CHAPTER 7

CONCLUSION AND PERSPECTIVES

This chapter summarizes the research contribution of the thesis. The directions for future work are also pointed out.

7.1 SUMMARY OF THE RESEARCH

The main topic of the thesis is the association rule mining with automatic specification of support thresholds. The following needs are identified:

- Mining with levelwise support thresholds
- Mining with itemwise support thresholds
- Rare Association Rule Mining

The goal is to find frequent and rare association rules of various kinds as listed below

- large frequent itemsets
- low support, high confidence association rules
- frequent itemsets involving rare items
- exact association rules

To achieve the goal and satisfy the identified needs, algorithms and some improvements over existing algorithms have been developed.
7.1.1 Mining with Levelwise Support Thresholds

The need for automatic specification of non uniform support thresholds at each level is investigated in Chapter 3, and two versions of the modified Apriori algorithm have been developed for the extraction of more number of large frequent itemsets.

First, a new technique namely DAS_Apriori is proposed to calculate support threshold at each level automatically by scanning the records in databases to generate sufficient number of frequent, interesting rules. The number of interesting rules to be mined is controlled by a user specified parameter. Next, another approach called DCS_Apriori is proposed to obtain the appropriate support thresholds at each level by using the previous level supports of itemsets. It assumes the support threshold to be considered in the current level depends on the support of itemsets at the previous level. Also, it should be less than the previous level support so that more number of large itemsets can be produced. It eliminates the need for the user specified parameter which is used for controlling the quantity of association rules. These algorithms are found to be useful in the extraction of large frequent itemsets.

7.1.2 Mining with Itemwise Support Thresholds

In chapter 4, an Automated_Apriori algorithm based on the CLS measure has been designed to calculate the itemwise support thresholds. The confidence and lift values are pushed into the support values and the itemwise support thresholds are calculated. These thresholds are used to prune the spurious frequent itemsets and helpful in generating more dependent, novel and relevant rules.
7.1.3 Rare Association Rule Mining

In chapter 5, Automated_Apriori_Rare algorithm has been designed to mine the frequent and rare association rules. This algorithm computes both levelwise and itemwise support thresholds for the generation of frequent and rare items. In Chapter 6, Exact_Apriori algorithm has been developed to establish a framework for the extraction of top rules with 100% confidence.

7.2 APPLICABILITY

The new algorithms proposed in this study suits for business data mining applications including market basket analysis, retail business, marketing and banking.

7.3 PERSPECTIVES

Future research may focus on filtering all rare association rules from the whole set of association rules. A theoretical framework should be designed to define the exact border that distinguishes the frequent and rare association rules.

Most well-known association rule mining algorithms have been developed with the assumption that the datasets can be accessed quickly and efficiently in the main memory. Though, the memory technology has improved rapidly, many huge datasets do not fit in available memory. Thus new algorithms with efficient data structures are needed for the efficient mining of frequent itemsets and association rules. It is necessary that such large databases can only be treated efficiently with parallel algorithms in a distributed environment. Parallel processing is ideally suited for addressing the issues of scalability.

Another challenge is to select the appropriate measures automatically to cater the need of various applications. This also requires the
development of algorithms with the capability of extracting the frequent and large itemsets using those measures. It would be interesting to develop such combined solutions for mining frequent and rare association rules in large databases.