Chapter 6

Conclusion and Future Scope

The computerization of several government and business activities and increasing use of bar codes for commercial products has contributed to explosive growth of data. This in turn demands for more powerful and affordable tools and technologies that can wisely transform the stored data into useful information which can be used for planning and decision making. Association rule mining has widely been accepted to be the most promising technology for discovering the associations and useful correlations among the data stored in the large database repositories. Mining useful, comprehensive and interesting associations is one of the most important tasks for several business houses. Theses association rules are useful for making policy decisions and marketing strategies. Classification rules, clustering and generalization are some other data mining techniques. An extensive study of literature on data mining technologies revealed that in spite of theoretically sufficient work reported in the
field of association rule mining, practically it lags behind much and needs improvement. Apriori is the most successful and popular technique for mining association rules from large databases. This approach also has some limitations. Repeated scanning of the database and generation of large number of intermediate itemsets are major hindrance factors affecting the usefulness of this approach.

A number of other approaches, that have been proposed over time, had tried to improve the efficiency of the association rule mining process. These approaches are either based on the horizontal layout of the database such as DHP, DIC and Partition or based on vertical layout databases such as FP-Tree, FP-Growth and Eclat to name a few. All these and many other approaches have tried to reduce the scanning time and simplifying the candidate generation process. Efficiency of any association rule mining approach depends upon the time that it takes to generate frequent itemsets and the time for generating association rule from these frequent itemsets. Thus any approach that focuses on this aspect may prove somewhat beneficial. An exhaustive survey of association rule mining technologies has been carried out by the authors of this thesis and found that still the problem of efficiency in terms of time consumed and storage space requires is not solved satisfactorily. The main factor that affects the performance of the existing approaches is that all these approaches use brute-force approach for generating frequent itemsets and building rules. All the possible itemsets are generated from scratch and then those which satisfy the requirement of minimum support and confidence are kept for use discarding the rest of the itemsets and rules. Thus a major portion of time consumed is associated with generating the itemsets and rules which ultimately prove uninteresting.
Ant Colony Optimization (ACO) is a subfield of swarm intelligence which is a popular meta-heuristic for solving combinatorial optimization problems. ACO is the problem solving approach which is based on the foraging behavior of real ant colonies. The ants in ant colonies make use of a feedback mechanism to communicate the information about the beneficial path. Thus for finding a path from nest to food source, ants cooperate with each other and discover the shortest path. Ants follow the trails marked by other ants and ultimately converge to a common and beneficial route. ACO is used for solving several NP-hard problems that are otherwise difficult to be solved. Travelling salesman problem, vehicle routing, network routing, scheduling and assignment problems are some of the hot areas where ACO has paved its roots. As mining of association rules can be viewed as the problem of finding optimal sets of rules and patterns that satisfies a set of constraints, ACO can be applied in this area.

The research work in this thesis is dedicated to designing novel approaches for mining association rules from data repositories using the ant colony meta-heuristic which are efficient in terms of search quality and search time. Efficiency in terms of search quality means the discovered rules are more interesting and comprehensive whereas efficiency in terms of time means that the rules and patterns are discovered in minimum possible time. The research work in this thesis presents three novel approaches for mining frequent itemsets and association rules. All the proposed approaches are graph based. The first two approaches are designed to discover the frequent itemsets from the database and the third approach is dedicated to discover the association rules directly without the need to generate frequent itemsets first. All the approaches have been evaluated practically on two major criteria viz. time consumed and number of itemsets/rules produced. The approaches have been
implemented in Microsoft Visual Studio 2010 using C#. Then the program is executed with real datasets taken from UCI repository and results were analyzed.

Existing approaches in the literature like Apriori and FP-tree uses a user-specified minimum support and minimum confidence threshold for detecting frequent itemsets. Apriori generates itemsets in an incremental manner. Every time an itemset is generated by joining the previous itemsets and selecting only the itemsets which have support higher than minimum support threshold specified by the user. Thus, if a different support threshold is specified, the algorithm has to re-run and the whole process is repeated from scratch. Similarly in FP-tree and other existing approaches, if the minimum support or minimum confidence threshold is changed, the complete procedure is started again. This problem is tackled in the graph based approach proposed in the Section 3.6. The proposed approach resolves the problem altogether by calculating the itemsets for all support values. The frequent itemsets are generated in such a way that in a single run all the frequent itemsets for all support thresholds are generated. Once the graph and the table for finding frequent itemsets are constructed successfully, we can find the most frequent itemsets of any given length (for every threshold value) without any sort of rework. After generation of the frequent itemset table, generating frequent itemset for a particular support threshold is just a matter of a table lookup. The proposed approach is efficient from the point of view of database scans. Most of the previous methods perform repeated scans of the database which, in turn, waste CPU time and results in degradation in the overall performance. The proposed approach when applied to the sample databases reported the same frequent itemsets in single database scan, as produced by several existing approaches in multi database scans. The cost of this improvement is the increase in the storage space needed for storing the graph and the table. With a graph of ‘n’
nodes, we need a matrix of order n*n and thus the need for storage space increases with increase in number of nodes. To minimize the storage requirement the proposed approach can be applied directly on small and medium size databases. For large databases, vertical partitioning can be used as a preprocessing step to filter out the infrequent items. This preprocessing step will make the numbers of nodes and the edges manageable. Also, sparse matrix can be used to store the graph because the entries in the lower diagonal part of the graph are empty. Thus with efficient implementation of matrix and table the proposed approach can be made more efficient.

The 2nd proposed approach follows the ACO approach for searching the frequent itemsets. The graph constructed in this approach is compact and thus it also overcomes the need of large storage space. The approach is very efficient for lower support values in large databases. The Apriori approach takes less time for generating frequent itemsets for high support values but the time increases very much for low support threshold values. But the proposed approach takes less time for both low and high support values. The experimental results proved that the approach is better than Apriori with respect to the time consumed for generating frequent itemsets. As far as number of frequent itemsets are concerned, the proposed approach is still better than Apriori because it generates all the high support maximal frequent itemsets. Apriori generates more candidate itemsets than the proposed approach but all the ultimate rules generated are much less. The proposed approach generates all maximal frequent itemsets and thus generates all high confidence rules.

The 3rd proposes approach is another improvement over the previous approach because it generates association rules directly without undergoing two-step process of generating
frequent itemsets and then generating association rules. This approach generates association rules in a single step. This is done by creating the rule at the time of checking the support of an item and placing the item in the antecedent or consequent part on the basis of the support values. The approach takes less time than Apriori approach and also generates all high confidence rules. Thus the proposed approaches for frequent itemsets mining and association rules mining are efficient both in terms of time consumed and number of rules generated.

The proposed approaches in the present research work have considered only numeric and categorical attributes. The approaches are not tested for continuous domains and complex attributes like image and geographic data. Such type of data needs to be converted to simple numeric attributes before applying these approaches. One future research direction may be to extend the work for applying these ACO based approaches for mining complex associations like multilevel association, multidimensional associations and associations in continuous data. The proposed approaches are not applicable for datasets containing advanced data such as spatiotemporal, multimedia and stream data. Thus the work can also be extended by modifying the proposed approaches for such complex database repositories. This work can provide a base for future research in this direction.