Association Rule Mining

The past few years have seen a tremendous interest in area of data mining. Data mining is generally thought of as the process of finding hidden, non trivial and previously unknown information in large collection of data. Association rule mining is an important component of data mining. Association rules are an important class of methods of finding regularities/patterns in data. It is perhaps the most important model invented and extensively studied by databases and data mining community. Association mining has been used in many application domains. One of the best known is the business field where discovering of purchase patterns or association between products is very useful for decision making and effective marketing. However in last few years application areas have increased significantly. Some examples of recent applications are: finding patterns in biological databases, extraction of knowledge from software engineering metrics, web personalization, text mining etc. Association rule mining can also play an important role in discovering knowledge from agricultural databases, survey data from agricultural research, data about soil and cultivation, data containing information linking geographical conditions and crop production to name a few. Such knowledge can assist in making decisions regarding selection of crops to be grown in particular area based in geographical conditions, increasing production of crop by selecting proper resources, selecting proper environment for crop production etc. This paper starts with basics of association rules. Further some of the problems with standard association mining methods are discussed. Focusing on some of these problems, the paper discusses extensions applied to generalize the association mining methods for efficient data mining and some of the new applications of association rule mining based on these extensions are discussed.

Basic Concept of Association Rules

Basics objective of finding association rules is to find all co-occurrence relationship called associations. Since it was first introduced in 1993 by Agrawal et. al, it has attracted a great deal of attention. Many efficient algorithms, extensions and applications have been reported. The classic application of association rule mining is market basket data analysis, which aims to discover how items purchased by customers in a supermarket (or store) are associated. Association rules are of form \( X \rightarrow Y \), where \( X \) and \( Y \) are collection of items and intersection of \( X \) and \( Y \) is null. For example we may find that “95 percent of customers who bought bread\( (X) \) also bought milk\( (Y) \)” A rule may contain more than one item in antecedent and consequent of rule. Every rule must satisfy two users specified constrains: one is measure of statistical significance called support and other is measure of goodness called confidence.

The problem of mining association rules can be stated as follows: Let \( I = \{i_1, i_2, \ldots, i_m\} \) be a set of items. Let \( T = (t_1, t_2, \ldots, t_n) \) be a set of transactions (the database), where each transaction \( t_i \) is a set of items such that \( t_i \subseteq I \). An association rule is an implication of the form, \( X \rightarrow Y \), where \( X \subseteq I \), \( Y \subseteq I \), and \( X \cap Y = \emptyset \)

\( X \) (or \( Y \)) is a set of items, called an itemset.

A transaction \( t_i \in T \) is said to contain an itemset \( X \) if \( X \) is a subset of \( t_i \) (we also say that the itemset \( X \) covers \( t_i \)). The support count of \( X \) in \( T \)(denoted by \( X.count \)) is the number of
transactions in $T$ that contain $X$. The strength of a rule is measured by its **support** and **confidence**.

**Support**: The support of a rule, $X \rightarrow Y$, is the percentage of transactions in $T$ that contains $X \cup Y$, and can be seen as an estimate of the probability, $\Pr(X \cup Y)$. The rule support thus determines how frequent the rule is applicable in the transaction set $T$. Let $n$ be the number of transactions in $T$. The support of the rule $X \rightarrow Y$ is computed as follows:

$$\text{support} = \frac{(X \cup Y).\text{count}}{n}$$

Support is a useful measure because if it is too low, the rule may just occur due to chance. Furthermore, in a business environment, a rule covering too few cases (or transactions) may not be useful because it does not make business sense to act on such a rule (not profitable).

**Confidence**: The confidence of a rule, $X \rightarrow Y$, is the percentage of transactions in $T$ that contain $X$ also contain $Y$. It can be seen as an estimate of the conditional probability, $\Pr(Y | X)$. It is computed as follows:

$$\text{confidence} = \frac{(X \cup Y).\text{count}}{X.\text{count}}$$

Confidence thus determines the **predictability** of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict $Y$ from $X$. A rule with low predictability is of limited use.

### Apriori Algorithm

A large number of association rule mining algorithms have been developed with different mining efficiencies. Any algorithm should find the same set of rules though their computational efficiencies and memory requirements may be different. The best known mining algorithm is Apriori algorithm. The Apriori algorithm works in two steps:

1. **Generate all frequent itemsets**: A frequent itemset is an itemset that has transaction support above minimum support.
2. **Generate all confident association rules from frequent itemsets**: A confident association rule is a rule with confidence above minimum confidence.

The Apriori algorithm relies on apriori or downward closure property to generate all frequent itemsets.

**Downward Closure Property**: If an itemset has minimum support, then every non empty subset of this itemset also has minimum support.

### Data Formats for Association Rule Mining

Market basket data sets are natural formats for generating association rules. However different data sets can be tailored to fit to the definition of transactional databases, so that association rules mining algorithms can be applied to them. For example text document can be seen as transaction data. Each document is a transaction and each distinctive word is an item. Mining can also be performed on relational tables. It is straightforward to convert a table data set to a transaction data set if each attribute in table takes categorical values. We simply change each value to an attribute-value pair.
Mining with Multiple Minimum Supports

The key element that makes association rule mining practical is the minsup threshold. It is used to prune the search space and to limit the number of frequent itemsets and rules generated. However, using only a single minsup implicitly assumes that all items in the data are of the same nature and/or have similar frequencies in the database. This is often not the case in real-life applications. In many applications, some items appear very frequently in the data, while some other items rarely appear. If the frequencies of items vary a great deal, we will encounter two problems:

1. If the minsup is set too high, we will not find rules that involve infrequent items or rare items in the data.
2. In order to find rules that involve both frequent and rare items, we have to set the minsup very low.

However, this may cause combinatorial explosion and make mining impossible because those frequent items will be associated with one another in all possible ways. This dilemma is called the rare item problem. Using a single minsup for the whole data set is inadequate because it cannot capture the inherent natures and/or frequency differences of items in the database. By the natures of items we mean that some items, by nature, appear more frequently than others. For example, in a supermarket, people buy FoodProcessor and CookingPan much less frequently than Bread and Milk. The situation is the same for online stores. In general, those durable and/or expensive goods are bought less often, but each of them generates more profit. It is thus important to capture rules involving less frequent items. However, we must do so without allowing frequent items to produce too many meaningless rules with very low supports and cause combinatorial explosion. One common solution to this problem is to partition the data into several smaller blocks (subsets), each of which contains only items of similar frequencies. Mining is then done separately for each block using a different minsup. This approach is, however, not satisfactory because itemsets or rules that involve items across different blocks will not be found. A better solution is to allow the user to specify multiple minimum supports, i.e., to specify a different minimum item support (MIS) to each item. Thus, different itemsets need to satisfy different minimum supports depending on what items are in the itemsets. This model thus enables us to achieve our objective of finding itemsets involving rare items without causing frequent items to generate too many meaningless itemsets. An interesting by-product of this extended model is that it enables the user to easily instruct the algorithm to generate only itemsets that contain certain items but not itemsets that contain only the other items. This can be done by setting the MIS values to more than 100% (e.g., 101%) for these other items. This capability is very useful in practice because in many applications the user is only interested in certain types of itemsets or rules.

Basic Association Rules: Problems, Solutions and New Applications

Most of the research efforts in the scope of association rules have been oriented to simplify the rule set and to improve performance of algorithm. But these are not the only problems that can be found and when rules are generated and applied in different domains. Troubleshooting for them should also take into consideration the purpose of association model and data they come from. Some of the major drawbacks of association rule algorithms are as follows:
Association Rule Mining

- Obtaining huge number of rules
- Obtaining non interesting rules
- Low algorithm performance
- Cannot incorporate domain/ user defined knowledge
- Not suitable for supervised learning

Some of the recent studies have focused on overcoming these limitations. Many algorithms for obtaining a reduced number of rules with high support and confidence have been produced. However these measures are insufficient to determine if discovered associations are really useful. An important property of discovered association rules is that they should be interesting and useful. Though interestingness of rule is a subjective aspect, many researchers have tried to come up with some ways of measuring of interest. It has been suggested that the rules are interesting if they are unexpected (unknown to user) and actionable (users can do something with them to their advantage). Further some other measures namely: any-confidence, all confidence and bond has been suggested as alternative measures of interestingness. Some authors have considered alternative measures of interest as: gini index, entropy gain or chi-squared for database or a measure of implication called conviction. Most of the approaches for finding interesting rules require user participation to articulate his knowledge or to express what rules are interesting for him. Systems have been developed to analyze the discovered rules against user’s knowledge. Discovered rules can be pruned to remove redundant and insignificant rules and further user’s evaluation can be used to rank the rules. Unexpected patterns discovered may represent “holes” in domain knowledge which needs to be resolved. These patterns can thus be used to refine already existing beliefs.

Traditionally, association analysis has been considered as an unsupervised technique, so it has been applied for knowledge discovery tasks. Recent studies have shown that knowledge discovery algorithms such as association rule mining can be successfully applied for prediction in classification problems. In such cases the algorithms used for generating association rules must be tailored to peculiarities of predictions in order to build effective classifiers. Some work has been done, where association mining algorithms have been extended so that they can be used for classification/ prediction. A proposal of this category is Classification Based on Association (CBA) algorithm. The algorithm consists of two parts, a rule generator for finding association rules and a classifier builder based on these rules. Main contribution of this algorithm is possibility of making prediction on any attribute in database. Moreover, new incomplete observations can be classified.

In conclusion we can say that association rule mining is an important area of data mining research and a comparatively a younger member of data mining community. In addition to finding co-occurrence relation between items, which is basic objective, the algorithm has been applied for diverse applications. Many extensions of standard methods have been proposed. A major research area on association rules is interestingness of discovered rules. In fact its potential has still to be tapped, so that it can be tailored to solve different types of data mining problems.