Mine Rule

Rosa Meo
Universitá degli Studi di Torino, Italy

Giuseppe Psaila
Universitá degli Studi di Bergamo, Italy

INTRODUCTION

Mining of association rules is one of the most adopted techniques for data mining in the most widespread application domains. A great deal of work has been carried out in the last years on the development of efficient algorithms for association rules extraction. Indeed, this problem is a computationally difficult task, known as NP-hard (Calders, 2004), which has been augmented by the fact that normally association rules are being extracted from very large databases. Moreover, in order to increase the relevance and interestingness of obtained results and to reduce the volume of the overall result, constraints on association rules are introduced and must be evaluated (Ng et al.,1998; Srikant et al., 1997). However, in this contribution, we do not focus on the problem of developing efficient algorithms but on the semantic problem behind the extraction of association rules (see Tsur et al. [1998] for an interesting generalization of this problem).

We want to put in evidence the semantic dimensions that characterize the extraction of association rules; that is, we describe in a more general way the classes of problems that association rules solve. In order to accomplish this, we adopt a general-purpose query language designed for the extraction of association rules from relational databases. The operator of this language, MINE RULE, allows the expression of constraints, constituted by standard SQL predicates that make it suitable to be employed with success in many diverse application problems. For a comparison between this query language and other state-of-the-art languages for data mining, see Imielinski, et al. (1996); Han, et al. (1996); Netz, et al. (2001); Botta, et al. (2004).

In Imielinski, et al. (1996), a new approach to data mining is proposed, which is constituted by a new generation of databases called Inductive Databases (IDBs). With an IDB, the user/analyst can use advanced query languages for data mining in order to interact with the knowledge discovery (KDD) system, extract data mining descriptive and predictive patterns from the database, and store them in the database. Boulicaut, et al. (1998) and Baralis, et al. (1999) discuss the usage of MINE RULE in this context.

We want to show that, thanks to a highly expressive query language, it is possible to exploit all the semantic possibilities of association rules and to solve very different problems with a unique language, whose statements are instantiated along the different semantic dimensions of the same application domain. We discuss examples of statements solving problems in different application domains that nowadays are of a great importance. The first application is the analysis of a retail data, whose aim is market basket analysis (Agrawal et al., 1993) and the discovery of user profiles for customer relationship management (CRM). The second application is the analysis of data registered in a Web server on the accesses to Web sites by users. Cooley, et al. (2000) present a study on the same application domain. The last domain is the analysis of genomic databases containing data on micro-array experiments (Fayyad, 2003). We show many practical examples of MINE RULE statements and discuss the application problems that can be solved by analyzing the association rules that result from those statements.

BACKGROUND

An association rule has the form $B \Rightarrow H$, where B and H are sets of items, respectively called *body* (the antecedent) and *head* (the consequent). An association rule (also denoted for short with rule) intuitively means that items in B and H often are associated within the observed data. Two numerical parameters denote the validity of the rule: *support* is the fraction of source data for which the rule holds; *confidence* is the conditional probability that H holds, provided that B holds. Two minimum thresholds for support and confidence are specified before rules are extracted, so that only significant rules are extracted.

This very general definition, however, is incomplete and very ambiguous. For example, what is the meaning of "fraction of source data for which the rule holds"? Or what are the items associated by a rule? If we do not answer these basic questions, an association rule does not have a precise meaning. Consider, for instance, the original problem for which association rules were initially proposed in Agrawal, et al. (1993)—the market baskets analysis. If we have a database collecting single purchase transactions (i.e., transactions performed by customers in a retail store), we might wish to extract association rules that associate items sold within the same transactions. Intuitively, we are defining the semantics of our problem—items are associated by a rule if they appear together in the same transaction. Support denotes the fraction of the total transactions that contain all the items in the rule (both *B* and *H*), while confidence denotes the conditional probability that, found *B* in a transaction, also *H* is found in the same transaction. Thus a rule

{pants, shirt} ⇒ {socks, shoes}
support=0.02 confidence=0.23

means that the items pants, shirt, socks, and shoes appear together in 2% of the transactions, while having found items pants and shirt in a transaction, the probability that the same transaction also contains socks and shoes is 23%.

Semantic Dimensions

MINE RULE puts in evidence the semantic dimensions that characterize the extraction of association rules from within relational databases and force users (typically analysts) to understand these semantic dimensions. Indeed, extracted association rules describe the most recurrent values of certain attributes that occur in the data (in the previous example, the names of the purchased product). This is the first semantic dimension that characterizes the problem. These recurrent values are observed within sets of data grouped by some common features (i.e., the transaction identifier in the previous example but, in general, the date, the customer identifier, etc.). This constitutes the second semantic dimension of the association rule problem. Therefore, extracted association rules describe the observed values of the first dimension, which are recurrent in entities identified by the second dimension.

When values belonging to the first dimension are associated, it is possible that not every association is suitable, but only a subset of them should be selected, based on a coupling condition on attributes of the analyzed data (e.g., a temporal sequence between events described in B and H). This is the third semantic dimension of the problem; the coupling condition is called *mining condition*.

It is clear that MINE RULE is not tied to any particular application domain, since the semantic dimensions allow the discovery of significant and unexpected information in very different application domains.

The main features and clauses of MINE RULE are as follows (see Meo, et al. [1998] for a detailed description):

- **Selection of the relevant set of data for a data mining process:** This feature is specified by the FROM clause.
- **Selection of the grouping features w.r.t., which data are observed:** These features are expressed by the GROUP BY clause.
- **Definition of the structure of rules and cardinality constraints on body and head, specified in the SELECT clause:** Elements in rules can be single values or tuples.
- **Definition of coupling constraints:** These are constraints applied at the rule level (mining condition instantiated by a WHERE clause associated to SELECT) for coupling values.
- **Definition of rule evaluation measures and minimum thresholds:** These are support and confidence (even if, theoretically, other statistical measures also would be possible). Support of a rule is computed on the total number of groups in which it occurs and satisfies the given constraints. Confidence is the ratio between the rule support and the support of the body satisfying the given constraints. Thresholds are specified by clause EXTRACTING RULES WITH.

MAIN THRUST

In this section, we introduce MINE RULE in the context of the three application domains. We describe many examples of queries that can be conceived as a sort of template, because they are instantiated along the relevant dimensions of an application domain and solve some frequent, similar, and critical situations for users of different applications.

First Application: Retail Data Analysis

We consider a typical data warehouse gathering information on customers' purchases in a retail store:

FactTable (TransId, CustId, TimeId, ItemId, Num,
Discount)
Customer (<u>CustId</u>, Profession, Age, Sex)

Rows in FactTable describe sales. The dimensions of data are the customer (*CustId*), the time (*TimeId*), and the purchased item (*ItemId*); each sale is characterized by the