CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1. CONCLUSION

Knowledge discovery from large data sets using classic data mining techniques has been proved to be difficult due to large size in both dimension and samples. In real applications, data sets often consist of many noisy, redundant, and irrelevant features, resulting in degrading the classification accuracy and increasing the complexity exponentially. Due to the inherent nature, the analysis of the quality of data sets is difficult and very limited approaches about this issue can be found in the literature. There is a need for quality of data, thus the quality of data is ultimately important. As a commonly used technique in data preprocessing, feature selection selects a subset of informative attributes or variables to build models describing data and then classification is done. By removing redundant and irrelevant or noise features, dimensionality reduction of the data for feature selection can improve the predictive accuracy and the comprehensibility of the predictors or classifiers. Various approaches have been used in this research for large number of datasets.

Initially, in this work an ensemble classification approach is introduced to classify the noisy and irrelevant features implanted in data sets and perceive the quality of the structure of data sets. An enhanced KNN method is used as the preprocessing approach to find the missing values from the whole dataset. Then the feature selection of the datasets is processed using

Enhanced Genetic Algorithm combined with Kernel PCA SVM Algorithm. Homogeneous and heterogeneous ensemble classification approaches are used in this research work for classification. In homogeneous ensemble classification model, Fuzzy KNN classifier is used. To increase the performance of heterogeneous ensemble classifier, the classification framework is proposed with the set of classifiers such as Fuzzy KNN, ANFIS, FRB etc.

Another approach proposed a feature selection ensemble approach which combines the outputs of multiple feature selectors approaches to improve the prediction accuracy for the subsequent classifier learning tasks. Three Novel implementation steps performed in feature selection ensemble concept imparts that there are construction of ensemble approach which may be applied to many subset evaluation techniques such as Hybrid genetic particle swarm optimization algorithm and Hybrid Artificial Bee Colony Algorithm and search algorithms for optimal feature subset results for prediction accuracy and evaluation of results after the classification is performed for larger dataset. Then the classification of the dataset is performed by using improved Fuzzy rule based classifier and Fuzzy Rough Positive Region based Nearest Neighbour. Experimental results are compared with the existing feature selection methods and classification methods for various large datasets. It shows that the proposed approaches are considerably more accurate than that of the existing methods.

7.2 FUTURE WORK

In high dimensional dataset, outlier detection is an important problem that has applications in many fields. High dimensional datasets are common in such applications. The major problem is that this research doesn’t focus on the outlier while performing the classification methods and feature selection methods. The future research will focus on the outlier detection

problem along with ensemble feature selection and ensemble classification methods.

Among the existing outlier detection methods, which use a Distance Based Outlier (DB-Outlier) detection, is one of the most generalizable and simplest approaches. The proposed work focuses on the future direction such as the similarity and distance based clustering methods to perform outlier detection before the classification of the task. It improves the feature selection accuracy and the classification accuracy It determined the outliers by calculating distance based between data points.

The result indicates that the proposed method works well in detecting the most suitable subspace based on users' standpoints. In spite of that, there are some factors which may lead to inaccurate results. Therefore, a more appropriate Fitness Value function can be tried to define the future work, based on the probability distribution which is also applicable to find the outliers in the dataset and this can be continued in the proposed research.

In these outlier detection methods the distance based measures cluster-distance bounds by optimizing the algorithm, the computational cost will be reduced and the dimensions and size of the data set can be scaled. Less number of random Inputs Outputs are focused over several recently proposed indexes. Because of the cluster with exact nearest neighbor search, the search time is reduced. Possibly by optimizing the clustering algorithm, the query cluster distance bounds can be further tightened, so as to optimize the cluster distance bounds. Future efforts would be directed toward this and other related problems.