Chapter 8

Conclusion and Future Works

Association rule mining in large databases is an important data mining technique and it is used extensively in the field of knowledge discovery. Some of the uses of association rules are understanding customers' buying patterns, detecting crime patterns in a particular city, detecting frauds in credit card systems, etc.

8.1 Finding Frequent Itemsets Plays an Important Role in Association Rule Mining

Generally, association rule mining is a two step process - finding frequent itemsets and finding association rules among the frequent itemsets. Out of these two steps, finding frequent itemsets is more important, difficult and challenging because if the frequent itemsets are available, finding association rules is a trivial task. There exist algorithms to find frequent itemsets from large databases. Among them, Apriori is one of the earliest and important algorithms. The algorithm is robust enough to find all frequent itemsets. The algorithm first finds candidate itemsets and then finds frequent itemsets from those candidate sets. However, it has been observed that the ratio of number of candidate itemsets to that of frequent itemsets is very high. In other words, very few candidate itemsets become actually frequent itemsets. There exist algorithms such as Bit_AssocRule, which use bitmap technique to find frequent itemsets. There also exist algorithms
which find frequent itemsets without candidate generation. \textit{FP-growth} is one such algorithm. However, it has been observed that performance of the algorithm degrades with the increase of minimum support.

### 8.1.1 Solutions Provided

Following solutions are provided to address the above problems.

- A faster version of \textit{Apriori} algorithm, which generates less number of candidate sets has been proposed.

- A modified and faster version of \textit{Bit-AssocRule} algorithm has also been proposed.

- To improve the performance of \textit{FP-growth}, a vertical partition based \textit{FP-growth} has been proposed.

### 8.1.2 Partitioning is Another Good Approach

Horizontal partitioning of databases is a good approach to find frequent itemsets from large databases. However, it does not perform well for high dimensional databases. This problem has been addressed by vertical partitioning of the databases. Vertical partitioning gives better results than horizontal partitioning in case of high dimensional databases.

### 8.2 Finding Frequent Itemsets for Dynamic Databases

One important feature of most of the databases is that they are dynamic in nature because records are added, updated, deleted very frequently. During the study of the algorithms, it has been observed that existing algorithms are not efficient to find frequent itemsets in dynamic databases. Moreover, to the best of our knowledge, no algorithm exists for distributed dynamic databases. Among the
existing algorithms for finding frequent itemsets in dynamic databases, Borders is the most important one. However, the algorithm scans the old database very frequently. To address this problem, a modified version of the algorithm has been proposed. The modified version does not require to scan the database frequently. Another problem is that Borders cannot be used directly for distributed dynamic databases. To solve this problem, a fully distributed version of Borders has been proposed. This distributed version can also be used for centralized database by partitioning the database and placing the partitions in different sites.

8.3 Feature Selection

Selecting relevant features is an important task in Decision Support Systems. There exist algorithms to select relevant features. These algorithms work for different types of databases and use different criteria to select relevant features. Moreover, these algorithms are very complex to implement. The thesis has reported a very simple algorithm to select relevant features using support count of the features. The algorithm is very simple and comparable to its counterparts in terms of relevant feature selections.

8.4 View Materialization

View materialization is an important technique used in data warehouse systems to reduce query response time. There exist algorithms for this purpose. All these algorithms have to work under some constraints such as disk-space constraint. Moreover, it has been observed that no algorithm has used support of the views for view selection. So, an algorithm has been proposed to select views for materialization. The important feature of the algorithm is that it has used density concept and support count of the views.
CHAPTER 8. CONCLUSION AND FUTURE WORKS

8.5 Future Works

Association rule mining is a vast area of research. So, it is not possible to cover every aspect of it in a stipulated period of time. Although it has been tried to cover as many aspects as possible, yet there are ample scopes for future works. Followings are some of the future works.

1. To develop a frequent itemset finding algorithm, which will generate the frequent itemsets without candidate generations for any high dimensional large market-basket databases and will also be capable of handling the minimum support condition i.e. the performance of the algorithm will not degrade even if value of minimum support count varies.

2. To extend the existing developments reported so far in the preceding chapters, to enable to work over spatial data, temporal data and any high dimensional categorical data.

3. To introduce soft-computing approach in the frequent itemset generation as well as in rule generation to discover more comprehensive and interesting patterns.

4. To develop better and robust dynamic rule mining algorithm over market-basket data as well as other huge data sources.

5. To incorporate data clustering technique or functional dependency approach to enable association mining over categorical and mixed types of data.

6. To analyze and develop quantitative association rule mining algorithms.

7. To develop a better feature selection technique using linear and non-linear manifolding techniques.

8. To explore the possibility of developing deviation analyzer (association rule mining based) for intrusion detection system.