Association Rule Mining of Relational Data

Anne Denton
North Dakota State University, USA

Christopher Besemann
North Dakota State University, USA

INTRODUCTION

Most data of practical relevance are structured in more complex ways than is assumed in traditional data mining algorithms, which are based on a single table. The concept of relations allows for discussing many data structures such as trees and graphs. Relational data have much generality and are of significant importance, as demonstrated by the ubiquity of relational database management systems. It is, therefore, not surprising that popular data mining techniques, such as association rule mining, have been generalized to relational data. An important aspect of the generalization process is the identification of problems that are new to the generalized setting.

BACKGROUND

Several areas of databases and data mining contribute to advances in association rule mining of relational data:

- **Relational Data Model:** underlies most commercial database technology and also provides a strong mathematical framework for the manipulation of complex data. Relational algebra provides a natural starting point for generalizations of data mining techniques to complex data types.
- **Inductive Logic Programming, ILP (Džeroski & Lavrač, 2001):** a form of logic programming, in which individual instances are generalized to make hypotheses about unseen data. Background knowledge is incorporated directly.
- **Association Rule Mining, ARM (Agrawal, Imielinski, & Swami, 1993):** identifies associations and correlations in large databases. Association rules are defined based on items, such as objects in a shopping cart. Efficient algorithms are designed by limiting output to sets of items that occur more frequently than a given threshold.
- **Graph Theory:** addresses networks that consist of nodes, which are connected by edges. Traditional graph theoretic problems typically assume no more than one property per node or edge. Data associated with nodes and edges can be modeled within the relational algebra.

Association rule mining of relational data incorporates important aspects of these areas to form an innovative data mining technique of important practical relevance.

MAIN THRUST

The general concept of association rule mining of relational data will be explored, as well as the special case of mining a relationship that corresponds to a graph.

General Concept

Two main challenges have to be addressed when applying association rule mining to relational data. Combined mining of multiple tables leads to a search space that is typically large even for moderately sized tables. Performance is, thereby, commonly an important issue in relational data mining algorithms. A less obvious problem lies in the skewing of results (Jensen & Neville, 2002). The relational join operation combines each record from one table with each occurrence of the corresponding record in a second table. That means that the information in one record is represented multiple times in the joined table. Data mining algorithms that operate either explicitly or implicitly on joined tables, thereby, use the same information multiple times. Note that this problem also applies to algorithms in which tables are joined on-the-fly by identifying corresponding records as they are needed. Further specific issues may have to be addressed when reflexive relationships are present. These issues will be discussed in the section on relations that represent a graph.

A variety of techniques have been developed for data mining of relational data (Džeroski & Lavrač, 2001). A typical approach is called inductive logic programming, ILP. In this approach relational structure is represented in the form of Prolog queries, leaving maximum flexibility to the user. While the notation of ILP differs from the

relational notation it can be noted that all relational operators can also be represented in ILP. The approach does thereby not limit the types of problems that can be addressed. It should, however, also be noted that while relational database management system are developed with performance in mind there may be a trade-off between the generality of Prolog-based environments and their limitations in speed.

Application of ARM within the ILP setting corresponds to a search for frequent Prolog queries as a generalization of traditional association rules (Dehaspe & De Raedt, 1997). Examples of association rule mining of relational data using ILP (Dehaspe & Toivonen, 2001) could be shopping behavior of customers where relationships between customers are included in the reasoning. While ILP does not use a relational joining step as such, it does also associate individual objects with multiple occurrences of corresponding objects. Problems with skewing are, thereby, also encountered in this approach.

An alternative to the ILP approach is to apply the standard definition of association rule mining to relations that are joined using the relational join operation. While such an approach is less general it is often more efficient since the join operation is highly optimized in standard database systems. It is important to note that a join operation typically changes the support of an item set, and any support calculation should therefore be based on the relation that uses the smallest number of join operations (Cristofor & Simovici, 2001). Equivalent changes in item set weighting occur in ILP.

Interestingness of rules is an important issue in any type of association rule mining. In traditional association rule mining the problem of rule interest has been addressed in a variety of work on redundant rules, including closed set generation (Zaki, 2000). Additional rule metrics such as lift and conviction have been defined (Brin, Motwani, Ullman, & Tsur, 1997). In relational association rule mining the problem has been approached by the definition of a deviation measure (Dehaspe & Toivonen, 2001). In general it can be noted that relational data mining poses many additional problems related to skewing of data compared with traditional mining on a single table (Jensen & Neville, 2002).

Relations that Represent a Graph

One type of relational data set has traditionally received particular attention, albeit under a different name. A relation representing a relationship between entity instances of the same type, also called a reflexive relationship, can be viewed as the definition of a graph. Graphs have been used to represent social networks, biological networks, communication networks, and citation graphs, just to name a few.

A typical example of an association rule mining problem is mining of annotation data of proteins in the presence of a protein-protein interaction graph (Oyama, Kitano, Satou, & Ito, 2002). Associations are extracted that relate functions and localizations of one protein with those of interacting proteins. Oyama et al. use association rule mining, as applied to joined relations, for this work. Another example could be association rule mining of attributes associated with scientific publications on the graph of their mutual citations.

A problem of the straight-forward approach of mining joined tables directly becomes obvious upon further study of the rules: In most cases the output is dominated by rules that involve the same item as it occurs in different entity instances that participate in a relationship. In the example of protein annotations within the protein interaction graph a protein in the "nucleus" is found to frequently interact with another protein that is also located in the "nucleus". Similarities among relational neighbors have been observed more generally for relational databases (Macskassy & Provost, 2003). It can be shown that filtering of output is not a consistent solution to this problem, and items that are repeated for multiple nodes should be eliminated in a preprocessing step (Besemann & Denton, 2004). This is an example of a problem that does not occur in association rule mining of a single table and requires special attention when moving to multiple relations. The example also highlights the need to discuss differences between sets of items of related objects are (Besemann, Denton, Yekkirala, Hutchison, & Anderson, 2004).

Related Research Areas

A related research area is graph-based ARM (Inokuchi, Washio, & Motoda, 2000; Yan & Han, 2002). Graph-based ARM does not typically consider more than one label on each node or edge. The goal of graph-based ARM is to find frequent substructures based on that one label, focusing on algorithms that scale to large subgraphs. In relational ARM multiple item are associated with each node and the main problem is to achieve scaling with respect to the number of items per node. Scaling to large subgraphs is usually irrelevant due to the "small world" property of many types of graphs. For most networks of practical interest any node can be reached from almost any other by means of no more than some small number of edges (Barabasi & Bonabeau, 2003). Association rules that involve longer distances are therefore unlikely to produce meaningful results.

There are other areas of research on ARM in which related transactions are mined in some combined fashion. Sequential pattern or episode mining (Agrawal & Srikant, 1995; Yan, Han, & Afshar, 2003) and inter-transaction mining (Tung, Lu, Han, & Feng, 1999) are two main