CHAPTER 6

CONCLUSION AND FUTURE WORK

The main contributions of this thesis are:

1. The proposal of a novel algorithm to fine-tune the tuples in a dataset that can help improve the prediction accuracy of the classifiers.

2. A spectrum of novel algorithm is presented for feature selection that can improve the detection performance of classifiers with the evaluation of the influence of feature pre-selection on the prediction accuracy of classifiers with a real world dataset and

3. The proposal of methods to handle the ambiguous classification case of equal class probability for each class of DT’s, NB and K-NN classifiers, which has the potential to further improve classification performance.

6.1 SUMMARY OF THE WORK DONE

A novel data pre-processing algorithm has been proposed to fine-tune the tuples in a dataset that identifies and removes the redundant data followed by correcting the conflicting data to enhance the classifier’s detection performance. The increase in the performance of the classifiers over this tuned data through an extensive empirical study on UCI repository datasets is demonstrated. The results proved that the claim is valid.
A novel evolving feature selection algorithms taking the advantages of Bayes theorem, Class association rule and Information gain to improve the predictive accuracy of classifier has been proposed. Bayes theorem and class association rule are used to discover dependency information among features. In addition to that, feature selection has been improvised by Information gain which selects features based on their importance. The proposed feature selection methods can remove unwanted features to improve classification performance. As shown in the experimental results, after feature selection by the Bayes theorem, Class association rule and Information gain to control false discovery rate, the classification performance of DTs and NB classifiers were significantly improved.

An attempt is made to develop a modification to standard decision tree algorithms to handle cases where a node has equal probability for more than one class and where two features have equal information gain with regard to the splitting criterion. The first kind of tie occurs when there are equal proportions of the target class outcome in the leaf node’s records that leads to a situation where majority voting can’t be applied. It is proposed to base the prediction of the result on the IFM, NB, k-NN and ARM methods. The second kind of tie occurs when two or more attributes with the same information gain are used for splitting an interior node during tree induction. This is resolved by the selection of the attribute that has already occurred in a prior splitting condition elsewhere in the tree induction. This heuristic procedure clearly avoids the addition of new features to the model. The attribute that had higher information gain ranking at a parent node could also be used to aid the proposed method’s accuracy and efficiency.

An interesting solution is provided to the inability of discriminating between cases when the chosen discriminating attributes do not provide enough information. The proposed solution provides an option to extend the
NB and $k$-NN classification beyond the maximum posterior classification using a form of PMM and IPMM.

A case study for threatening e-mail classification through DTs, NB, SVM and NN is presented. A decision-tree based classification method named *Ad Infinitum* to detect e-mails that contain information on terrorism is proposed. The proposed classification method is a revised version of DT’s by integrating user’s domain knowledge. The proposed method on an e-mail corpus showing a superior performance to traditional decision tree induction algorithms was tested. The data preprocessing steps, the training data, the decision tree produced, and performance validation are reported on classifying emails regarding terrorism.

### 6.2 CONCLUDING REMARKS

This thesis proposes to improve the detection performance of the basic classification algorithms through Dataset Tuning, Feature Selection and Exception Handling. Classification is one of the most common data mining tasks used frequently for data categorization and analysis in the industry and research. Real-world data mining sometimes mainly deals with noisy information like redundant and conflicting data, irrelevant and redundant features. These data may degrade the performance of any classification algorithms. Inductive learning algorithms, in general, perform well on data that have been pre-processed to reduce complexity. By themselves they are not particularly effective in reducing data complexity while learning difficult concepts. Feature construction has been shown to reduce complexity of space spanned by input data. An iterative algorithm for enhancing the performance of any classification algorithms through the use of dataset tuning and feature selection as a pre-processing step is proposed in this thesis. The procedure on three learning methods, namely C4.5, NB and lazy learner is applied and
shown as improvement in performance. Handling Exceptions can also increase the performance. The results on the UCI datasets show that the proposed modification procedure can provide better predictive accuracy over the existing methods in the literature.

Some directions of future work include:

1. Experiments to investigate these findings on text classification.
2. Extending these methodologies to medical informatics.
3. Experiments to find out the other sources in order to improve classifier performance further.
4. Extending this work to construct a more robust and scalable classifier.