CHAPTER 6

CONCLUSIONS AND FURTHER RESEARCH

6.1 SUMMARY OF CONTRIBUTIONS

In this thesis, we have investigated the issues in Association Rule Mining as the efficiency of the algorithms and the appropriateness of the results produced as output. We summarize our contributions based on the contributions in the areas below.

6.1.1 Efficiency of Algorithms

Numerous algorithms have been proposed in the recent past for the efficient mining of association rules, each in turn claiming to outperform its predecessors on a set of standard databases. In our thesis, our approach was to extend the applicability of the rules generated in one location to suit the needs of another location. The methodology proposed utilizes a partitioning strategy to determine the support of itemsets in the required combinations. In our model, we use the knowledge of the existence of the fine partitioned groups or the factions. The consistency of the factions for all situations is used for obtaining the rules for new situations. The change in proportions of factions, according to our methodology leads to new rules. A different method is to vary the rule mining process by including the locales even in the core mining process. The idea behind this is to obtain rules that are qualified by the values of the corresponding locales for these rules, e.g., (For Men: Diaper ⇒ beer), and (For Women: Diaper ⇒ ice cream). This approach does not require the identification of the factions beforehand.
The problems of this approach are manifold. First, this requires the recognition of the factors (disjoint locale values) by the mining process—potentially additional attributes to be handled by the mining technique (which typically scale exponentially with number of attributes). These locales, however, could be different in nature from which the association rules are mined. For example, the attributes could be quantitative (such as income, age, etc.) and those attributes involved in rules could be categorical. Numerous methods have been proposed to handle quantitative attributes, but all involve significant changes to the rule mining algorithm for categorical attributes.

Second, the nature of association between the factors and the items could be very different from those between the factors or between the items. The nature of the relation between the factors and the rules is intuitively causative. A number of techniques have been proposed in contemporary research that addresses discovery of causative rules between items. However, these techniques could be adapted to look for causal relations only between the factors and the rules, while looking for the required association rules between the items themselves. This implies that various measures are utilized for the identification of patterns between the factors between the items alone. This would significantly increase the mining complexity, even if it were to be feasible.

Moreover, it may not be easy to interpret directly, the qualified rules discovered by the above process. Qualifying a rule in terms of values for the causative factors may need to be simplified or eliminated when trying to get the whole picture of the set of rules for a new situation. For example, there might be a number of rules, where the qualifiers could be more complicated, like for Women in the ages 20-25 in rural areas or for Men in the ages 35-45 who watch news on television. For any decision maker, deciding which one of the rules is more dominating (and take further action based on it) may require
further processing to customize the qualifications for the new situation under consideration. Such customized rules are those that are produced directly using the CD approach.

In view of the additional complexity and large number of unsolved problems that need to be tackled together with this approach, we have decided to pursue the simple approach of Coalescent Data Set for obtaining the association rules for a new situation. Our experiment results indicate that a significant improvement in the accuracy of the rules can be obtained with the simple CD approach.

The mining of association rules derived from data has been investigated in data mining. Most of the research is focused on optimizing the process, developing methods to handle different kinds of data, developing appropriate significance metrics, and improving the user’s control of the process. However, the means to relate the discovered association rules is an equally important role in the knowledge discovery process. In this work, we provide a model to deal with this issue during the data mining phase. In particular, we provide an approach to apply the association rules discovered from the data set of one situation to another situation with different characteristics. We provide a model that distinguishes the difference between different situations using the concept of factions or fine-partitioned groups. Different situations have their factions in different proportions. The behavior of a particular faction is consistent across situations. However, different sets of rules are caused by the different proportions of the factions for different situations. Using this model, given the data set for one situation, we give a simple approach for estimating the set of rules derived for the second situation. Our CD approach requires no modification to the core mining process and, thus, can be applied to a variety of association rule mining techniques.
When patterns are a requisite for a new situation, using the patterns from the available dataset of the source situation is one option, if only the dataset for the first one is available. The results are erroneous when the relative proportion of factions for both situations is quite different. Using our techniques to obtain rules from the Coalescent Dataset can dramatically reduce these errors.

We have shown that the errors could be aroused by directly applying association rules derived from the data set of one situation to another situation. We have demonstrated the effectiveness of the CD approach in deriving the rules for the second situation with a high degree of accuracy. Hence, it can be seen that our approach provides an easy, yet powerful means to apply correctly the association rules from one situation to another. In our work, we narrowed down on the problem of extending the applicability of association rules from one situation to another. This problem is a significant characteristic of any rule-generating technique. It is an open problem for extending the Coalescent Dataset approach to other rule mining techniques.