7.1 Conclusion

The theoretical underpinnings of Rough Set Theory are explored in the thesis in a data mining point of view. Rough Set properties such as indiscernibility, reduct, core, discernibility matrix etc. are applied effectively in attribute reduction, rule mining and evaluation, missing value handling and clustering. This results in several novel algorithms in these selected areas.

Attribute Reduction

In attribute reduction three different RST based feature selection approaches namely positive region based attribute reduction, discernibility matrix based approach and Proportional Rough Set (PRS) relevance method are discussed in detail. The expensive reduct computation is the major limitation of all these RST based feature selection approaches. As an alternative to these methods two different extended discernibility matrix based models are suggested.

Based on these models three novel feature selection algorithms are proposed. First feature selection algorithm, DFRM is designed to perform feature selection in medium sized datasets having small number of decision classes. The algorithm is implemented and applied on an artificial Car dataset and on a real world dataset consisting of the signs and symptoms of the Learning Disabilities in school age children. The experimental results show the success of the proposed approach in selecting the significant
features from the data domain. The efficiency of the method is evaluated by comparing the performance with the PRS relevance approach.

The number of decision classes and the size of the input dataset are two important factors affecting the performance of the proposed DFRM algorithm. To overcome this, a novel feature selection algorithm ADDM is proposed. In ADDM, as a preprocessing step, we introduced a filter based approach to remove outlier points thereby improving the quality of the input dataset. The core idea behind this ADDM is the introduction of the Centre of Gravity (CoG) of a set of points to reduce the complexity of the overall feature selection process and a discernibility matrix based feature relevance weighting model to rank the features of the dataset. The proposed algorithm is applied on a set of features extracted from handwritten character images. The effectiveness of the reduction process is evaluated by comparing the classification performance with original set of features and ADDM reduced feature set. For comparison purpose a reduct based PRS relevance approach is applied on the same dataset. The results show that the proposed ADDM outperforms the computationally expensive PRS relevance approach. The drastic reduction in the size of the dataset ensures the scalability of the algorithm to handle high dimensional datasets.

With the help of various feature selection strategies such as Mean Selection, Half Selection and Selection by Threshold, the feature selection can be automated. Incorporating this idea, an improved version of the ADDM is presented. The performance of the different feature selection strategies is analyzed and it is observed that Selection by Threshold gives better performance compared to others. A comparison of the efficiencies of various selection strategies shows that the proposed Discernibility Matrix
based approach gives maximum efficiency compared to the PRS relevance approach in the case of Mean Selection strategy.

**Rule Mining and Evaluation**

As a part of this study an RST based rule induction algorithm namely Learning from Examples Module Version 2 (LEM2) is implemented with a modification. It is applied on a dataset related to farming and rules are generated and analyzed.

As a post processing operation to rule mining, effective methods are required to extract interesting, relevant and novel rules automatically. An objective and domain dependent measure called *Degree of Rule Significance* is introduced in the thesis as a new rule evaluation measure. Based on this measure, a Rule Ranking (RR) algorithm is designed to rank and extract significant rules. In RR algorithm the idea of decision relative discernibility matrix in RST is used to eliminate the expensive reduct computation. The effectiveness of the *Degree of Rule Significance* in ranking the generated decision rules is established through experiments. Further the efficiency of this new measure and the resulting RR algorithm is established with the help of two different Rough Set based approaches. In the first approach, reduct/core attributes are generated from the original dataset. Then the effectiveness of the proposed measure in selecting significant rules is measured in terms of the presence of these reduct/core attributes in the top ranked rules. In the second approach, reduct/core rules are generated by constructing a new decision table by considering the generated decision rules as conditional attributes. The presence of these reduct/core rules in the set of ranked rules with high *Degree of Rule Significance* proves that the proposed method provides an automatic and effective way of ranking rules.
Chapter 7

Missing Value Handling

The indiscernibility relation of RST provides a well understood formal model to handle missing values. Different RST based paradigms for handling missing values are discussed. A detailed analysis of a popular RST based missing value handling method namely Characteristic Set based approach is carried out. Considering its drawbacks, an improvement in this method using a modified similarity relation is suggested. Through experiments it is proved that the proposed approach is better than that of the Characteristic Set based approach.

Clustering

In this thesis some conventional clustering methods and their limitations are reviewed in detail. A new clustering model is proposed, which is a synergic combination of RST and Graph Theory. In the proposed approach, for cluster generation, the rigid requirement of indiscernibility is relaxed, and at each stage of clustering, to unfold the real structure of the clusters the idea of complete graph is employed. It is shown that the proposed algorithm is effective for generating natural clusters from small and medium sized datasets.

7.2 Future works

In this study we mainly concentrated on classical Rough Set Theory. Hence the main limitation of the proposed Rough Set based approaches included in this thesis is the restrictive requirement that all data is discrete. There is no way to consider real-valued or noisy data. To address this issue, we would like to extend the proposed works by integrating it with powerful soft computing techniques viz. Fuzzy sets, Neural networks etc.
It would be highly beneficial to investigate the performance of the proposed attribute reduction methods by employing it in other application domains such as character recognition, medical imaging and bio-metric based detection.

By experimenting the proposed methods with variety of datasets it is possible to extend the evaluations to more application domains.

In the present study, the main focus is application of Rough Sets in data mining. Hence, the proposed algorithms are compared only with the other Rough Set based approaches contributed in the respective areas. It is interesting if these methods are compared with other existing non Rough Set based data mining methods.

In order to conduct various experiments, in this study, only limited size datasets are used. Because of the explosive growth of available information, a series of experiments and investigations are necessary to establish the potential utility of the proposed methods in large datasets.

In clustering, the algorithm proposed as a part of the study is tested only with an example dataset. An extensive study of this Rough Set based clustering approach is left as future work.

In MODILEM2 algorithm, to form the decision table used for rule mining, the actual attribute values are mapped on to some intervals and all these intervals are crisp in nature. In order to overcome this drawback, we are focusing on designing a novel algorithm by incorporating soft computing technique such as fuzzy logic, which may lead to better results.