Conclusion

In this study the focus has been on exploring the applicability of the Rough Set Theory in the three step data mining process considered for this purpose. The Application of RST has been observed and evaluated firstly in data preprocessing, namely in discretization. Secondly, as a technique for granulation for granulation tree based association rule mining, and finally, for the interpretation of the mined knowledge by way of analyzing the clusters.

7.1 Summary

Knowledge Discovery in Databases is a complex process of knowledge extraction from large quantity of data with the goal of obtaining meaning and consequently understanding of data, as well as to acquire new knowledge. In Chapter 1 a discussion on KDD and its various steps has been presented. Definition, tasks and common techniques of data mining are listed to present a framework for this research. A brief review of the basic concepts and definition of Rough Set Theory have also been presented in this chapter.

A detail survey of literature on data mining has been presented in Chapter 2 with special emphasis on Clustering and Association Rule Mining. In this chapter a modified 3 step data mining process model of the model proposed by Wu et al. has been considered. Data preprocessing, mining the data, and evaluation-validation &
interpretation-presentation are the three steps considered under this model for this work.

A two phase technique for discretization based on clustering and Rough Set Theory in Chapter 3 was proposed. The observations from the experiments based on the parameters of evaluation indicate good results comparable with the best as recorded in the literature. For $N$ the number of objects in the database, the algorithmic complexity for discretization was observed to be of the order of $N \log N$. The performance of the proposed technique was compared with seven discretization algorithms namely, equal-width, equal frequency, Patterson-Niblett, IEM, Maximum Entropy, CADD and CAIM has been compared with the proposed technique, using the two evaluation parameters CAIR value and the total number of intervals. Results of the proposed scheme exhibited a good trade off between CAIR value and the total number of intervals compared to other discretization algorithms.

Granulation tree, a novel approach for granular computing based on clustering and Rough Set Theory presented has been in Chapter 4. It is a set theoretic approach for its application in data mining. The rough set theoretic concept positive region has been applied for construction of the granulation tree in case of dealing with numeric values of attributes. The structure Granulation Tree has also been defined in this chapter to present the result of the granulation process. The to evaluate the result of the proposed method the four evaluation parameters proposed in the Chapter are, 1) number of Granules 2) number of fine granules 3) number of Granules containing objects of more than one class, and 4) number of object which cause the result 3. The
proposed technique is semi-supervised, exhibited the flexibility of selecting the level of granule, quality of granule for a suitable application.

In Chapter 5 a new technique Granulation Tree based Quantitative Association Rules (GTQARs) to mine Quantitative Association Rules has been presented. GTQARs method works in three phases. In first phase granulation tree of the given database is constructed. In the second phase frequent itemsets are generated from the granulation tree, and finally in third phase quantitative association rules are obtained from the frequent itemsets using standard methods. The salient feature of the proposed system is that, for different sets of support and confidence the method yielded results using the same granulation tree. Therefore without re-executing the entire algorithm different set of quantitative association rule for different support and confidence were possible to mine.

Clustering, one of the major data mining functionalities has been taken up Chapter 6. An Attribute based Hierarchical Clustering algorithm (AHC) and the application of Rough Set Theory in cluster analysis has been proposed in this chapter. The complexity of AHC is bounded by $N^2$; where $N$ is the number of objects. Preprocessing the continuous attributes using the discretization scheme proposed in Chapter 3, AHC can be applied to dataset with attributes of different types. Since objects are not clustered using the distance measure, in this method arbitrary shape clusters can be obtained. Method for Cluster analysis based on rough set theoretic concept reducts. RST based evaluation of attribute significance for results of AHC are noted to be higher than the results corresponding DBSCAN. The interpretability of
the clusters yielded by AHC, using the RST concept reducts has been observed to be better than the clusters produced by DBSCAN. Using the reducts for the clusters obtained by AHC, characterizing or differentiating clusters, require less number of attribute as compared to DBSCAN.

The two concepts of rough set theory applied in this study are the concept of positive region $\text{POS}_B(D)$ and the reduct. The concept $\text{POS}_B(D)$ was exploited to optimize the relevance of attribute in $B$, the attribute considered for discretization with respect to a decision attribute $D$. The concept of positive region is also used to generate the granulation tree of the database with numeric attributes. In this work Granulation tree based method has been proposed for quantitative association rule mining. The granulation tree approach lead to encouraging results for quantitative association rule mining. The concept of reduct has been found to be useful for cluster analysis. By characterizing a cluster and discriminating pairs of clusters based on reducts, has been useful in obtaining reasons for existence of cluster and difference between clusters. Finally, an attribute based hierarchical clustering algorithm has also been proposed which satisfies a good trade off between a number of criteria such as ability to handle attribute of different types, discover clusters of arbitrary shape, minimal input of domain knowledge and result not affected by the order of data input.
7.2 Future Work

- Some other parameters like entropy may be used to measure the goodness of discretization. Some new measures for goodness of discretization can be suggested for unlabeled data.

- To evaluate the granulation process and to qualify granules some parameters may be identified.

- Attribute relevance analysis can be embedded in the granulation algorithm to further improve the performance by reducing the number of granules and also may increase the percentage of good granules. Some heuristics may also be used to obtain good granules of variable size.

- The use of granulation tree for other data mining techniques/tasks may be explored.

- To improve the efficiency of the proposed GTQARs method may be implemented in parallel programming paradigm.