ABSTRACT

This thesis presents a spectrum of novel solutions for enhancing the performance of classification tasks in Knowledge Discovery Process by taking into account, redundant and inconsistent data, irrelevant and redundant features and exceptions while performing classification tasks (equal probability and information gain) that can arise often in real world datasets. These characteristics are very often neglected by state-of-the-art classification algorithms. The main focus of this thesis is Knowledge Discovery in Databases (KDD). Four different domains related to KDD namely, Dataset Tuning, Feature Selection, Handling Exceptions in Classification algorithms and Classification of Threatening E-mail have been investigated.

In real world datasets, lots of redundant and conflicting data exist that affect the performance of the classification algorithms. They have to be removed to increase the efficiency and the accuracy of the classifiers. Dataset tuning is a pre-processing method used to fine-tune the tuples in a dataset for noise elimination. Dataset tuning is done to remove redundant data and to correct the conflicting data.

Feature selection is a fundamental problem in data mining to select relevant features and cast away irrelevant and redundant features based on some evaluation criteria. It is well known that correlated and irrelevant features may degrade the performance of the classification algorithms. In this thesis, feature selection algorithms such as Bayes Feature Selector, Class
Association Rule- Information Gain Feature Selector and Bayes Theorem - Information Gain Feature Selector are proposed.

Classification is a primary data mining task aimed at learning a function that classifies a database record into one of the several predefined classes based on the values of the record attributes. The Decision Tree, Naive Bayes, k-Nearest Neighbour and Support Vector Machine are the popular machine learning methods for classification. However prior classification algorithms do not handle exceptions namely, (i) when the information gain for two or more attributes is the same in decision tree induction algorithm (ii) when a leaf node has equal probability for more than one class in decision tree induction algorithm (iii) when the posterior probabilities estimated by Naive Bayes are equal and (iv) when a tie between classes can occur in a k-Nearest Neighbour classifier.

Overall this thesis deals with the performance of different classification algorithms and the impact of Dataset Tuning, Feature Selection and Exception Handling method on DTs, NB, SVM and k-NN classifiers. The significant improvement in performance was seen with k-NN classifier by dataset tuning. In addition to that, the use of the Bayes theorem, Class association rule, and Information gain in feature pre-selection phase statistically improved the performance of DTs, NB, SVM and NN classifiers over the conventional feature selection methods. Besides, the proposed exception handling methods have the potential to further improve classification performance.