this work can be listed as follows:

* Corresponding author. Tel.: +90 312 2105518.
E-mail addresses: yusuf.kavurucu@ceng.metu.edu.tr (Y. Kavurucu), senkul@ceng.metu.edu.tr (P. Senkul), toroslu@ceng.metu.edu.tr (I.H. Toroslu).
URL: http://www.ceng.metu.edu.tr/karagoz/ (P. Senkul).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.04.011

1. The selection order of the target instance (the order in the target relation) may change the resulting hypothesis set. In each coverage set, the induced rules depend on the selected target instance and the covered target instances in each step do not have any effect on the induced rules in the following coverage steps. To overcome this problem, first, all possible values for each argument of a relation are determined by executing simple SQL statements in the database. Instead of selecting a target instance, those values for each argument are used in the generalization step of CRIS. By this way, the generated rules do not depend on the instance selection order and induced rule quality is improved.
2. This technique facilitates the generation of transitive rules, as well. When the target concept has common attribute types with only some of the background predicates, the rest of the predicates (which are called *unrelated relations*) can never take part in hypothesis. This prevents the generation of transitive rules through such predicates. In CRIS, since all target instances are considered together, there is no distinction for related and unrelated relations and hence transitive rules can be induced.
3. Better rules (higher accuracy and coverage) can be discovered by using aggregate predicates in the background knowledge. To do this, aggregate predicates are defined in the first-order logic and used in CRIS. In addition, numerical attributes are handled in a more accurate way. The rules having comparison operators on numerical attributes are defined and used in the main algorithm.
4. CRIS utilizes primary key-foreign key relationship (if exists) between the head and body relations in the search space as a pruning strategy. If a primary-foreign key relationship exists between the head and the body predicates, the foreign key argument of the body relation can only have the same variable as the primary key argument of the head predicate in the generalization step.
5. The main difficulty in relational ILP systems is searching in intractably large hypothesis spaces. In order to reduce the search space, a confidence-based pruning mechanism is used. In addition to this, many multi-relational rule induction systems require the user to determine the input–output modes of predicate arguments. Instead of this, we use the information about relationships between entities in the database if given.
6. Muggleton shows that [48], the expected error of an hypothesis according to positive versus all (positive and negative) examples do not have much difference if the number of examples is large enough. Most ILP-based concept learning systems input background facts in Prolog language; this restricts the usage of ILP engines in real-world applications due to the time-consuming transformation phase of problem specification from tabular to logical format. The proposed system directly works on relational databases, which contain only positive information, without any requirement of negative instances. Moreover, the definition of confidence is modified to apply Closed World Assumption (CWA) [53] in relational databases. We introduce type relations to the body of the rules in order to express CWA.

In [31], the contribution presented in the first item of the above list was introduced without performance evaluation. In [33], the basics of aggregate predicate usage are presented. In this work, the features of CRIS are elaborated in more detail with performance evaluation results on several data sets. In [29,28,30,32], features of another concept discovery system developed by our research group, namely C^2D, are presented. Although, CRIS and C^2D have common properties such as use of only positive instances, the concept discovery algorithm of CRIS has different properties and advantages which are presented and discussed and evaluated in this work.

This paper is organized as follows: Section 2 gives preliminary information about concept discovery in general and the concepts employed in CRIS. Section 3 presents the related work. Section 4 describes the proposed method. Section 5 presents the experiments to discuss the performance of CRIS. Finally, Section 6 includes concluding remarks.

2. Preliminaries

In this section, basic terminology in concept discovery and basics for concept representation and discovery are introduced.

2.1. Basics

A concept is a set of patterns which are embedded in the features of the instances of a given target relation and in the relationships of this relation with other relations. In this work, a concept is defined though concept rules.

Definition 1. [Concept rule] A concept rule (or shortly rule) is an association rule (range-restricted query extension). It is represented as "$h \leftarrow b$", where h is the head of the rule and b denotes the body of the rule.

Definition 2. [Target relation] A target relation is a predicate that corresponds to the concept to be discovered. The instances of the target relations have to be correctly covered by the discovered pattern. If the discovered pattern is in the form of rules (as in this work), target relation appears in the head of the rule. In recursive rules, it may take part in the body part, as well.

Definition 3. [Background relation] A background relation is a predicate that is different than the target relation and involves in the concept discovery. When discovered pattern is in the form of rules, a background relation may appear in the body part of the rule.

In Table 1, the relation given in the first column, *ancestor*, is the target relation. The content of the first column constitutes the target instances. For this example, one of the concepts rules defining the concept is "ancestor(A, B) $\leftarrow$ parent(A, B)".

We use the first-order logic as the language to represent data and patterns. The concept rule structure is based on query extension. However, to emphasize the difference from classical clause and query, we firstly present definitions for these terms.

Definition 4. [Clause] A clause is a universally quantified disjunction $\forall(l_1 \vee l_2 \vee \ldots \vee l_n)$. When it is clear from the context that clauses are meant, the quantifier $\forall$ is dropped. A clause $h_1 \vee h_2 \vee \ldots \vee h_p \vee b_1 \vee b_2 \vee \ldots \vee b_r$, where the h_i are positive literals and the b_j are negative literals, can also be written as $h_1 \vee h_2 \vee \ldots \vee h_p \leftarrow b_1 \wedge b_2 \wedge \ldots \wedge b_r$, where $h_1 \vee h_2 \vee \ldots \vee h_p$

Table 1
The database of the ancestor example with type declarations.

Concept instances	Background facts
a(*kubra*, *ali*).	*p*(*kubra*, *ali*).
a(*ali*, *yusuf*).	*p*(*ali*, *yusuf*).
a(*yusuf*, *esra*).	*p*(*yusuf*, *esra*).
a(*yusuf*, *aysegul*).	*p*(*yusuf*, *aysegul*).
a(*kubra*, *yusuf*).	
a(*kubra*, *esra*).	
a(*kubra*, *aysegul*).	
a(*ali*, *esra*).	
a(*ali*, *aysegul*).	