Association Rule Mining

Vasudha Bhatnagar
University of Delhi, India

Sarabjeet Kochhar
University of Delhi, India

INTRODUCTION

Data mining is a field encompassing study of the tools and techniques to assist humans in intelligently analyzing (mining) mountains of data. Data mining has found successful applications in many fields including sales and marketing, financial crime identification, portfolio management, medical diagnosis, manufacturing process management and health care improvement etc..

Data mining techniques can be classified as either descriptive or predictive techniques. Descriptive techniques summarize / characterize general properties of data, while predictive techniques construct a model from the historical data and use it to predict some characteristics of the future data. Association rule mining, sequence analysis and clustering are key descriptive data mining techniques, while classification and regression are predictive techniques.

The objective of this article is to introduce the problem of association rule mining and describe some approaches to solve the problem.

BACKGROUND

Association rule mining, one of the fundamental techniques of data mining, aims to extract interesting correlations, frequent patterns or causal structures among sets of items in data.

An association rule is of the form X→ Y and indicates that the presence of items in the antecedent of rule (X) implies the presence of items in the consequent of rule (Y). For example, the rule *{PC, Color Printer}* → *{computer table}* implies that people who purchase a PC (personal computer) and a color printer also tend to purchase a computer table. These associations, however, are not based on the inherent characteristics of a domain (as in a functional dependency) but on the co-occurrence of data items in the dataset. Thus, association rule mining is a totally data driven technique.

Association rules have been successfully employed in numerous applications, some of which are listed below:

1. **Retail market analysis:** Discovery of association rules in retail data has been applied in departmental stores for floor planning, stock planning, focused marketing campaigns for product awareness, product promotion and customer retention.
2. **Web association analysis:** Association rules in web usage mining have been used to recommend related pages, discover web pages with common references, web pages with majority of same links (mirrors) and predictive caching. The knowledge is applied to improve web site design and speed up searches.
3. **Discovery of linked concepts:** Words or sentences that appear frequently together in documents are called linked concepts. Association rules can be used to discover linked concepts which further lead to the discovery of plagiarized text and the development of ontologies etc..

The problem of association rule mining (ARM) was introduced by Agrawal et al. (1993). Large databases of retail transactions called the market basket databases, which accumulate in departmental stores provided the motivation of ARM. The basket corresponds to a physical retail transaction in a departmental store and consists of the set of items a customer buys. These transactions are recorded in a database called the transaction database. The goal is to analyze the buying habits of customers by finding associations between the different items that customers place in their "shopping baskets". The discovered association rules can also be used by management to increase the effectiveness of

Figure 1. Boolean database and corresponding transaction database

TID	A	B	C	D	E
10	1	1	1	0	0
20	1	0	1	0	0
30	1	0	1	1	0
40	0	1	1	0	1
50	1	0	1	0	1

TID	Items
10	A, B,C
20	A, C
30	A, C, D
40	B, C, E
50	A, C, E

advertising, marketing, inventory management and reduce the associated costs.

The authors in (Agrawal et al., 1993) worked on a *boolean database* of transactions. Each record corresponds to a customer basket and contains transaction identifier (*TID*), transaction details and a *list of items* bought in the transaction. The list of items is represented by a boolean vector with a *one* denoting presence of corresponding item in the transaction and *zero* marking the absence. Figure 1 shows the *boolean database* of five transactions and the corresponding transaction database.

The problem of finding association rules is to find the columns with frequently co-occurring ones in the boolean database. However, most of the algorithms for ARM use the form of transaction database shown on the right. We give the mathematical formulation of the problem below.

MATHEMATICAL FORMULATION OF THE ARM PROBLEM

Let $I = \{i_1, i_2, ..., i_n\}$ denote a set of items and D designate a database of N transactions. A transaction $T \in D$ is a subset of I i.e. $T \subseteq I$ and is associated with a unique identifier *TID*.

An *itemset* is a collection of one or more items. X is an itemset if $X \subseteq I$. A transaction is said to contain an itemset X if $X \subseteq T$. A *k-itemset* is an itemset that contains k items.

An *association rule* is of the form $X \rightarrow Y$ *[Support, Confidence]* where $X \subset I$, $Y \subset I$, $X \cap Y = \varnothing$, and *Support* and *Confidence* are rule evaluation metrics.

Support of an itemset X is the fraction of transactions that contain X. It denotes the probability that a transaction contains X.

$$\text{Support}(X) = P(X) = \frac{\text{No. of transactions containing } X}{\text{Total number of transactions in } D}$$

Support of a rule $X \rightarrow Y$ in D is 's' if s% of transactions in D contain $X \cup Y$, and is computed as:

$$\text{Support}(X \rightarrow Y) = P(X \cup Y) = \frac{\text{No. of transactions containin } X \cup Y}{\text{Total number of transactions in } D}$$

Support indicates the extent of prevalence of a rule. A rule with low support value represents a rare event.

Confidence of a rule measures its strength and provides an indication of the reliability of prediction made by the rule. A rule $X \rightarrow Y$ has a confidence 'c' in D if c% of transactions in D that contain X also contain Y. It is computed as the conditional probability that Y occurs in a transaction, given X is present in the same transaction, i.e.

$$\text{Confidence}(X \rightarrow Y) = P(Y/X) = \frac{P(X \cup Y)}{P(X)}$$

Example 1: Consider the example database shown in Figure 2 (a). Here, I = {A, B, C, D, E}. Figures 2 (b) and 2 (c) show the computation of support and confidence for a rule.