MINING INFREQUENT ITEMSETS USING APRIORI ALGORITHM

Ms.L.Jayachitra, Mrs.S.Lakshmi Sridevi

1M.Phil student, Department of MCA, Hindustan University, Padur, Chennai
2Assistant Professor, Department of MCA, Hindustan University, Padur, Chennai

Abstract: The aim of Association Rule Mining is to find the correlation between data Items based on frequency of occurrence. Infrequent Item set mining is a variation of frequent item set mining where it finds the uninteresting patterns i.e., it finds the data items which occurs very rarely. Considering weight for each distinct item in a transaction independent manner adds effectiveness for finding frequent item set mining. Several articles related to frequent and weighted infrequent item set mining were proposed. This paper focus on reviewing various Existing Algorithms related to frequent and infrequent item set mining which creates a path for future researches in the field of Association Rule Mining.

Keywords: Infrequent Item set mining, Association Rule Mining, weight, Correlation.

1. INTRODUCTION

1.1 Introduction:

Knowledge Discovery and Data Mining (KDD) is playing an important role in extracting knowledge in this era of data overflow. KDD consists of many methods and techniques that can be applied to different data to extract knowledge. Some of the methods include association, classification, and clustering [14].

Association rule mining is the discovery of association relationships among a set of items in a dataset. Association rule mining has become an important data mining technique due to the descriptive and easily understandable nature of the rules. Although association rule mining was introduced to extract associations from market basket data [14], it has proved useful in many other domains (e.g. microarray data analysis, recommender systems and network intrusion detection). In the domain of market basket analysis, data consists of transactions where each is a set of items purchased by a customer. A common way of measuring the usefulness of association rules is to use the support-confidence framework introduced by [14].

Support of a rule is the percentage of transactions that carry all the items in the rule, and the confidence is the percentage of the transactions that carry all the items in the rule among those transactions that carry the items in the antecedent of the rule.

The problem of association rule mining can be stated as: Given a dataset of transactions, a threshold support (minsupport), and a threshold confidence (minconfidence); Generate all association rules from the set of transactions that have support greater than or equal to minsupport and confidence greater than or equal to minconfidence.

1.2 Association Rule:

Initially it was largely motivated to understand the market basket data, the results of which allowed companies to understand purchasing behavior and, as a result, better target market audiences. ARM is user centric as the objective is the
elicitation of interesting rules from which new knowledge can be derived. ARM is to facilitate the discovery, heuristically filter, and enable the presentation of these inferences or rules for subsequent interpretation by the user to determine their usefulness. ARM has been divided into two phase of process as follows:

**Phase 1**: Identify the sets of frequent items or item sets or pattern within the set of transaction using user-specified support threshold.

**Phase 2**: Generate inferences or rules from these above patterns using user-specified confidence threshold.

The above two phases are generated strong association rules from dataset. The first phase is called frequent item set construction or mining. That is extremely computational expensive than phase 2. The second phase is called association rule generation. That is, straight forward process. This phase computational complexity is negotiable to compare with first phase. There are two major problems in second phase. The first problem is rule quantity means that algorithms can produce large number of rules. The second problem is rule quality means that, all the rules are not interesting. The support and confidence measures play a vital role to filter unwanted item sets and rules from the mining process.

### 1.3 Types of Association Mining:

#### 1.3.1 Positive Association Rule Mining:

The classical association rules consider only items enumerated in transactions of the dataset. The positive relationship can be found between the set of items. The rules are generated from the positive related items. These rules are referred to as positive association rules. Most of the algorithms were developed for generating positive associations between items. These are useful to decision making [16].

The positive rules are classified as follows:

1. Boolean association rule
   a. Quantitative [1]
   b. Constrained rules [3]
   c. Sequential rules [4]
2. Qualitative association rule [1]
3. Spatial association rule
4. Temporal association rule

#### 1.3.2 Negative Association Rule Mining:

Negative association rules also consider the same items, but in addition the item also considers which were absent from transactions. The negative rules are generated from infrequent item sets. These rules play some important role in decision-making [16]. These are useful in market basket analysis to identify products that conflict with each other or products that complement each other. This is a difficult task, due to the fact that there are essential differences between positive and negative rule mining.

Brin et al [5] mentioned for the first time in the literature the notion of negative relationships.

The authors have used statistical chi-square test to verify the independence between two variables. The authors have also used correlation measure to determine the nature (positive or negative) of the relationship. The strong negative rules are mined by Savasere et al [6]. They combined positive frequent item sets with domain knowledge in the form of taxonomy.

#### 1.3.3 Constraint based Association Rule Mining:

The constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the rules. By doing this lots of cost of mining those rules that turned out to be not interesting can be saved. Usually constraints are provided by users. The constraints are classified as follows:

1. Knowledge based constraints [7]
2. Data constraints [8]
2. SYSTEM

2.1 Existing System:
In the existing system the Apriori algorithm was used to find the frequent item sets. For finding the frequent item sets in the medical database makes more time to spend the treatment with the healthier people.

So doctor’s time was wasted in the existing system. So we are moving to find the infrequent data in the proposed system.

2.2 Proposed System:
In the proposed algorithm is finding infrequent item sets using the Apriori property is applicable that is if we find an item set as infrequent then all its superset are considered as infrequent but here these super sets are not to be pruned as in Apriori and they are considered in to the solution as infrequent k- item sets.

Since the proposed system uses medical application to find the infrequent diseases in order to give the high response to abnormal diseases.

In the study of finding a better treatment approach for a special disease, researchers would like spend more time on studying an abnormal case rather than reading the millions of records of healthy people. In this scenario, more effort has been put into the development of infrequent item set mining.

3. CONCLUSION

Since the late 1990s, more and more researchers have realized the importance of infrequent patterns with the increasing demands from applications of anomaly detection, especially in medicine, genetics, molecular biology and network security.

4. FUTURE WORK

The proposed algorithm works efficiently to find the infrequent item sets. It finds all the infrequent item sets within one database scan. But the proposed method not considers any pruning strategy. It is better to implement any pruning strategy to improve the complexity of the proposed method.

REFERENCES


[22] Agrawal, Rakesh; and Srikan, Ramakrishnan; Fast algorithms for mining association rules in large databases.