D

Distance-Based Methods for Association Rule Mining

Vladimír Bartík
Brno University of Technology, Czech Republic

Jaroslav Zendulka
Brno University of Technology, Czech Republic

INTRODUCTION

Association rules are one of the most frequently used types of knowledge discovered from databases. The problem of discovering association rules was first introduced in (Agrawal, Imielinski & Swami, 1993). Here, association rules are discovered from transactional databases –a set of transactions where a transaction is a set of items. An association rule is an expression of a form $A \Rightarrow B$ where A and B are sets of items. A typical application is market basket analysis. Here, the transaction is the content of a basket and items are products. For example, if a rule *milk* $\wedge$ *juice* $\Rightarrow$ *coffee* is discovered, it is interpreted as: "If the customer buys milk and juice, s/he is likely to buy coffee too." These rules are called *single-dimensional Boolean association rules* (Han & Kamber, 2001). The potential usefulness of the rule is expressed by means of two metrics – support and confidence.

A lot of algorithms have been developed for mining association rules in transactional databases. The best known is the Apriori algorithm (Agrawal & Srikant, 1994), which has many modifications, e.g. (Kotásek & Zendulka, 2000). These algorithms usually consist of two phases: discovery of frequent itemsets and generation of association rules from them. A frequent itemset is a set of items having support greater than a threshold called minimum support. Association rule generation is controlled by another threshold referred to as minimum confidence.

Association rules discovered can have a more general form and their mining is more complex than mining rules from transactional databases. In relational databases, association rules are ordinarily discovered from data of one table (it can be the result of joining several other tables). The table can have many columns (attributes) defined on domains of different types. It is useful to distinguish two types of attributes.

A *categorical attribute* (also called nominal) has a finite number of possible values with no ordering among the values (e.g. a country of a customer).

A *quantitative attribute* is a numeric attribute, domain of which is infinite or very large. In addition, it has an implicit ordering among values (e.g. age and salary of a customer).

An association rule (*Age* = [20…30]) $\wedge$ (*Country* = "Czech Rep.") $\Rightarrow$ (Salary = [1000$...2000$]) says that if the customer is between 20 and 30 and is from the Czech Republic, s/he is likely to earn between 1000$ and 2000$ per month. Such rules with two or more predicates (items) containing different attributes are also called *multidimensional association rules*. If some attributes of rules are quantitative, the rules are called quantitative association rules (Han & Kamber, 2001).

If a table contains only categorical attributes, it is possible to use modified algorithms for mining association rules in transactional databases. The crucial problem is to process quantitative attributes because their domains are very large and these algorithms cannot be used. Quantitative attributes must be *discretized*.

This article deals with mining multidimensional association rules from relational databases, with main focus on distance-based methods. One of them is a novel method developed by the authors.

BACKGROUND

There are three basic approaches regarding the treatment of quantitative attributes (Han & Kamber, 2001). First one uses a predefined set of ranges (or, in general, a concept hierarchy) to replace the original numeric values of a quantitative attribute by ranges that represent intervals of values. This discretization occurs prior to applying a mining algorithm. It is *static* and

predetermined. In the second approach, quantitative attributes are initially discretized statically. The resulting ranges are then combined during the mining algorithm. Therefore, the discretization process is *dynamic*. The third approach tries to define ranges based on semantic meaning of the data. This discretization is dynamic too. It considers distance between data points. Discretization and mining methods based on this approach are referred to as *distance-based methods*.

There are two basic methods of *static discretization*: equi-depth and equi-width discretization. Equi-depth discretization lies in creating intervals that contain the same number of values. Equi-width discretization creates ranges of the same size. These methods are very simple, but the result of discretization can be unsuitable because some important associations may be lost. This is caused by the fact that two very near values can be in two different ranges. Sometimes one cluster of values, which might be represented by one predicate in an association rule, is divided into two ranges.

The following two methods are representatives of the second approach mentioned above.

A method for mining quantitative association rules was proposed in (Srikant & Agrawal, 1996). The number of intervals, which will be created, is determined by means of a measure referred to as K-partial completeness, which guarantees the acceptable loss of information. This measure is represented by a number K, which is higher than 1. This value is used to determine number of intervals N:

$$N = \frac{2}{minsup \cdot (K-1)} \qquad (1)$$

where *minsup* is a minimum support threshold.

After initial equi-depth discretization, neighboring intervals are joined together to form intervals having sufficient support. These intervals are then used to discover frequent itemsets.

Zhang extended the method to *fuzzy quantitative association rules* in (Zhang, 1999). Here, association rules can contain fuzzy terms too. In the phase of discretization, creating of intervals is combined with creating of fuzzy terms over quantitative attributes.

A method for mining *optimized association rules* proposed in (Fukuda, Morimoto, Morishita & Tokuyama, 1996) uses no additional measure for the interestingness of an association rule. Because the minimum support and confidence thresholds are antipodal, either rules referred to as optimized support rules or optimized confidence rules are obtained with this method. The goal of the optimized support rules is to find rules with support as high as possible that satisfy minimum confidence. Similarly, the optimized confidence rules maximize confidence of the rule while satisfying the minimum support requirement. The method uses the initial equi-depth discretization of quantitative attributes. Then the method continues with finding optimal interval for a given form of an association rule. The method discovers association rules referred to as constraint-based association rules where the form of rules to be mined must be defined, for example: $(Age \in [v_1, v_2]) \wedge (Country = X) \Rightarrow (Salary \in [v_3, v_4])$. Another method for mining optimized association rules is proposed in (Xiaoyong, Zhibin & Naohiro, 1999).

MAIN FOCUS

Distance-based methods try to respect semantics of data in such a way that the discretization of quantitative attributes reflects the distances between numeric values. The intervals of numeric values are constructed dynamically to contain clusters of values lying close to each other. The main objective of distance-based methods is to minimize loss of information caused by discretization.

The distance-based approach can either be applied as a discretization method or as part and parcel of a mining algorithm.

The first distance-based method was proposed in (Miller & Yang, 1997). Here, the clustering methods are used to create intervals. This algorithm works in two phases; in the first one, interesting clusters over quantitative attributes are found (e.g., with clustering algorithm Birch (Zhang, Ramakrishnan & Livny, 1996) and these clusters are used to generate frequent itemsets of clusters and association rules containing them.

The result of the process is a set of association rules of a form $C_{X1} \wedge ... \wedge C_{Xn} \Rightarrow C_{Y1} \wedge ... \wedge C_{Yn}$, where C_X and C_Y are clusters over quantitative attributes X and Y. Xi and Yj are pairwise disjoint sets of quantitative attributes. Except minimum support and confidence, an association rule must meet the condition of an *association degree*. To determine the value of the association degree of a rule $C_{Xi} \Rightarrow C_{Yi}$, we need to know if values of the attribute X in the cluster C_{Yi} are inside the cluster C_{Xi}