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** (Dzeroski & 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 (Dzeroski & 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
Related Content

A Survey of Feature Selection Techniques
[www.igi-global.com/chapter/survey-feature-selection-techniques/11077?camid=4v1a](www.igi-global.com/chapter/survey-feature-selection-techniques/11077?camid=4v1a)

Vertical Data Mining
William Perrizo, Qiang Ding, Qin Ding and Taufik Abidin (2005). Encyclopedia of Data Warehousing and Mining (pp. 1181-1184).
[www.igi-global.com/chapter/vertical-data-mining/10776?camid=4v1a](www.igi-global.com/chapter/vertical-data-mining/10776?camid=4v1a)

Agent-Mediated Knowledge Acquisition for User Profiling
[www.igi-global.com/chapter/agent-mediated-knowledge-acquisition-user/7711?camid=4v1a](www.igi-global.com/chapter/agent-mediated-knowledge-acquisition-user/7711?camid=4v1a)

Hierarchical Document Clustering
[www.igi-global.com/chapter/hierarchical-document-clustering/10938?camid=4v1a](www.igi-global.com/chapter/hierarchical-document-clustering/10938?camid=4v1a)