CHAPTER 1

INTRODUCTION

The data dimensionality is growing rapidly in a broad spectrum of application domains which poses a great challenge to traditional data mining and machine learning tasks in terms of scalability, effectiveness and speed. Without modifying existing traditional data mining and machine learning algorithms, feature selection can alleviate the problem of high dimensionality. Accordingly it has been well recognized as a key step during the process of knowledge discovery from the high dimensional data and has attracted growing interest from both the academia and the industry in the past. By removing irrelevant and redundant features from the high dimensional data based on certain feature evaluation measures, feature selection can help in enhancing the generalization capability of learning models, speed up the learning process and improve the model interpretability. A great variety of feature selection algorithms has been developed with a focus on improving the predictive accuracy of learning models while reducing dimensionality and model complexity.

Although it is important to evaluate the performance of feature selection algorithms based on the generalization ability of models built on the selected features, researchers are shifting their focus towards more comprehensive evaluation parameters. One such parameter is the stability of feature selection algorithms. It is the insensitivity of the outcome of a feature selection algorithm to small changes in the training set. Of late it is being considered as a potential evaluation parameter as it is critical, particularly for application domains where feature selection is used as a knowledge discovery tool. Biomarker discovery for cancer diagnosis is one such example. In such type of application domains a good feature selection algorithm should not only help to achieve better prediction performance but also produce stable results even if the training data varies slightly.
1.1 HIGH-DIMENSIONAL DATA AND FEATURE SELECTION

These days, researchers are impressed by the huge amount of data generated by high throughput technologies. But they are unable to work on this data as it is accumulating at a rate unmatchable by human’s capability to process it. In many scientific and application domains, the introduction of delicate and complicated technologies allows people to perform deeper and broader investigation on huge volume of data. For example, high-throughput genomic and proteomic technologies are widely used in cancer research to erect better predictive models of diagnosis, prognosis and therapy, to find out and describe key signaling networks and to find new targets for drug development (Clarke et al. 2008); a combination of information retrieval and high-dimensional learning schemes in modern text classification systems can be better adopted for automated population of hierarchical catalogues of Web resources, filtering of spam, categorizing documents and even automated ranking of essays.

In the new scenario, data is presented in an unprecedented large scale which makes the traditional machine learning approaches ineffective. In order to reduce the complexity of data mining problems and adapt existing machine learning approaches to the new scenario without losing the effectiveness of data, data preprocessing can be used. Data preprocessing established itself as a key step during the entire knowledge discovery process. Among various data preprocessing techniques, feature selection is widely used due to its good theoretical and empirical properties.

Feature selection has been a promising area of research and development since the 1970s in statistical pattern recognition, machine learning, and data mining. They are widely applied to many domains such as text classification, image retrieval, intrusion detection, and genomic analysis (Golub et al. 1999). A great number of feature selection algorithms have been developed with a focus on improving classification accuracy while reducing dimensionality and model complexity.

In the medical field, the identification and validation of molecular biomarkers (genes) for cancer diagnosis, prognosis, and therapeutic targets is an important problem in cancer genomics. It is essential to choose a small number of highly discriminative genes for validation because of the prolonged, costly and effortful nature of clinical and biological validations (Pepe et al. 2001). For example, the DNA microarray technology
allows to monitor the expression levels of a large number of genes at once. In the recent years, analysis of gene expression data is one of the major topics in health informatics (Ng and Pei 2007). For instance, classification of DNA microarray data allows the detection of unseen patterns in gene expression profiles and increases the opportunity for accurate cancer classification. The main problem in classifying gene expression data is the curse of dimensionality problem (small number of samples compared to large number of genes) (Ong Huey Fang et al. 2011) that caused data over-fitting in many of the studies conducted on this data, which demands additional validation. To overcome this issue, feature selection is used to identify the most informative differentially expressed genes and to remove the irrelevant genes (Shang and Shen 2006).

1.2 PROBLEM STATEMENT

Although the feature selection techniques aid in improving the classification accuracy, one of the problems with existing feature selection algorithms is the variation in the selected feature subset in the successive runs on the same dataset with small changes. Ultimately, the unsteadiness of feature selection results will decrease the confidence of experts on the outcome.

While using feature selection methods, not only model performance but also the robustness of feature selection process is important. In this work we focus on feature selection for cancer classification from microarray data which is a good example for high dimensional data. For example, in microarray data analysis, determination of features (genes) that are differentially expressed in cancerous cells could assist domain experts in designing and planning more appropriate treatments for cancer patients. But the instability of feature selection techniques will reduce the confidence in discovered markers (genes).

Since the SVM-RFE, an embedded feature selection algorithm has outperformed most of its counterparts in the field of cancer classification we have used it as the baseline. From the literature it is found that the curse of dimensionality (relatively small number of samples in high dimensional data) and sample variations in the training dataset are main reasons for instability in microarray data. Likewise the greediness in choosing the optimal feature set is an instability issue of SVM-RFE.
1.3 OBJECTIVE OF THE RESEARCH

In order to address the instability issues caused by microarray data and SVM-RFE, this work proposes a more stable feature selection framework that uses bootstrapping and K-fold cross validation techniques. But the proposed framework increases the computational complexity of the selection process due to its highly iterative nature.

Hence the objectives are

(i) To formulate a new re-ranking strategy for improving the feature selection stability of SVM-RFE.
(ii) To propose a new feature elimination strategy for reducing the computational complexity of SVM-RFE.
(iii) To improve the classification accuracy of the classifier built from the selected features.

The problems identified in the feature selection process and the solutions proposed are stated below.

Problem 1

To enhance the stability of the feature selection algorithm (SVM-RFE) which in turn improves the generalizing ability of the SVM classifier built using the selected features for cancer classification.

Proposed Solutions

(i) To overcome the main instability issue of the microarray data (the curse of dimensionality problem) balanced bootstrapping technique and k-fold Cross Validation are used.

(ii) A new re-ranking strategy is formulated. The genes that occupied the top most position more number of times by the iterative stable feature selection framework are declared as the top ranking genes. This method has increased computational complexity of the feature selection process. So further work is needed to strengthen the re-ranking strategy and to bring down the computational complexity.
Problem 2

To propose a more stable re-ranking strategy that can identify highly class discriminative genes than the previously proposed re-ranking strategy and to bring down the computational complexity.

Proposed Solutions

(i) Here k-fold Cross Validation is used to generate various training sets from the same set of samples. This is to overcome the curse of dimensionality problem.

(ii) Cumulative Ranking Score (CRS) is a parameter formulated to determine the class discriminative ability of each gene and genes having high CRS are considered as highly discriminative genes. Compared to the previous approach, here the feature selection framework was iterated only less number of times which reduced the computational complexity comparatively. But still the computational complexity is more.

Problem 3

To propose a fast and stable SVM-RFE by incorporating the stability factor into the algorithm design.

Proposed Solution

(i) The SVM is trained only once using all the training samples. In the following iterations, only the samples (Support Vectors) that have the chance of being misclassified are used for training. This reduces the time taken for identifying the highly class discriminative features drastically.

(ii) In every iteration of the proposed method, effort is taken to eliminate more number of less class discriminative features. This reduces the computational time.

(iii) A parameter called co-efficient of large margin is formulated and used along with the weight vector of the features for ranking them. This improves the stability of the feature selection process and the generalization ability of the classifier built.
1.4 ORGANIZATION OF THE THESIS

The rest of this thesis is organized as follows.

Chapter 2

It presents the literature review on three broader areas (i) Reported work on SVM and SVM-RFE for Cancer Classification (ii) Reported work on Stable Feature Selection (iii) Reported work on Fast SVM-RFE

Chapter 3

In this chapter the fundamental biological terminologies, the experimental steps of typical microarray, application of microarray data in medicine and biology and the challenges faced in microarray data analysis are discussed.

Chapter 4

Here the discussion about the objectives of feature selection, requirement of feature selection in microarray data analysis, classifier and feature selection algorithm used in this thesis and limitation of SVM-RFE are included.

Chapter 5

This chapter focuses on the importance of stable feature selection, reasons for instability in feature selection, stable feature selection approaches and need of fast feature selection method.

Chapter 6

This chapter presents the robust feature selection frame work, overall flow of the framework, the re-ranking strategy used, and graphical representation of occurrence the top 10 genes, proof of biological relevance from gene ontology and literature proof of the top 10 genes.
Chapter 7

In this chapter the description of the general framework of accurate and stable feature selection process, computation of Cumulative Ranking Score, graphical representation of the results, comparison of classification accuracy and literature proof of the top 10 genes are presented.

Chapter 8

Discuss the general framework of the proposed accelerated stable SVM-RFE, the hypothesis used for stable feature selection, performance comparisons in terms of generalization ability, computational complexity and stability are presented here.

Chapter 9

This chapter concludes the thesis and outlines future research directions.