CHAPTER 6

RESULTS AND CONCLUSION

6.1 INTRODUCTION

Feature-selection methods are particularly welcome in interdisciplinary collaborations because the selected features retain the original meanings domain experts are familiar with. The rapid developments in computer science and engineering allow for data collection at an unprecedented speed and present new challenges to feature selection. Wide data sets, which have a huge number of features but relatively few instances, introduce a novel challenge to feature selection. This type of data supports a vast number of models. Examples of such data are microarrays, transaction logs, and Web data. Unlabeled data presents another challenge; the lack of class labels compounds the already difficult problem of feature selection. Feature-selection methods attempt to explore intrinsic properties of data by employing statistics or information theory. This chapter presents the results and conclusions drawn from new algorithms, new approaches for different types data set, namely categorical and text data.

6.2 DISCUSSION OF THE RESEARCH OUTCOME

This thesis proposes three different feature selection approaches. Along with ACO and J4.8, a total of six types of hybrid classifiers generate classification rules in data sets.
Three feature selection algorithms namely FCBF, DDC-DIC and MIEF are based on information theoretic approaches and they considered on different feature interaction. The Fast Correlation based Feature selection algorithm (FCBF) have interaction between feature and class attribute. The second algorithm namely Decision dependent and Decision Independent algorithm (DDC-DIC) consider the interaction between pair of features \( F_i, F_j \). The third algorithm Mutual information based Evaluation Function (MIEF) considers the joint mutual information between class and pair of features \( I(Y, F_i, F_j) \). It eliminates noisy and redundant features. The problems of higher computation time of ACO and excessive rule generation of J4.8 classifier are overcome by the proposed six hybrid classifier schemes. The performances of classifiers are compared with respect to the classification accuracy \( (C_A) \), the number of rule generated \( (NR_G) \) and computation time \( (T_C) \).

We have compared the performance of ACO, J4.8 with six hybrid classifiers in six public domain categorical data sets. The results show that, concerning classification accuracy, the number of rule generated and computation time of all six hybrids classifiers obtained somewhat better results in all data sets than ACO and J4.8.

Regarding the number of rule generation for large dataset, namely, Coil 2000 (Tic-Tac) and US Census data, the hybrid J4.8 classifier is approximately six times better than J4.8 classifier. The computation time of hybrid ACO classifier is reduced approximately 50 times as compared to ACO classifier for the slightly improved classification accuracy for Coil and US Census data.

Among the hybrid classifiers, the performance of hybrid ACO classifiers are better than hybrid J4.8 classifiers for medium and large dataset (G2 and G3), with respect to classification accuracy \( (C_A) \) and standard
deviation. Regarding the number of rules generated, all the three hybrid ACO classifiers provide simple rules than hybrid J4.8 classifiers for all six data set. For US Census dataset, the number of rules generated for the hybrid ACO classifiers are approximately ten times less than that of hybrid J4.8 classifiers.

Therefore, overall one can say that hybrid ACO is comparable to hybrid J4.8 with respect to classification accuracy, and give much simpler (smaller) rule lists than hybrid J4.8. On the other hand, hybrid ACO has slightly higher computation time than hybrid J4.8 classifier. Therefore, hybrid ACO, and hybrid J4.8 classifiers seems particularly advantageous when they are important to minimize the number of discovered rules and reduce the computation time in order to improve comprehensibility of the discovered knowledge. It can be argued that this point is important in many data-mining applications, where discovered knowledge will be shown to a human user as a support for intelligent decision making.

This thesis also proposes two additional feature selection algorithms based on feature relevance and interaction. The first algorithm information theoretic based interact (IT-IN) algorithm is a two stage algorithm which concern the feature relevance and consistency. The second algorithm, Dependency based Interact algorithm (DIA) is a three stage algorithm, which concerns the relevance, redundancy, and consistency of the features. To evaluate the classification accuracy of IT-IN and DIA, four well known classifiers namely ELM, Naive Bayes, J4.8 and SVM are used for ten UCI repository categorical datasets. The proposed IT-IN and DIA algorithms select relevant features and efficiently handle feature interaction. They rank the features to overcome the feature order problem through a special hashing data structure. This is used to avoid repeated scanning of a data set by taking advantage of the inherent properties of the C- Contribution measure, resulting significant computation time.
In Naive Bayes classifier, the performance of IT-IN is better than Interact, FCBF, Relief and CFS algorithms. IT-IN has 7 maximal values of classification accuracy, over ten datasets whose performance is significantly better than Unselect, where the win/tie/loss for IT-IN is 7/0/3. Among the proposed IT-IN and DIA algorithm, the performance of DIA is better than IT-IN. The classification accuracy of most of the data set and the average accuracy of the DIA is higher than IT-IN. The win/tie/loss for DIA is 8/0/2.

In J4.8 classifier the performance of IT-IN is distinct than other four feature selectors in average value and win/tie/loss (6/1/3). The classification accuracy of IT-IN and DIA algorithm, the performance of DIA and IT-IN are same in Lung Cancer, Iris, Tic-tac, Soy, Wine, and Zoo. The win/tie/loss for DIA is similar to IT-IN.

In SVM classifier the distinctness of proposed IT-IN is higher than the other four selectors where the win/tie/loss is 7/0/3. IT-IN and DIA algorithms, they are equally contributing in Lung Cancer, Breast cancer, Iris, US Census, Tic-tac, and Wine. The classification accuracy of DIA is slightly higher in Zoo and Splice data set and the win/tie/loss with unselect feature is similar to IT-IN.

In ELM classifier the predominance of IT-IN is also distinct and better than other four selectors in the aspects of average value and win/tie/loss. The win/tie/loss of proposed IT-IN with ELM classifier is 9/0/1. This outperforms than other feature selector in same classifier as well as other classifiers. The average classification accuracy of ELM with other remaining feature selection methods is higher than other classifier schemes. The performance of DIA with ELM classifier is higher in Lung Cancer, CMC, Iris, US Census, soy and Splice data set and it provides equal performance in Breast Cancer and Tic Tac data. For Wine and Zoo data the IT-IN performs better than DIA. The win/tie/loss with unselect feature is similar to IT-IN.
The overall all classification schemes IT-IN and DIA have the highest classification performance over three-quarter datasets in all classifiers, which is higher than other feature selection algorithms. Among IT-IN and DIA the performances of DIA superior to IT-IN which provide the three stage filtering approach and more suitable for Lung cancer, Breast Cancer, iris data analysis.

The other dimension of this thesis is concerned on text data. It proposes a new “term weighting scheme” called Modified Inverse Document Frequency (MIDF) to improve the performance of text categorization. The document represented in MIDF is trained using the SVM classifier with linear, polynomial and Radial Basis Function kernels (RBF), and Neural Network. The experiments are carried out in Reuters-21578 corpora. The performance of TF.IDF, WIDF, and MIDF based SVM TC (linear, RBF kernel) schemes are compared with TF.IDF, WIDF, MIDF based NN based TC. The performances of macro-averaged F1-measure and categorization cost for the top ten categories are compared for three term weighting schemes.

From the experimental results, the overall performance macro-averaged F1-measure and cost measure of proposed MIDF term weighting scheme with SVM–RBF classifier is better than other existing TF.IDF and WIDF term weighting schemes with SVM-Linear kernel and NN approaches. In SVM Linear Kernel, the F1 measure of the proposed MIDF is 2% higher than TF.IDF and 10% higher than WIDF for most of the categories. The macro average of MIDF is 6% better than TF.IDF and more than 25% higher than WIDF approach. At the same time the cost measure for False alarm is negligibly 10% less than TF.IDF approach and 30% lower than WIDF.

In SVM RBF Kernel, the F1 measure of the proposed MIDF is approximately 10% higher than TF.IDF and approximately 6% higher than WIDF for most of the categories. The macro average of MIDF is 25% better than TF.IDF and more than 13% higher than WIDF approach. At the same
time the cost measure for False alarm is negligibly 31% less than TF.IDF approach and 16% lower than WIDF.

In NN classifier, the performance of MIDF is better than TF.IDF and WIDF with respect to F1 measure and Cost measure for top ten categories. But overall performance of the SVM-RBF is superior to other method in terms of F1 measure and cost.

This thesis also focuses on feature selection for email spam filtering approaches for Ling spam public mail corpora datasets. In this thesis the proposed Modified Inverse Document Frequency (MIDF) is used as term weighting scheme and FCBF based feature selection is to select subset of spam features. The efficiency and effectiveness of proposed spam filter is demonstrated through extensive experiments compared with other term weighting schemes TF.IDF, and WIDF with feature selection algorithms namely IG and SU. The performance of the proposed spam filter is compared with different classifier namely, Naive Bayes, J4.8, SVM and ELM classifiers. To improve the performance of spam filter a new approach, called weighted term frequency (WTF) is used to increase the weights of spam words, which is present in subject or content of the spam documents.

Experimental results show that, the proposed FCBF based feature selection with MIDF provides 2% better performance in Naive Bayes and J4.8 classifier. In SVM (Linear, Poly, RBF) classifiers the performance of all term weighting schemes with three feature selection approaches are more are less similar.

In ELM classifier all the term weighting schemes with three feature selection algorithms provide better accuracy than other classifier schemes, which is 1% higher than other schemes. Among the three term weighting schemes MIDF, TF.IDF, WIDF and three feature selection approaches, FCBF based MIDF schemes is superior to all other schemes.
6.3 CONCLUSION

From the experimental results of six types of hybrid classifiers on categorical data set, it is found that the information theoretical based algorithms namely FCBF, DDC-DIC, MIEF with ACO and J4.8 plays the major role for better classification accuracy, number of rule generation, and reduction of computation time of hybrid classifiers. Hence the hybrid classifiers are recommended for large data set.

The Proposed IT-IN and DIA feature selection algorithms select minimal features than the FCBF, DDC-DIC and Relief algorithms and provide better classification accuracy for Naive Bayes, J4.8, SVM and ELM than the other feature selection algorithms. The computation time for IT-IN and DIA are reduced by the hashing data structure which is used to reduce the repeated scanning of data.

The proposed MIDF based term weighting scheme with SVM classifier provide better result than existing TF.IDF and WIDF based term weighting schemes. As the online computation time of MIDF method is inherently less, F1 measure is high and false alarm cost is very minimum. Hence MIDF based TC scheme can be recommended for applications where there is a large flow of dynamic information that needs to be organized.

In text categorization the proposed term weighting scheme MIDF with FCBF based feature selection which also improves the classification accuracy of ELM, SVM, J4.8 and Naive Bayes classifiers. Hence it is more suitable for spam filtering approach. From the thesis it is found that information theoretic based feature selection namely FCBF, DDC-DIC, MIEF and the proposed IT-IN, DIA performs major role in numerical, categorical and text data types which select only relevant data and remove the redundancy to improve the classifier accuracy.
6.3.1 **Scope for Further Research**

There are further scopes for extension in the following area.

- Symmetrical uncertainty measure only handles nominal or discrete values, our current method requires continuous values be discretized, which opens the opportunity to investigate how different discretization methods affect the performance of FCBF, DDC-DIC, MIEF, IT-IN and DIA algorithms. It would be interesting to explore measures that can handle all types of values or ways of combining different measures under the framework of relevance and redundancy analysis. And different hybrid classifier schemes can be developed for Text categorization.

- Apply different heuristic function for feature relevance in proposed IT-IN and DIA algorithms and they can be extended for text categorization and Spam filtering.

- The Spam Filtering approach can be further developed using different heuristic function for feature evaluation metrics with different feature selection algorithms are the promising approach in spam filtering.

- The topics of spam emails change constantly. Thus, adaptive learning ability of single-class learning techniques will be highly relevant and desirable in the spam filtering context, as well as in similar application scenarios.

- In this study, we use only SVM, Naive Bayes, J4.8 and ELM as base classifiers in spam filtering. Inclusion and empirical evaluation of different base classifiers represents an interesting research direction.