STUDIES ON SOME GENE SELECTION METHODS FOR CANCER CLASSIFICATION BASED ON MICROARRAY DATA

A thesis

submitted in partial fulfillment of the requirement for the Degree of

Master of Computer Science and Engineering

of

Jadavpur University

By

Manish Babu

Registration No.: 128998 of 2014-15

Examination Roll No.: M4CSE1612

Under the Guidance of

Prof. Kamal Sarkar

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

2016

STUDIES ON SOME GENE SELECTION METHODS FOR CANCER CLASSIFICATION BASED ON MICROARRAY DATA

A thesis

submitted in partial fulfillment of the requirement for the Degree of

Master of Computer Science and Engineering

of

Jadavpur University

By

Manish Babu

Registration No.: 128998 of 2014-15

Examination Roll No.: M4CSE1612

Under the Guidance of

Prof. Kamal Sarkar

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

2016

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Recommendation

This is to certify that the dissertation entitled "Studies on Gene Selection Methods for Cancer Classification based on Microarray Data" has been carried out by Manish Babu (University Registration No.: 128998 of 2014-15, Examination Roll No.: M4CSE1612) under my guidance and supervision and be accepted in partial fulfillment of the requirement for the Degree of Master of Computer Science and Engineering. The research results presented in the thesis have not been included in any other paper submitted for the award of any degree in any other University or Institute.

……

Prof. Kamal Sarkar(Thesis Supervisor)
Department of Computer Science and Engineering
Jadavpur University, Kolkata-32

Countersigned

……

Prof. Debesh Kumar Das
Head, Department of Computer Science and Engineering,
Jadavpur University, Kolkata-32.

……

Prof. Sivaji Bandyopadhyay
Dean, Faculty of Engineering and Technology,
Jadavpur University, Kolkata-32.

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Approval*

This is to certify that the thesis entitled "Studies on some Gene Selection Methods for Cancer Classification based on Microarray Data" is a bona-fide record of work carried out by Manish Babu in partial fulfillment of the requirements for the award of the degree of Master of Computer Science and Engineering in the Department of Computer Science and Engineering, Jadavpur University during the period of June 2015 to May 2016. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

…….

Signature of Examiner 1

Date:

…….

Signature of Examiner 2

Date:

*Only in case the thesis is approved

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis entitled "Studies on some Gene Selection Methods for Cancer Classification based on Microarray Data" contains literature survey and original research work by the undersigned candidate, as part of his Degree of Master of Computer Science & Engineering.

All information have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Manish Babu
Registration No: 128998 of 2014-15
Exam Roll No.: M4CSE1612
Thesis Title: Studies on some Gene Selection Methods for Cancer Classification based on Microarray Data.

……..

Signature with Date

Acknowledgement

I would like to start by thanking the holy trinity for helping me deploy all the right resources and for shaping me into a better human being.

I would like to express my deepest gratitude to my advisor, **Prof. Kamal Sarkar**, Department of Computer Science and Engineering, Jadavpur University for introducing me to the wonderful world of Computational Biology. I deeply thank to him for his admirable guidance, care,belief, patience and for providing me an excellent atmosphere for doing research. Our numerous scientific discussions and his many constructive comments have greatly improved this work.

I would like to thank **Prof. Debesh Kumar Das**, Head, Department of Computer Science and Engineering, Jadavpur University and **Prof. Sivaji Bandyopadhyay**, Dean, Faculty of Engineering and Technology, Jadavpur University for providing me with moral support at times of need.

I would like to thank Mr. Soumya Prakash Rana, my lab mate for helping me to understand a number of concepts, the fruit full conversations, advices, and especially for the laughs, arguments and the great atmosphere in the lab.

I would like to thank Mr. Tapas Nayak, Mr. Joy Mahapatra and Mr. Om Prakash Kumar for always extending a helping hand at times of need and for always motivating me.

Most importantly none of this would have been possible without the love and support of my family. I want to thank my parents, my brother and my sister for

giving me the best support a family can give and for cleverly putting things in perspective. I specially want to thank my mother whose forbearance and whole hearted support helped this endeavor succeed. Finally I would also like to mention my grandparents without whose aegis this work could have never been able to flourish.

This thesis would not have been completed without the inspiration and support of a number of wonderful individuals — my thanks and appreciation to all of them for being part of this journey and making this thesis possible.

……..

Manish Babu
Registration No: 128998 of 2014-15
Exam Roll No.: M4CSE1612
Department of Computer Science & Engineering
Jadavpur University

Abstract

Microarray technology has enriched the study of gene expression in such a way that scientists are now able to measure the expression levels of thousands of genes in a single experiment. Microarray gene expression data gained great importance in recent years due to its role in disease diagnoses which help to choose the appropriate treatment plan for patients.A key challenge in biomedical studies in recent years is the classification of samples into categories such as cases and control (individuals who carry some illness and others who do not). This is done by first learning how to classify, based on a training set containing labeled samples from the two populations, and then predicting the label of new samples. Each sample consists of gene expression measurements.This technology has shifted a new era in molecular classification, interpreting gene expression data remains a difficult problem and an active research area due to their native nature of "high dimensional low sample size".An important sub-problem in such studies is that of gene selection.In microarray data, the number of features (gene expression levels) far exceeds the number of samples. Standard classifiers do not work well in such situation. Selecting only the genes that are most relevant for the discrimination between the two categories helps in constructing better classifiers, both in terms of accuracy and in terms of efficiency.

This thesis aims on a comparative study of state-of-the-art feature selection methods, classification methods, and the combination of them, based on gene expression data. We compared the efficiency of two different classification methods including: support vector machines and k-nearest neighbor, and eight different feature selection methods, including: t-test, Chi-Square test, Information Gain,

mRMR, relief-F, SVM-RFE, Genetic Algorithm and Differential Evaluation. Accuracywas used to evaluate the classification performance. Two well known gene expression data sets Leukemia and Colon Tumor were used for this study.Different experiments have been applied to compare the performance of the classification methods with and without performing feature selection. Results revealed the important role of feature selection in classifying gene expression data. By performing feature selection, the classification accuracy can be significantly boosted by using a small number of genes.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Gene Expression Microarray Data

In every living organism, there are basic hereditary units known as genes. Genes are segment deoxyribonucleic acid (DNA) that holds genetic information for encoding specific cellular ribonucleic acid (RNA) and proteins. The central dogma of molecular biology describes how proteins are produced from DNA which is divided into two main steps illustrated in Figure 1.1. The first step, known as transcription, describes that a gene in DNA is expressed by transferring its coded information into messenger ribonucleic acid (mRNA). The second step, known as translation,

describes that proteins are produced based on information from the mRNA. This process of transcription and translation that allows a gene to be expressed as proteins is known as gene expression [1].

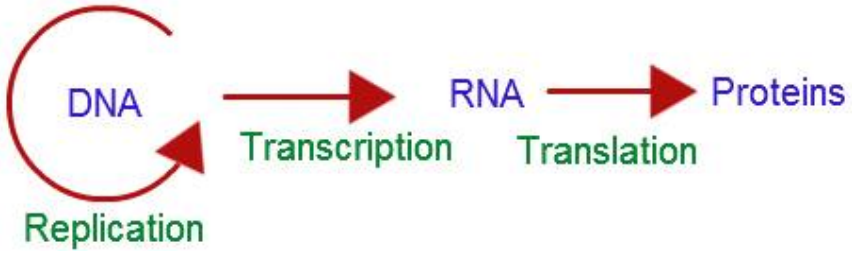


Figure 1.1: Central Dogma of Molecular Biology
(Image Courtesy: http://hyperphysics.phy-astr.gsu.edu/hbase/organic/imgorg/cendog.gif)

Although almost every cell in the body of an organism contains an exact same copy of the DNA, mRNA level varies over time and different cell types as well as varies within cells under different conditions. The amount of mRNA being expressed plays an important role. The more mRNA produced, the more proteins produced. The level of mRNA is used as a measure of gene expression. In other words, gene expression level indicates the amount of mRNA produced in a cell during protein synthesis. Some tumors occur because of mutation of certain genes and it is reflected in the change of the expression level of these certain genes, which means the genes are expressed abnormally in particular cells, either being up-regulated (express in a higher amount), down-regulated (express in a lower amount) or not being expressed [2]. The difference between the gene expression levels produces a specific profile for each gene. There are a number of experimental techniques to measure gene expression such as expression vector, fluorescent hybridization, and DNA microarray.

The microarray technology produces large datasets with expression values for thousands of genes (2,000-20,000), but with only a small number of samples. The data are usually organized in a matrix of n rows and m columns, which is known as a gene expression profile. The rows correspond to genes and the columns

correspond to the samples. Figure 1.2 shows an illustration of a gene expression profile.

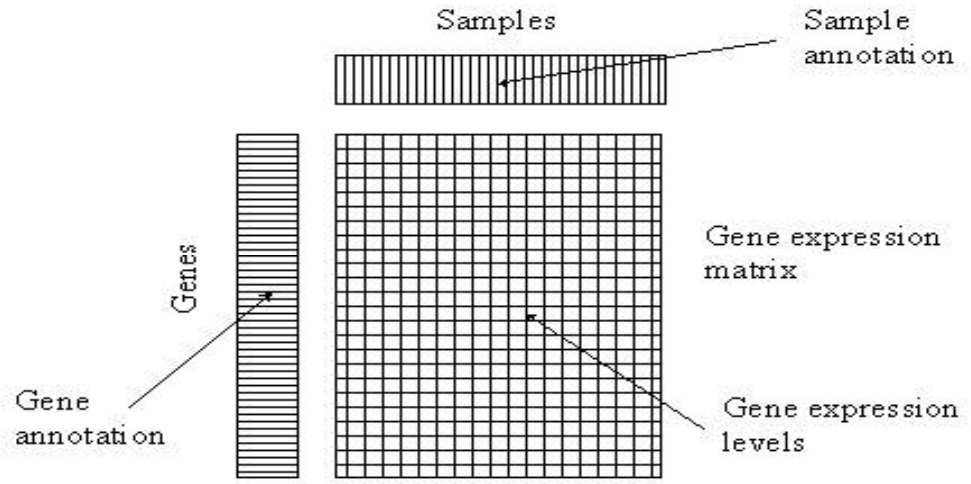


Figure 1.2: An illustration of gene expression profile.
(Image Courtesy: http://www.people.vcu.edu/~mreimers/OGMDA/gene_expression_matrix.gif)

1.2 Importance of Microarray Data

DNA microarray technology is one of the fastest-growing new technologies that have empowered the study of gene expression in such a way that scientists can now measure the expression levels of large numbers of genes in a single experiment rather than performing several experiments and gathering data for a single gene at a time [3]. If we try to analyze different genes in a biological lab, it will take around 2-3 months to identify whether a particular person is affected by a particular disease or not. Within that period, may be patient's problem become critical, which could have been improved or cured if it would have been found at earlier and proper

medication would have been used. If we use gene expression microarray data and perform computational procedure then maybe we would be able to find in few days that the person is affected by a particular disease or not. This also helps to discover the roles played by specific genes in the development of diseases. It can also be helpful in finding out the outcome of a treatment on a particular person. Particularly, DNA microarray enables scientists and physicians to understand the pathophysiological mechanisms in diagnoses and choosing the appropriate treatment plan for patients at the gene expression level.

It is essentially important to analyze the gene expression profiles. For instance, in analyzing cancer tumors, the biologists hope to distinguish and select genes that are responsible for the growth of tumor cells from the experiment. This information can be used to identify and classify a new patient's sample into their class. However, considering the amount and complexity of the gene expression data, it is impossible to analyze the $n \times m$ gene expression matrix manually.

Thus, computational methods are urgently needed. Computationally speaking, the problem is a classification problem. The goal is to find a way to differentiate healthy samples from samples with diseases, or differentiate samples with different level of diseases. Due to the nature of the problem, which is "high dimension low sample size", traditional classification methods often do not perform well. Dimension reduction or feature selection is essential in such problems.

1.3 Usage of Microarray Data for Cancer Classification

Cancer is one of the world's most serious diseases in modern society and a major cause of death worldwide. Traditional diagnostics methods are based mainly on the morphological and clinical appearance of cancer, but have limited contributions as cancer usually results from other environmental factors. There are several causes of cancer (carcinogens) such as smoke, radiation, synthetic chemicals, polluted water, and others that may accelerate the mutations and many undiscovered causes. On the other hand, a need to select the most informative genes from wide data sets, removal of redundant &uninformative genes and decreases noise, confusion and complexity and increase the chances for identification of diseases and prediction of various outcomes like cancer types is mandatory [3]. One of the challenging tasks in cancer diagnosis is how to identify salient expression genes from thousands of genes in microarray data that can directly contribute to the phenotype or symptom of disease [4]. The development of array technologies indicates the possibility of early detection and accurate prediction of cancer. Through these technologies, it is possible to get thousands of gene expression levels simultaneously through arrays, and also the ability to make use to know and find out whether it is cancer or not, and if it is then classify cancer [5]. Thus, there is a need to identify the informative genes that contribute to a cancerous state. An informative gene is a gene that is useful and relevant for cancer classification [6]. Cancer classification, which can help to improve health care of patients and the quality of life of individuals, is essential for cancer diagnosis and drug discovery [4]. Cancer classification refers to the process of constructing a model on the microarray dataset and then distinguishing one type of samples from other types within the induced model[4]. Microarray is a device or a technology used to measure expression levels of thousands of genes simultaneously in a cell mixture, and finally produces a microarray data, which is

also known as gene expression data. The task of cancer classificationusing microarray data is to classify tissue samples into related classes of phenotypes such as cancer versus normal [7].

1.4 Importance of Gene Selection

One of the main problems with the Gene Microarray Data is the "Curse of Dimensionality ". Gene Microarray data contains very less numbers of samples n but very high numbers of genes m. With the "curse of dimensionality" of gene expression microarray data, it is common that a large number of genes are not informative for cancer classification because they are either irrelevant or redundant. Only a small number of genes may be important[17]. Early and accurate detection and classification of cancer is critical to the wellbeing of patients. The need for a method or algorithms for cancer identification is important and has a great value in providing better treatment and this can be done through analysis of genetic data. For practical use,an algorithm has to be fast and accurate as well as easy to implement, test, and maintain. The optimal algorithm for a given task would have adequate performance with minimal implementation complexity [18].For the above reasons, feature/gene selection techniques become apparently needed for both domains: biology and computation/statistics. Selecting gene markers that present the maximum discriminative power between cancerous and normal cells or between different types of cancerous cells is a central step towards understanding the underlying biological process which becomes one of the vital research areas in microarray data analysis. The identification of a small number of key genes can help biologists to get a better understanding of the disease mechanism and have a simpler principle for diagnosis. On the other hand, computational and statistic experts are more concerned in dealing with the noisy data with redundant features and avoiding the over-fitting issues, which often happen when there are only small number of

samples. Thus, feature selection can significantly ease computational burden of the classification task and can reduce the number of genes by removing the meaningless features not just without the loss of the classification accuracy but even significantly enhance it [5-10].

1.5 Problem Statement

Cancer diagnosis nowadays is based on clinical evaluation and physical examination and also refers to medical history. But this diagnosis takes a long time. It might be too late to cure the patient if a tumor is found in its critical stage. Also, it is very important for diagnostic research to develop diagnostic procedures based on inexpensive microarray data that have adequate number of genes to detect diseases. The classification of gene expression data is challenging due to the enormous number of genes relative to the number of samples. It is common that a large number of genes are not informative for classification because they are either irrelevant or redundant. For that, it is significant to find whether a small number of genes are sufficient for gene expression classification. This thesis focuses on a comparative study of the state-of-the-art gene selection and classification algorithms and proposes an effective way to combine some of them together to achieve a more stable performance.

1.6 Thesis organization

The **Chapter1** of the thesis contains the preliminary concepts of molecular biology which starts from the central dogma and a brief discussion about gene expression data. And finally gives a short idea about Machine Learning for gene expression data and problem statement. **Chapter 2** discuss about the related work, that are already done in this field.**Chapter 3** explains basic details about normalization,

various gene selection and classification methods. It also give an overview of evaluation methods. **Chapter 4** describes evaluation and the experimental results. Finally, this thesis ends with conclusion, discussion and future works that can be done in **Chapter** 5.

Chapter 2

Related Work

Many machine learning methods have been introduced into microarray classification to attempt to learn the gene expression data pattern that can distinguish between different classes of samples in recent years.

The work done by Mehdi [19] evaluated and compared the efficiency of different classification methods, including SVM, neural network, Bayesian classification, decision tree (J84, ID3) and random forest methods. Also, a number of clustering methods including K-means, density-based clustering, and expectation maximization clustering were applied to eight different binary (two class) microarray datasets. Further, the efficiency of the feature selection methods including support vector machine recursive feature elimination (SVM-RFE), Chi-squared, and correlation-based feature selection were compared. Ten-fold cross validation was used to calculate the accuracy of the classifiers. First the classification methods were applied to all datasets without performing any feature selection. In most datasets SVM and neural networks performed better than other classification methods. Then the effect of feature selection methods was examined on the different classification methods. Various number of genes were tested (500, 200,100, and 50) and the top 50 genes were selected because it gave a good accuracy, consumed less processing time, and required less memory configurations comparing to others. Almost in all cases, the accuracy performance of classifiers was improved after applying feature selections methods to the datasets. In all cases

SVM-RFE performed very well when it was applied with SVM classification methods.

Liu, Li and Wong [20] presented a comparative study of five feature selection methods using two datasets (Leukemia and Ovarian cancer). The feature selection methods are: entropy-based, Chi-squared, t-statistics, correlation-based, and signal-to-noise statistic. The top 20 genes that have the highest score in chi-squared, t-statistics, and signal-to-noise were selected, and all the features recommended by correlation-based were selected. For entropy, features having an entropy value less than 0.1 were selected if existed, or the 20 features with the lowest entropy values were selected otherwise. The effectiveness of these features was evaluated using KNN, C4.5, naïve Bayes, and SVM classifiers. SVM reported the least error rate among the other classification methods when applied to the datasets without feature selection. When applying feature selection on the datasets, the accuracy performance of the four classifiers was greatly improved in most cases.

The work by Tao Li [21] studied and compared the results of multiclass classification using many feature selection and classification methods on nine multiclass gene expression datasets. The "RankGene" software was used to select the informative and related genes on the training set with eight methods supported in this software: information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, one-dimensional SVM, and t-statistics. The top 150 ranked genes in every dataset were selected. The multiclass classifiers that were used to evaluate the selected genes were : SVM, KNN, naive Bayes and decision tree. In the experiments the original partition of the datasets into training and test sets was used whenever information about the data split was available. Otherwise four-fold cross validation was applied. They concluded that the SVM had the best performance in all the datasets. The KNN classifier gave reasonably good performance on most of

the datasets which means it is not problem-dependent. Other interesting discussions of their report were that it was difficult to choose the best feature selection method, and the way that feature selection and classification methods interacted seemed very complicated. Due to the separation of the gene selection part from the classification part, there is no learning mechanism to learn how those component interact with each other.

Li, Zhang and Ogihara[22] has done a comparative study of eight feature selection techniques and seven classification techniques on nine different gene expression microarray data sets. The feature selection methods are information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, t-statistics, and one dimensional SVM. The classifiers that are used to evaluate the selected genes are J4.8, Naïve Bayes, kNN, SVM1 vs all, SVM Random, SVM Exhaustive and SVM Pair wise. They have used four fold cross validation to compare the accuracies achieved using the above classifiers. At first they calculated accuracies on all data sets without applying any feature selection technique. In this case SVM gave best result. Various numbers of genes [50,100,150,200,250] are selected and their performances are compared. Finally top 150 genes are selected by used gene selection techniques. SVM is resulted as the overall best classifier. kNN gave good performance on most of the data sets. There was not any clear view of best gene selection technique. None of gene selection method gave best result on all the datasets. Information gain has the superb performance on the ALL dataset; max minority performs the best on the SRBCT dataset; and sum of variances, t-statistic, and one-SVM achieve the best result on the MLL-leukemia and Lymphoma datasets. Overall, the methods sum of variances, t-statistics, and one dimensional SVM appear to have similar performance.

The work done by Heba Abusamra [23] included the study of feature selection and classification methods on two data sets (Freije and Phillips). He has used eight feature selection methods and three classification methods. The feature selection methods that he used are information gain, twoing rule, sum minority, max minority, gini index, sum of variances, one-dimensional SVM, and t-statistics. And the classifiers that are used to measure the performance of above gene selection methods are Support Vector Machines, k- Nearest Neighbors and Random Forest used 5-fold cross validation on the test set. For finding the required number of genes he selected various numbers of genes [20, 50, 100, 150, 200, 250] using eight gene selection methods and their performances are compared using above three classification methods. The top 20 genes that have the highest score were selected because it performed well, consumed less time, and required less memory configurations comparing to others. With only twenty features selected from Freije by information gain, gini index, sum of minority methods, or 1-dimentional SVM, SVM maintained the same accuracy as that by using all features. Gini index and 1-dimentional SVM achieved the best accuracy in 3-NN. For random forest, the accuracy performance was improved after applying feature selection mostly in all cases. For the Phillips data set, almost in all cases, the accuracy performance of SVM was improved after applying feature selections. With only twenty features, SVM had the highest accuracy with t statistics and gini index. Almost for the other feature selection methods it maintained the same accuracy as that by using all features. The performance of random forest had not improved much after performing feature selection.

Xing, Jordan and Karp [24] presented a comparative study of three feature selection methods and three classification methods for high dimensional microarray data. They used Leukemia data set for study and compare the results. Feature

selection methods those were used are Unconditional Mixture Modeling, Information Gain Ranking and Markov Blanket filtering. They used Gaussian Classifier, Logistics regression and k-Nearest neighbors for evaluating the performances. They used above classifiers on 2 to 100 number of genes on test set. They found that kNN(k=3) gave lowest error with 40 genes, Gaussian classifier and Logistic regression gave lowest error with 10 genes. kNN and Logistic classifier gave same results and the highest accuracy. Almost same results are obtained, when Leave-one-out cross validation is used to calculate the accuracies.

Mundra and Rajapakse [25] has done a comparative analysis on three feature selection methods on four data sets. They used MRMR, SVM-RFE and SVM-RFE with MRMR methods for feature selection. These methods are applied on four data sets i.e Colon Tumor, Leukemia, Hepato and Prostate. Different number of genes are selected using those methods for evaluation from each data sets. SVM-RFE with MRMR selected less number of genes from all the some data sets. SVM-RFE with MRMR gave best results in case of relevancy and accuracy. MRMR gave better results than SVM-RFE for all data sets.

Jayger, Sengupta and Ruzzo [26] compared classification done with five different test statistics: Fisher, Golub, Wilcoxon, TNoM, and t-test on three different publicly available datasets, Golub, Notterman and Alon. For the evaluation of the above gene selection methods, they have used LOO-CV and SVM. At first they calculated the accuracy with all genes and compared with the resultant accuracy after applying gene selection methods. There was no clear winner between the methods and it depends largely on the dataset and parameters used. All the proposed feature selection methods find a subset that has better LOOCV performance than the currently used approaches.

Karaboga and Akay [27] has done a comprehensive comparative study on the performances of well-known evolutionary and swarm-based algorithms for optimizing a very large set of numerical functions is presented. They have compared Evolution strategies, genetic algorithm, differential evolution algorithm, particle swarm optimization algorithm. From the results obtained in this work, it can be concluded that the performance of ABC algorithm is better than or similar to that of these algorithms although it uses less control parameters and it can be efficiently used for solving multimodal and multidimensional optimization problems.

In [28]Alshaman, Badr and Alohali proposed an innovative feature selection algorithm, minimum redundancy maximum relevance (mRMR), and combine it with an ABC algorithm, mRMR-ABC, to select informative genes from microarray profile. They evaluated the performance of the proposed mRMR-ABC algorithm by conducting extensive experiments on six binary and multiclass gene expression microarray datasets which are Colon, Leukemia1, Lung, SRBCT, Lymphoma and Leukemia2. They compared the proposed mRMR-ABC algorithm with various other gene selection techniques like Artificial Bee Colony(ABC), GA etc. They reimplemented two of mRMR techniques for the sake of a fair comparison using the same parameters. These two techniques are mRMR when combined with a genetic algorithm (mRMRGA) and mRMR when combined with a particle swarm optimization algorithm (mRMR-PSO).They have used Support Vector Machine (SVM) for calculating the accuracy. Different numbers of genes are used for different gene selection methods and data sets. The experimental results proved that the proposed mRMR-ABC algorithm achieves accurate classification performance using small number of predictive genes when tested using both datasets and compared to various gene selection methods.

Huawen, Lei and Huijie[4] compared EGSG (Ensemble Gene Selection by Grouping) with three other gene selection methods FCBF, mRMR and ECRP. They had applied these gene selection techniques on five data sets Breast Cancer, CNS(Central Nervous System), Colon Cancer, Leukemia and Prostate. They used two well known classification methods Naïve Bayes and kNN to evaluate the performance of these gene selection methods. They had done LOO-CV, 5-fold and 10-fold cross validation on above datasets. They found that their proposed method EGSG is comparable and effective. It not only leaded to better classification accuracy but also show higher stability. One of main disadvantage of EGSG is found, it selects more number of genes than mRMR.

Chanho and Sung-Bae[5] had done a comparative analysis on seven gene selection methods and six classifiers using Lymphoma and Colon Cancer data. The gene selection methods, they used are Pearson correlation coefficient , Spearman correlation coefficient , Euclidean Distance, Cosine coefficient, Information Gain, Mutual Information and Signal to noise ratio. To evaluate the performances, Multi-layer perceptron (MLP), kNN, SVM, Structure Adaptive Self Organizing Map (SASOM) and Genetic algorithm to search Optimal Esemble. They found that IG yields good performance in gene selection and kNN(cosine) yields superior performance in classification.

Ji-Gang and Hong-Wen [14] proposed a gene selection method based on Bayes error. The proposed method is Based Bayes error Filter (BBF), to select relevant genes and remove redundant genes of microarray data. They had used five publicly available datasets Colon, DSLBCL, Leukemia, Prostate and Lymphoma for evaluation. They had used kNN and SVM for measuring the accuracies. They found that BBF method can effectively perform gene selection with low classification error

rates. They also found that this method selects very less number of genes as compare to others.

In [16], Xing, Jordan and Karp showed that is high dimensional data, feature selection methods are essential if the analyzer want to analyze their data. They had studied a generative Gaussian classifier, regression classifier and kNN and they found that all of them performed better in reduced feature space than the full feature space. They have used three feature selection techniques namely Unconditional Mixture Modeling, Information Gain Ranking and Markov Blanket Filtering. kNN gave the best accuracy.

Chapter 3

Methodology

3.1 Data Normalization

The gene expression data generated from the DNA microarray technology cannot be directly used for analysis because there might be inconsistencies between the scales of measurements in different conditions or samples. Different factors may produce these inconsistencies. For instance, different equipment has been used to prepare and process microarray gene expression data sets. Such data are collected at different times or with the use of different methodologies for preprocessing the individual arrays [29]. Normalization is a necessary step to effectively analyze gene expression data, as well as to reduce unwanted variation in expression measurements and allow data to be comparable in order to find actual changes. There are a number of widely used normalization methods, such as z-score, min-max, and quantile normalization [30].

Min-max normalization rescales the features or outputs from one range of values to a new range of values lie in the closed interval from 0 to 1. The rescaling is often accomplished by using a linear interpretation formula, as in Equation (I).

$$Min - Max(x) = \frac{x - Min_x}{Max_x - Min_x} \quad \text{(I)}$$

Where Min_x and Max_x are the minimum and maximum value of x, respectively.

3.2 Used Gene Selection Methods

As mentioned earlier, the high dimensionality of gene expression data remains challenging. Among thousands of genes whose expression levels are measured, some genes are not informative for classification. Thus, there is a need to select some genes that are highly related to particular classes for classification, known as informative genes. This process is known as gene selection, which is also referred to as feature selection in machine learning. Many gene selection algorithms have been applied in gene expression data analysis. Some of the most widely applied methods in literature are briefly described as follows [19, 31].

T-test[32]is one of the most popularly used filter methods for feature selection. This method measures the statistical significance of a difference of a particular feature between the two classes. A t-Test is a statistic that checks if two mean of genes are reliably different from each other. It is inferential statistic that allows us to make inferences about the population beyond our data. A big t-value means different groups and a small t-value means similar groups. Each t-value has a corresponding p-value. The p-value is the probability that the pattern of data in the sample could be produced by random data. From experiment it is found that p-value should be <=0.5 i.e less than or equal to 5% chance then there is no real difference. Genes with the largest t-statistics are then selected. T-statistic is computed using Equation (II).

$$t - test = \frac{|\mu_i^+ - \mu_i^-|}{\frac{(\sigma_i^+)^2}{n_+} + \frac{(\sigma_i^-)^2}{n_-}} \quad \text{(II)}$$

Chi-squared test[20]is also a popular filter method that can be used for gene selection. The value of $\chi 2$-Statistic is computed for each gene individually with respect to the classes. Similar to Information Gain, eachnumeric gene is discretized before computing $\chi 2$-Statistic. For each gene X_i, $\chi 2$-Statistic is defined as

$$\chi^2 = \sum_{x \in X_i} \sum_{c \in C} \frac{\left(n_{(x \in X_i \,\&\, c \in C)} - e_{(x \in X_i \,\&\, c \in C)}\right)^2}{e_{(x \in X_i \,\&\, c \in C)}} \qquad \text{(III)}$$

where $n_{(x \in X_i \& c \in C)}$is the number of samples or patients in Xi for class c whose value is x. The expected frequency$e_{(x \in X_i \& c \in C)}$is defined as

$$e_{(x \in X_i \,\&\, c \in C)} = \frac{n_{x \in X_i} * n_{c \in C}}{n} \qquad \text{(IV)}$$

where $n_{x \in Xi}$ denotes the number of samples in X_i with value x and$n_{c \in C}$ represents the number of samples of class c. n is the total numberof samples. The genes are selected based on the sorted valuesof $\chi 2$-Statistic for all features.

Information gain is a univariate filter method that has been extensively used in microarray data to identify informative genes, and it has been reported to be the superior gene selection technique by [24, 33]. This method computes how well a given feature separates the training samples according to their class labels. It measures the expected reduction in entropy caused by splitting the samples according to a particular feature. For feature *F* and samples *S* the information gain is computed by

$$I(F,S) = \sum_{v=values(F)} \frac{|S_v|}{|S|}(Entropy(S) - Entropy(S_v)) \quad \text{(V)}$$

$$\text{Where} \quad Entropy(S) = \sum_{i=0}^{k} -p_i \log p_i$$

Relief-F[34] is an uncomplicated and efficient method used as a pre-processing feature (gene) subset selection method, to assess the quality of the genes that have very high dependencies between the genes [35]. Relief-F is capable of dealing with multiclass datasets and is an efficient method to deal with noisy and incomplete datasets. It can be used to estimate the quality and identify the existence of conditional dependencies between attributes effectively.The main concept of the Relief-F, is to assess the quality of genes based on their values to differentiate among instances that are close with each other. Given a randomly chosen instance Ins_m from class L, the Relief-F searches for K of its nearest neighbours, from the same class known as nearest hits H, and also K nearest neighbours from each of the different classes, called nearest misses M. Later it updates the quality estimation Wi for gene i depending on their values for Ins_m, H, M. If the instance Ins_m and the others in H have dissimilar values on gene i, then the quality estimation W_i is reduced. In contrast, if instance Ins_m and those in M have dissimilar values on the gene i, then W_i is increased. The entire process is iterated n times, which is set by users. To update W_i the Equation1is used as follows:

$$W_i = W_i - \frac{\sum_{k=1}^{K} D_{Hk}}{nK} + \sum_{c=1}^{C} P_c \sum_{k=1}^{K} \frac{D_{MCk}}{n_c K} \quad \text{(VI)}$$

Where, n_c = # of instances of class c

P_c = Probability of class c

Support Vector Machine Recursive Feature Elimination (SVM-RFE) was proposed in [36] to do gene selection for cancer classification. Nested features are

selected based on backward elimination manner, which starts with all the features and remove the features one by one that has smallest ranking criteria. At each step the weight vector w is generally considered as the ranking criteria. Each step of SVM-RFE procedure is as follows

1. Train the classifier
2. Compute the ranking criteria
3. Remove the feature with smallest rank

Maximum Relevance Minimum Redundancy (mRMR) was introduced in [35]. The mRMR filter method selects genes with the highest relevance and minimally redundant with the target class [37]. In mRMR, the Maximum Relevance and Minimum Redundancy of genes are based on mutual information[38]. Given g_i, which represents gene i and c represents the class label, the mutual information of g_i and c is defined in terms of their probability frequencies of appearances $P(g_i)$, P(c), and $P(g_i,c)$ as follows:

$$I(g_i, c) = \sum\sum p(g_i, c) * \ln \frac{p(g_i,c)}{p(g_i)p(c)} \quad \text{(VII)}$$

The Maximum Relevance method selects the highest top m genes, which have the highest relevance correlated to the class labels from the descent arranged set of $I(g_i,c)$.

$$\max\left(\frac{1}{|S|}\right) \sum_{g_i \in S} I(g_i, c) \quad \text{(VIII)}$$

However, it is well known to the researchers that "the m best features are not the best m features", because the correlations among those top genes might also be high [14*]. Therefore, Minimum-Redundancy criterion is introduced by [14*] in order to remove the redundancy features. The following is the Minimum Redundancy criterion:

$$\min_{S}\left(\frac{1}{S^2}\right)\sum_{g_i,g_j} I(g_i, g_j) \tag{IX}$$

Equation (3) shows that the mutual information between each pair of genes is taken into consideration. The (MRMR) filter combines both optimization criteria of Eqs. (X and XI).

A sequential incremental algorithm to solve the simultaneous optimizations of optimization criteria of Eq. (2 and 3) is explained as follows: Assume, we have G which represent a set of genes and also have S_m-1, the gene set with m−1 genes, and then the task is to choose the m−th gene from the set $\{G-S_m-1\}$. This feature is chosen by increasing the single-variable relevance minus redundancy function.

$$\max_{g_i \in G-S_{m-1}} [I(g_i; c) - \frac{1}{m-1}\sum_{g_i \in S_{m-1}} I(g_j, g_i)] \tag{X}$$

The $m-th$ features also can be chosen by maximizing the single variable relevance divided by redundancy function

$$\max_{g_i \in G-S_{m-1}} [\frac{I(g_i;c)}{\frac{1}{m-1}\sum_{g_i \in S_{m-1}} I(g_j,g_i)}] \tag{XI}$$

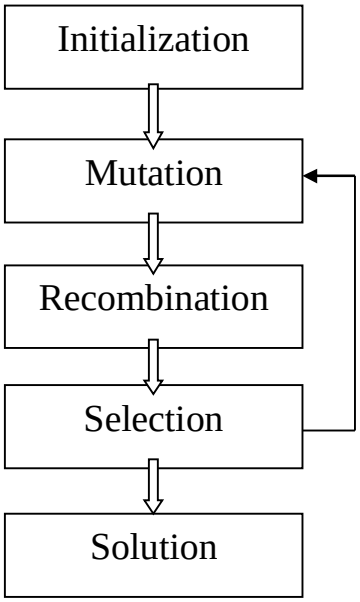
Genetic Algorithm(GA) [39] is an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomized, GA is by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space.

Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solution (offspring) are selected according to their fitness – the more suitable they are the more chances they have to reproduce. This is repeated until some condition is satisfied.

Outline of Genetic Algorithm

1) [Start] Generate random population of *n* chromosomes (suitable solutions for the problem)
2) [Fitness] Evaluate the fitness *f(x)* of each chromosome *x* in the population
3) [New population] Create a new population by repeating following steps until the new population is complete
    a) [Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
    b) [Crossover] With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
    c) [Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome).
    d) [Accepting] Place new offspring in a new population
4) [Replace] Use new generated population for a further run of algorithm
5) [Test] If the end condition is satisfied, stop, and return the best solution in current population
6) [Loop] Go to step 2

Differential Evaluation is introduced by Storn and Price [40]. It is an evolutionary algorithm like Genetic Algorithm. It is slight modification version of Genetic Algorithm.GA works on bits or characters but DE works on vectors. It has all the steps of GA like Initialization, Selection, Crossover and Mutation. But the method of doing these steps are different as discussed below. The principal difference between Genetic Algorithms and Differential Evolution is that Genetic Algorithms rely on crossover, a mechanism of probabilistic and useful exchange of information among solutions to locate better solutions, while evolutionary strategies use mutation as the primary search mechanism.



Here Recombination is nothing but crossover.

Outline of Differential Evaluation :-

1) For each candidate ($X_{1...NP-1}$) in the population.
2) Choose three distinct parents at random (they must differ from each other and i)
3) Perform Mutation

Add a weighted difference vector between two population members to a third member

$X^g_c` = X^g_c + F * (X^g_b - X^g_a)$

4) Perform Crossover (Recombination) :-For every variable j in X_i

$Xt_{i,j} = X^g_c`$ if rand < CR or n = rand(D) + 1

$X^g_{i,j}$ otherwise

Where CR = Crossover Rate and rand(D) generates random integer in the range of 1..D-1

5) Perform Selection

$X^{g+1}_i = Xt_i$ if $f(Xt_i) > f(X^g_i)$

X^g_i otherwise

3.3 Used Classification Methods

Different classification algorithms in machine learning have been applied to predict and classify different tumor types using gene expression data in recent research. Classification is the problem of identifying to which different category a new observation belongs. If there are two categories or classes, then the classification is known as the binary classification problem; and if there are three or more classes, then it is called the multi-classification problem. The main process of classification in machine learning is to train classifier to accurately recognize patterns from given

training samples and to classify test samples with the trained classifier. Building a classifier that is as accurate as possible in classifying new samples is challenging for several reasons. If the training set is relatively small, then it is less likely to capture the underlying distribution of the data. Another problem is the complexity of the model and its generalizing capabilities. If the classifier is too simple it may fail to capture the underlying structure of the data. However, if the classifier is too complex and there are too many free parameters, it may incorporate noise in the model, leading to over-fitting, where the learned model highly fits the training set, but performs poorly on test samples. As mention earlier, there are two main types of learning schemes in machine learning, i.e., supervised learning and unsupervised learning. Here we focus on supervised binary classification methods. We summarize below some of the most commonly used supervised learning algorithms. They are among the top ten supervised classification methods identified by the Institute of Electrical and Electronics Engineers (IEEE) [41].

Support vector machine (SVM) was introduced by Vapnik in [42], has been widely used as an efficient tool for classification and regression problems. It is based on simple but powerful idea, yet is a very popular classification method. SVM classifiers are generally binary based. If the data is linearly distributed then SVM computes the hyperplane that maximizes the margin between training samples and the class boundary. In contrast, if the data is not linearly distributed then the samples are mapped into a multi-dimensional space in which such a hyperplace can be constructed. This mapping process is generally called the kernel function. The formal basic of SVM is briefly explained below. Starting with the simplest type of SVM, which is linear classification. In linearly separable cases, a separating hyperplane H is considered to be the best if its margin is maximized. The margin is the distance between two parallel hyperplanes, each of which sits on the support vectors of one category. To demonstrate this idea let us consider Figure 2.1. For the

same training set, different separating hyperplanes can be found, but the aim is to find the optimal hyperplane that does not only separate training samples of one class from the other, but also separates the testing samples as accurately as possible.

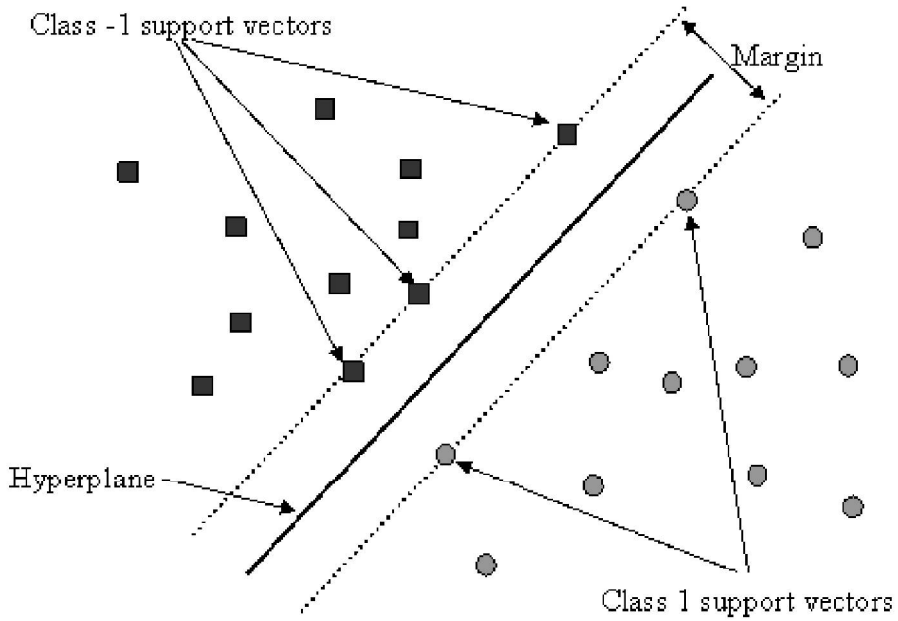


Figure 3.1: Illustration of separating hyperplane and margins.
(Image Courtesy :http://ivplab.cs.thu.edu.tw/research/mi/03/index.htm.)

General hyperplane equation can be expressed as $w`x_i + b = 0$where *x* is the vector of a data point, *w* is a weight coefficient vector, and *b* is a bias term. The hyperplane should separate the data, so that $w`x_i + b \geq 1$for all $x_i$the of one class, and $w`x_i + b \leq 1$ for all the $x_j$of the other class. However, in many practical problems there is no linear boundary separating the classes. In such cases, the technique of "kernel" is

used to map the training samples from the original space into a higher dimensional space, and to learn a separator in that space.

SVM classifier is widely used in bioinformatics due to its high accuracy, theoretical guarantees regarding over-fitting, ability to deal with high-dimensional data such as gene expression, and flexibility in modeling diverse sources of data [43, 44].

K-Nearest Neighbor (KNN) is one of the simplest but widely used machine learning algorithms. It was first introduced by Fix and Hodges[45] and it has been extensively studied and discussed in respect to classification. Generally in this algorithm "distance" is used to classify a new sample from its neighbors. The most common class among its *K* distance wise nearest neighbors will be assign to this new sample. Figure 2.2 helps to illustrate the algorithm in details.

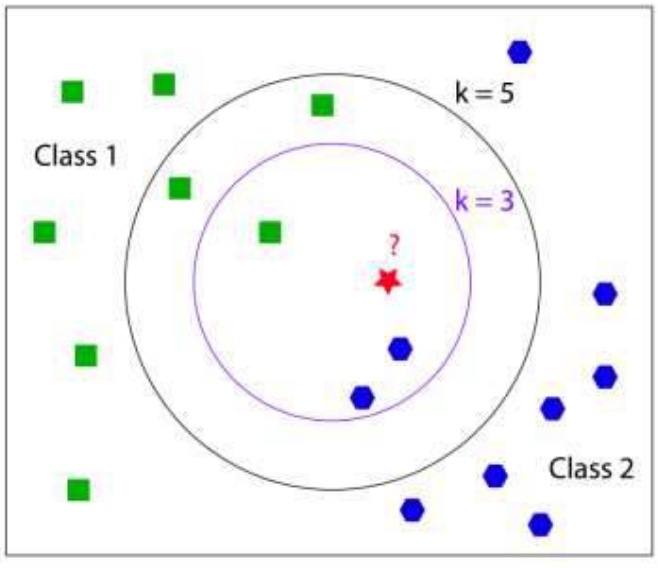


Figure 3.2: Illustration of the kNN algorithm.

(Image Courtesy
:http://141.61.102.17/perseus_doku/doku.php?id=perseus:activities:matrixprocessing:learning:classificationprocessing)

For a given data point, KNN assumes that all samples correspond to points in the *d*-dimensional space. For example, let the samples be points in the two-dimensional space and with "+" or "-" (positive and negative, respectively) class label as shown in the figure. To classify an unknown sample *x*, we measure the distance (using some distance metric, e.g. the Euclidean distance) from *x* to every other training sample. The K smallest distances are identified, then *x* is classified with class label that has the majority of its K neighbors. For instance, in Figure 2.2, if *K*=1 then *x* will be classified as Class 2. However, if *K*=3 and *K*=5 then *x* will be classified as Class 2 and Class 1 respectively. In case of tie, *x* is assigned randomly.

This example introduces the problem of choosing a proper *K* (the number of neighbors). If *K* is too small, then the result can be sensitive to noise. On the other hand, if *K* is too large, then the neighborhood may include too many points from other classes. The best choice of *K* depends upon the data. While there is no rule for choosing the best *K*, a common way is to try several possible values for *K* and select the one with the lowest error estimation. In binary classification problems, it is useful to choose *K* to be an odd number as this avoids tied votes.

KNN is considered a non-parametric lazy learning algorithm. Non-parametric technique means that it does not make any assumptions on the underlying data distribution. Lazy algorithm means that there is no explicit training phase. For high-dimensional datasets, dimension reduction is usually performed prior to applying KNN algorithm in order to avoid the effects of the curse of dimensionality [46, 47].

3.4 Accuracy Estimation Methods and Evaluation

Generally speaking, classification suffers in bioinformatics due to the "small samples in conjunction with large numbers of features" nature of bioinformatics data. When there are insufficient data to split the samples into training and testing sets, estimating and reporting classification accuracy becomes a problem. There are external methods for error estimation that utilize the original data without incurring substantial bias [48].

Hold-out is considered as the simplest portioning that splits data to two or more partitions as cross validation. This method splits data randomly to two unequally sized groups. The largest partition is used for training while the second is used for testing. This method is usually preferable and does not take time to compute. However, the holdout estimate of accuracy depending heavily on the split. This drawback appears clearly in case if we have sparse data.

Cross validation is a statistical method of evaluating and comparing learning algorithms by repeatedly partitioning the given data set into two disjoint subsets: the training and the test subset. The training subset is used to build the classifier, then the samples belong to the test subset is used to test the trained classifier. The process is repeated with several partitions and gives an estimate of the classification performance. The most common form of cross-validation is k-fold cross-validation [31].

- K-fold cross validation: The k-fold cross validation partitions the given data set into k equally sized subsets. Then, training is done on k-1 subsets and testing is done on the remaining subset. This process is repeated k times(folds) with each subset is taken to be a test set in turn.
- Leave One Out Cross Validation: In this method, we use k-fold cross validation where k is equal to number of samples in the data set. In

this method k-1 samples are used as training set and only one sample used for testing. This process is repeated for all the samples. This method is computationally expensive as it requires the construction of n different classifiers. However it is more suitable for smaller datasets.

There are several ways to measure the effectiveness of classification classifying unlabeled samples. First, I will introduce the confusion matrix. A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual class labels in the test data. The matrix is *k*-by-*k*, where *k* is the number of classes. In binary class problem, confusion matrix consists of four cells, True Positive (TP), False Positive (FP), False Negative(FN) and True Negative(TN) which derive all the other measures.

		Predicted Label	
		Positive	**Negative**
Actual Label	**Positive**	True Positive (TP)	False Negative(FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 3.1: The Confusion Matrix

Where *TP* is the number of positive samples that are classified as positive in the test set; *FP* is the number of negative samples that are classified as positive. *FN* is the number of positive samples that are classified as negative; and *TN* is the number of negative samples that are classified as negative. For each fold of cross validation, we

keep count these four values to construct the confusion matrix. After constructing the confusion matrix, different measures can be computed. Some of them are briefly described below.

Accuracy refers to the percentage of correct predictions made by the model, when compared with the actual class labels in the test set.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(XII)}$$

Precision (also called sensitivity, hit rate, and recall) is the proportion of predicted positive instances that are actually positive.

$$Precision = \frac{TP}{TP+FP} \quad \text{(XIII)}$$

Recall also called sensitivity is the proportion of positive instances that are correctly classified as positive.

$$Recall = \frac{TP}{TP+FN} \quad \text{(XIV)}$$

F-Score is a measure of test's accuracy.

$$F - Score = 2.\frac{Precision.Recall}{Precision+Recall} \quad \text{(XV)}$$

Chapter 4

Evaluation and Results

4.1 Data Sets

Two well known datasets are used for this research work which are

4.1.1 Colon Tumor Data: This is a collection of gene expression profiles of 62 samples. These samples are collected from colon-cancer patients. Every sample contains 2000 genes expressions. Among them 40 tumor biopsies are from tumors labeled as "negative" and 20 normal labeled as "positive" are from healthy parts of the colons of the same patients. I have divided the data in two parts Training data and Test data. Among 62 samples, I have used 42 samples as Training data and the remaining 20 samples as Test data. This data was also not normalized, so before using them, at first I have done min-max normalization on the given data. The raw gene data can be found at http://microarray.princeton.edu/oncology/affydata/index.html. And the processed data in the format of .data and .names can be found at http://datam.i2r.a-star.edu.sg/datasets/krbd/ColonTumor/ColonTumor.html.

4.1.2 Leukemia ALL-AML Data: This group of gene expression profiles contains 72 samples. These data are collected from Leukemia patients. This data contains two group of samples Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). This data was not normalized, so to scale the given data, I have performed min-max normalization on the data. Every sample contain 7129 genes expressions. Training data set contain 38 samples and the testing data set contain 34 samples.The raw gene data can be found at http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi. And the processed data in the format of .data and .names can be found at http://datam.i2r.a-star.edu.sg/datasets/krbd/Leukemia/ALLAML.html.

Dataset	# of Samples	Classes	Samples in each class	# of Genes
Colon Tumor	62	negative	40	2000
		positive	22	
Leukemia Data	72	ALL	47	7129
		AML	25	

Table 4.1: Summary of two microarray datasets

4.2 Evaluation:

Hold Out and LOO-CV are used. In this experiment, feature selection is applied to the training set and the selected features are tested on the testing set. In this research, R and WEKA(Waikato Environment of Knowledge Analysis) tool are used. The calculated results of Hold Out and LOO-CV using R and WEKA respectivelyare reported.

4.3Experimental Results:

The classification methods were first applied to both datasets without performing any feature selection. Results of the Hold-out using R are shown in Figure 4.1 and Table 4.3. And results of the Leave One Out Cross Validation (LOO-CV) are shown in Figure 4.2 and Table 4.4. kNN performed better than SVM on both the datasets without any gene selection. In case of Colon Data, accuracy achieved by SVM and kNN (k=6) on Hold out are 65% and 75% respectively whereas in case of Leukemia data, accuracy achieved by SVM and kNN (k=5) are 58.8% and 67.6% respectively. Both the classifiers on Hold out gave very less accuracy, which is not acceptable for diagnosis purpose. If we talk about LOO-CV accuracy then on colon data, accuraciesare overall similar but on leukemia dataset kNN gave very high accuracy. On Colon Dataset, SVM and kNN (k=6) gave 63% and 72.5% accuracy, whereas on Leukemia Dataset, SVM gave 65.3% accuracy and kNN (k=5) gave 83.3% accuracy.

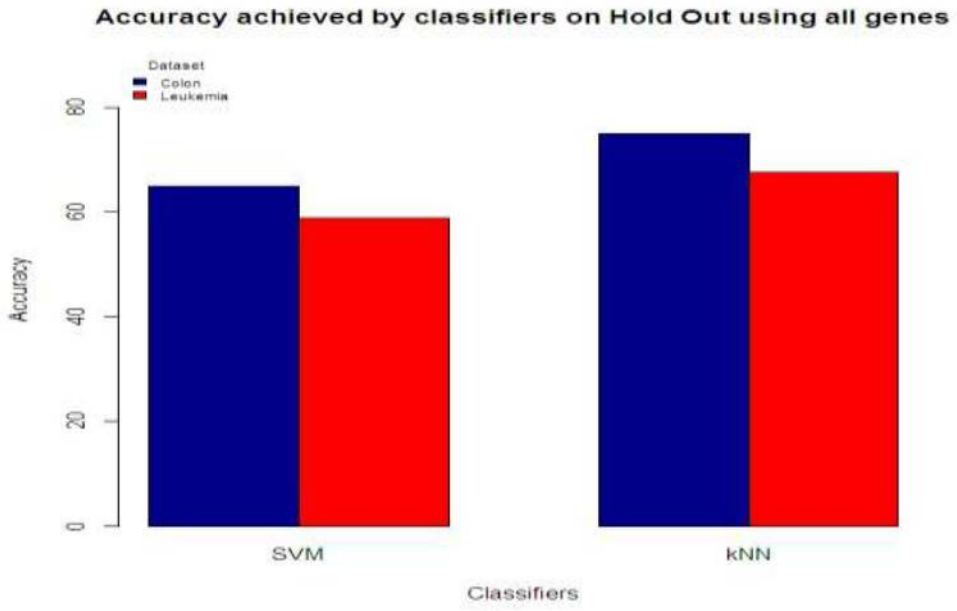


Figure 4.1: Hold out accuracy achieved by classification methods without performing gene selection.

	SVM				kNN			
	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy
Colon	0.325	0.5	0.4	65	0.86	0.64	0.74	75
Leukemia	0.294	0.5	0.37	58.8	0.73	0.62	0.67	67.65

Table 4.2: Hold Out accuracy of the classification methods without any gene selection.

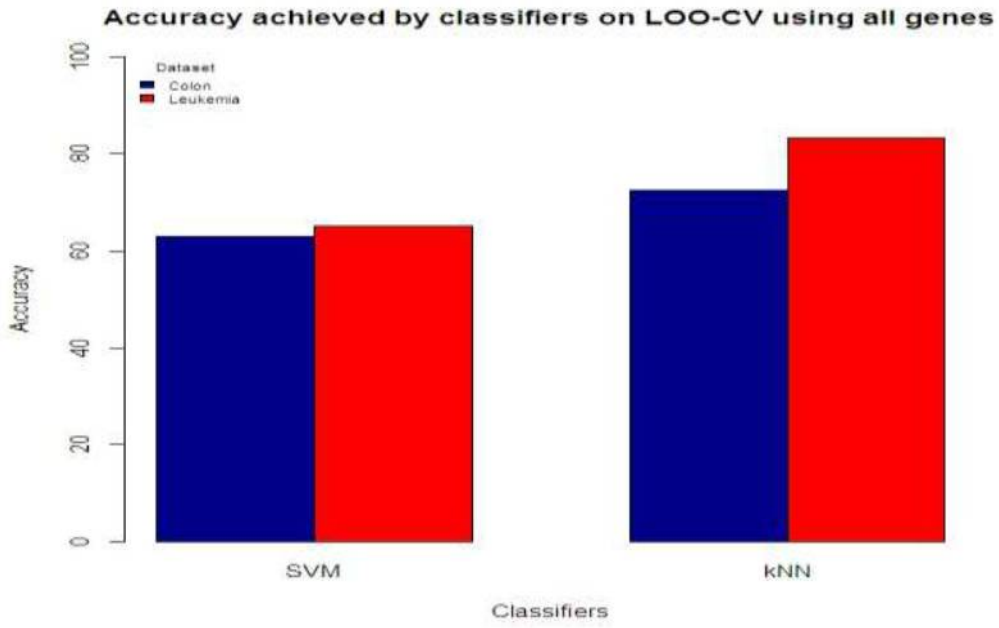


Figure 4.2: LOOCV accuracy achieved by classification methods without performing gene selection.

	SVM				kNN			
	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy
Colon	0.325	0.5	0.4	65	0.86	0.64	0.74	75
Leukemia	0.326	0.5	0.4	65.27	0.87	0.77	0.81	83

Table 4.3: LOOCV accuracy of the classification methods without any gene selection.

In next experiment, we applied the classification methods to both data sets after selecting genes using various gene selection methods.

Results on the Colon Dataset:

To find out the optimal number of genes, we have selected different number of genes 10, 12, 14,.., 26, 28, 30, 50 using 8 gene selection methodsusing R. After analyzing the achieved accuracies, we have found out that none of the number of genes gave best result for all the gene selection methods. So we have selected top 50 genes for calculating the accuracies using SVM and kNN. By performing gene selection, accuracies are highly improved on Colon Data in almost all cases. The accuracy achieved by top 10 to 30 using SVM and kNN are summarized in the Figure 4.3 and Figure 4.4 respectively.

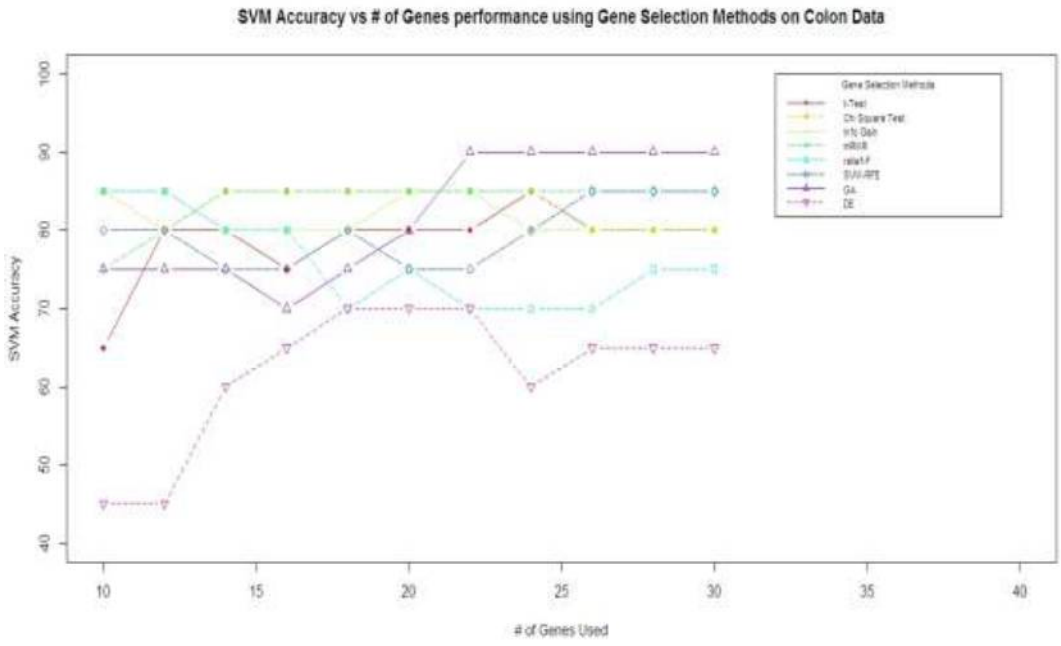


Figure 4.3: Hold Out accuracy achieved by gene selection with different number of genes using SVM on Colon data.

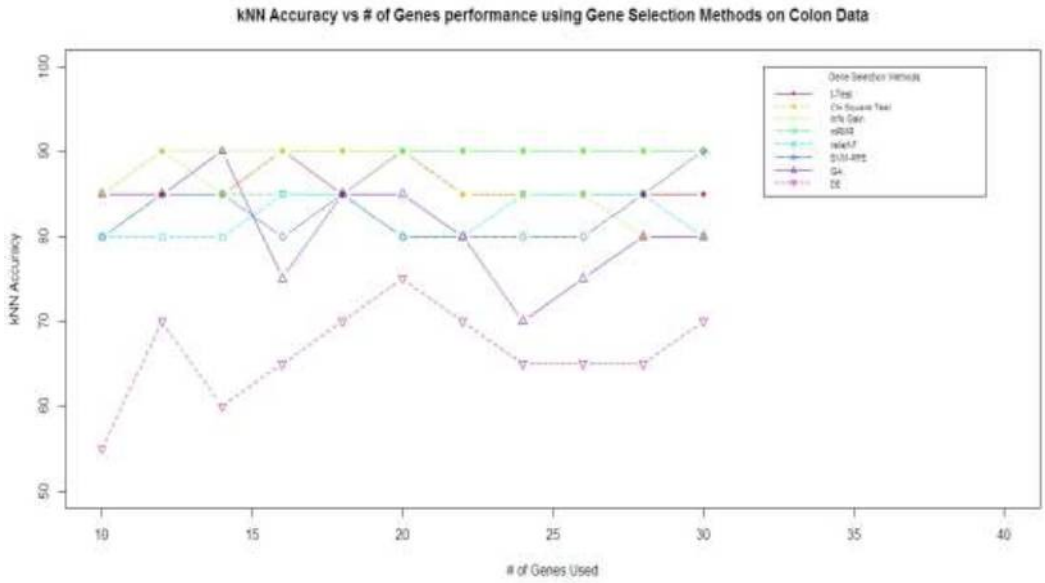


Figure 4.4: Hold Out accuracy achieved by gene selection with different number of genes using kNN (k=6) on Colon data.

We have selected top 50 genes using different gene selection methods in R. With different values of k, k=6 gave the overall best result on Colon Data, so we have selected k=6 for the experiment. The hold out performance of SVM and kNN (k=6) vary in different gene selection methods. With only 50 genes selected by t-Test and relief-F, SVM achieved the highest accuracy of 85%. Using genes selected by DE, SVM achieved the lowest accuracy of 70% whereas for 50 selected by other gene selection methods SVM achieved 80% accuracy. Relief-F achieved the best accuracy in kNN. Using genes selected by GA and DE, kNN gave the lowest accuracy of 80%. Average accuracy achieved by SVM and kNN using gene selected by all the gene selection methods are 80% and 87.5%. By averaging the accuracy achieved by SVM and kNN for all the used gene selection methods, Relief-F gave the best average accuracy of 90%. By analyzing the results, we can say that kNN gave the best result in case of classification methods and Relief-F gave the best result in case of used gene selection methods.

Figure 4.5 and Table 4.3 summarizes the hold out performance of gene selection using top 50 genes and classification methods achieved using R.

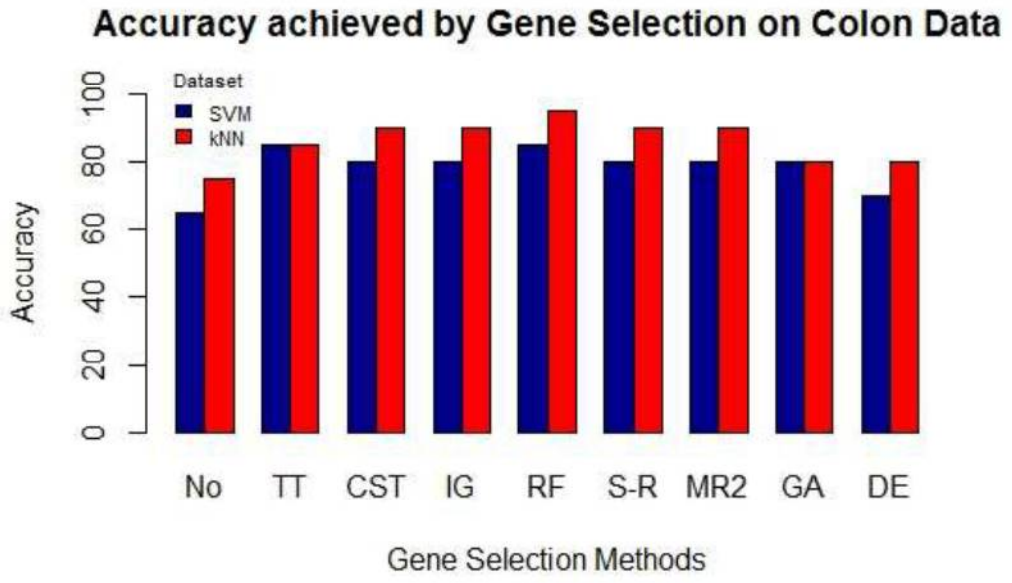


Figure 4.5: Hold Out accuracy achieved by gene selection using top 50 genes and classification methods on Colon data. No(without feature selection), TT(t-Test), CST(Chi Square Test), IG(Information Gain), RF(Relief-F), S-R(SVM-RFE), MR2(mRMR), GA(Genetic Algorithm) and DE(Differential Evaluation).

Gene Selection Methods	Classification Methods	
	SVM Accuracy	kNN Accuracy
t-Test	85	85
Chi Square Test	80	90
Information Gain	80	90
Relief-F	85	95
SVM-RFE	80	90
mRMR	80	90
GA	80	80
DE	70	80

Table 4.4: Hold Out accuracy achieved by gene selection using top 50 genes and classification methods on Colon data.

LOOCV accuracy is also calculated. For that, WEKA tool is used. WEKA doesn't contain all the gene selection methods that are used in this research, only those methods which are present in WEKA are compared using LOOCV. Gene selection methods compared using WEKA are Chi Square Test, Information Gain, Relief-F,SVM-RFE, mRMR and GA. Using these methods on Colon Tumor Data, top 50 genes are selected except GA. GA has selected 764 genes from Colon Data. All the gene selection methods are evaluated on the basis oftheir accuracies calculated usingSMO (Sequential Minimal Optimization) SVM and kNN (k=6). In WEKA, SVM-RFE with kNN gave the best accuracy of 88.7%. Figure 4.6 and Table 4.4 summarizes the performance of LOOCV of gene selection methods and classification methods achieved using WEKA.

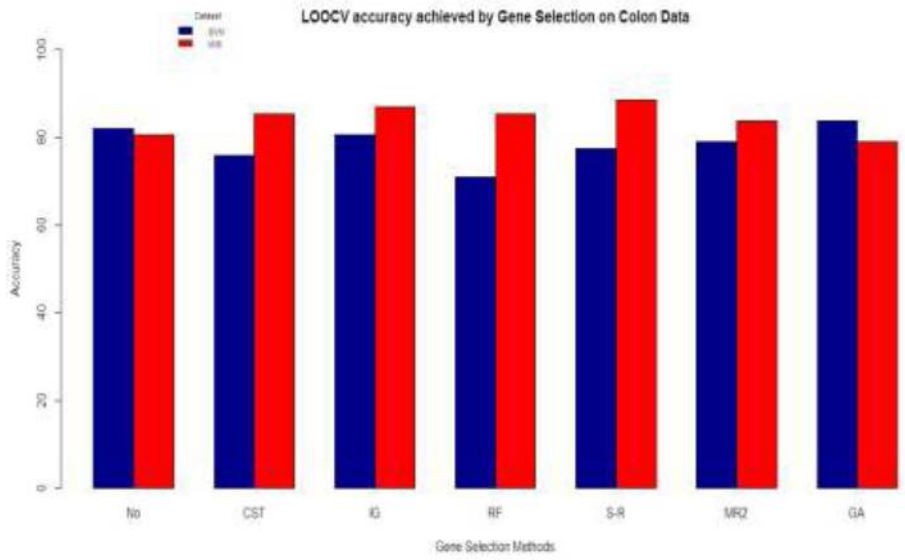


Figure 4.6: LOOCV accuracy achieved by gene selection using optimal # of genes and classification methods on Colon data. No(without feature selection), CST(Chi Square Test), IG(Information Gain), RF(Relief-F), S-R(SVM-RFE), MR2(mRMR) and GA(Genetic Algorithm).

Gene Selection Methods	Classification Methods	
	SVM Accuracy	kNN Accuracy
Chi Square Test	75.80	85.50
Information Gain	80.64	87.00
Relief-F	71.00	85.50
SVM-RFE	77.42	88.70
mRMR	79.00	83.87
GA	83.87	79.00

Table 4.5: LOOCV accuracy achieved by gene selection using optimal # of genes and classification methods on Colon data.

Results on Leukemia Data Set:

For the Leukemia Dataset also, we used different number of genes 10,12,14,…,26,28,30 and 50 to find out the optimal number of genes.Here also top 50 genes gave overall best performance. Almost in all cases, the accuracy performance of SVM and kNN (k=5) were improved after applying genes selection. We have used RBF kernel SVM. RBF kernel SVM uses two parameters gamma and cost. We calculated the accuracies achieved using different values of gamma and cost. RBF kernel SVM gave best accuracy with gamma=0.02 and cost=100. And for kNN, there is only parameter that had to be tuned and that is k. So we calculated the accuracies with different values of k and k=5 is selected after comparing the achieved accuracies. The accuracy achieved by top 10 to 30 using SVM and kNN are summarized in the Figure 4.6 and Figure 4.7 respectively.

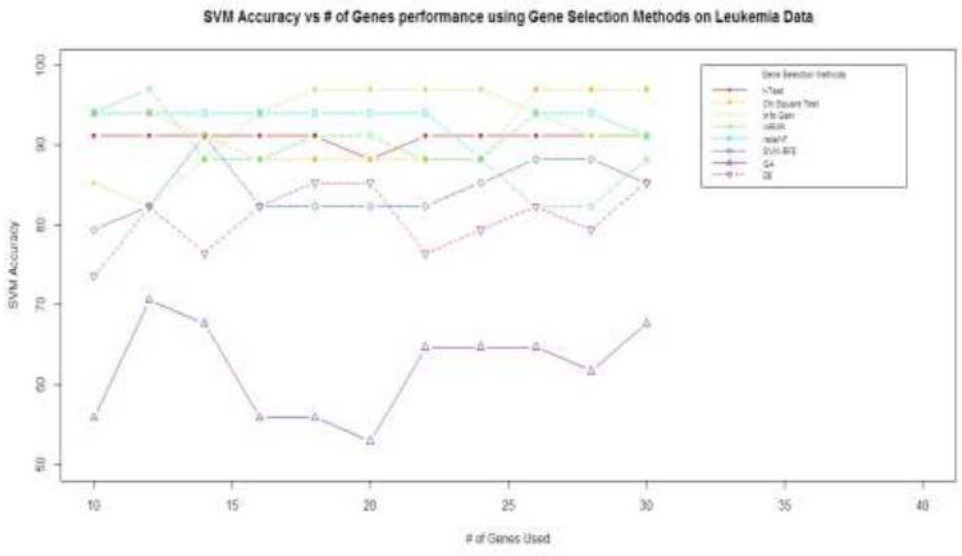


Figure 4.7: Hold Out accuracy achieved by gene selection with different number of genes using SVM on Leukemia data.

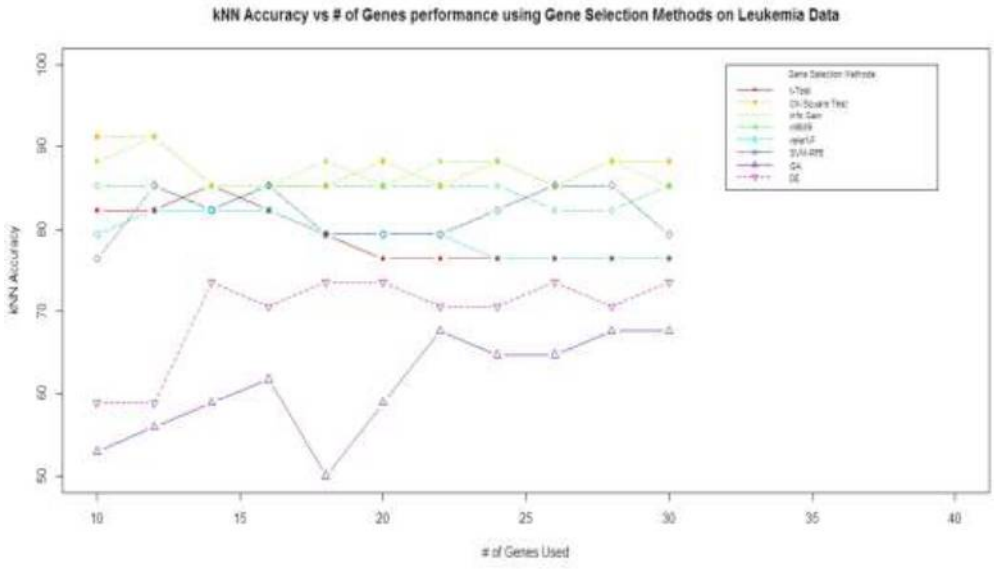


Figure 4.8: Hold Out accuracy achieved by gene selection with different number of genes using kNN (k=5) on Leukemia data.

In case of R, with only top 50 features, SVM had the highest accuracy of 94% using Information Gain and Minimum Redundancy and Maximum Relevance. For Genetic Algorithm and Differential Evaluation, it gave very low accuracy of 61% and 76.5% respectively. The accuracy of GA has not improved much. Almost for the other gene selection methods, it maintains the same accuracy of 91%. kNN (k=5) has not given very good results as compare to SVM. The best accuracy achieved by kNN is 85% using genes selected by Chi Square Test, Information Gain and SVM-RFE. For most of the other gene selection methods, it gave accuracy of 80%. As like SVM, it gave low accuracy for GA and DE methods, which are 70.5% and 73.5%. But in case of GA, it gave better result as compare to SVM. Out of SVM and kNN, SVM gave the best average accuracy of 86% and among all gene selection methods, Information Gain gave the best average accuracy of 89.5. Chi Square Test & SVM-RFE gave second highest average accuracy of 88%. Looking at the average accuracies of the classification methods and gene selection methods, we can say SVM and Information Gain is the winner according to achieved accuracy.

Figure 4.8 and Table 4.5 summarizes the performance of gene selection using top 50 genes and classification methods.

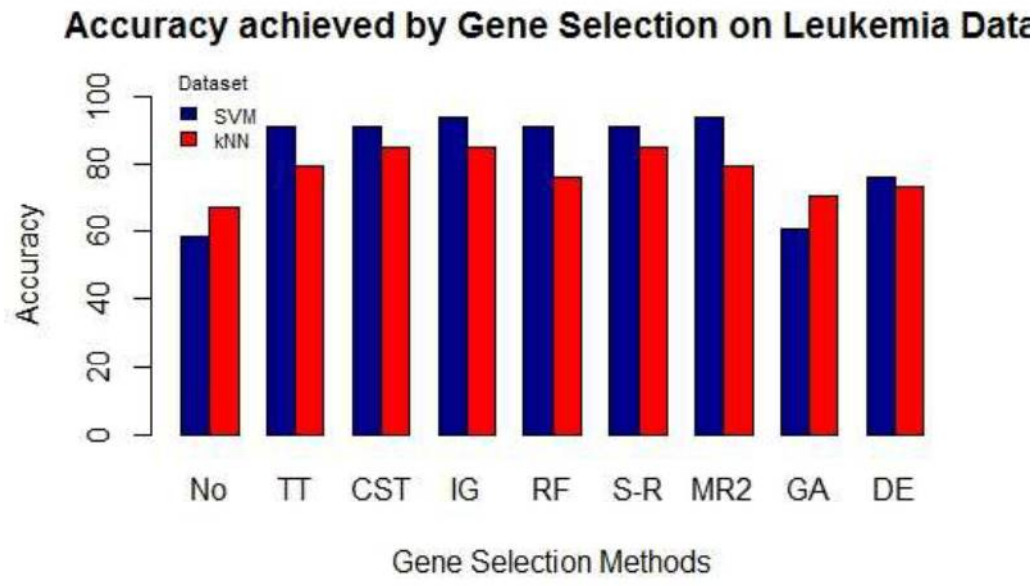


Figure 4.9: Hold Out accuracy achieved by gene selection using top 50 genes and classification methods on Leukemia data. No(without feature selection), TT(t-Test), CST(Chi Square Test), IG(Information Gain), RF(Relief-F), S-R(SVM-RFE), MR2(mRMR), GA(Genetic Algorithm) and DE(Differential Evaluation).

Gene Selection Methods	Classification Methods	
	SVM Accuracy	kNN Accuracy
t-Test	91.00	79.50
Chi Square Test	91.00	85.00
Information Gain	94.00	85.00
Relief-F	91.00	76.50
SVM-RFE	91.00	85.00
mRMR	94.00	79.50
GA	61.00	70.50
DE	76.50	73.50

Table 4.6: Hold Out accuracy achieved by gene selection using top 50 genes and classification methods on Leukemia Data.

For Leukemia dataset, LOOCV is also used using WEKA. SVM-RFE with kNN and GA with SMO SVM performed very well.They gave very high accuracy of 98.61%. Without Gene selection SMO SVMalso gave 98.615 accuracy. GA selected 1287 genes. GA with kNN gave the worst accuracy of 84.72%. The calculated accuracies are summarized in Figure 4.10.

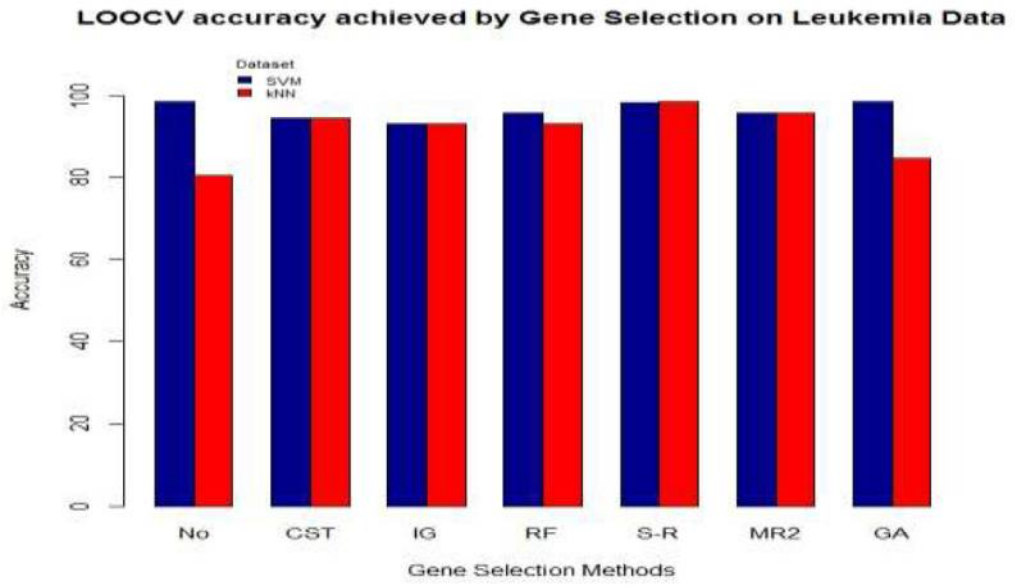


Figure 4.10: LOOCV accuracy achieved by gene selection using optimal # of genes and classification methods on Leukemia data. No(without feature selection), CST(Chi Square Test), IG(Information Gain), RF(Relief-F), S-R(SVM-RFE), MR2(mRMR) and GA(Genetic Algorithm).

Gene Selection Methods	Classification Methods	
	SVM Accuracy	kNN Accuracy
Chi Square Test	94.40	94.44
Information Gain	93.00	93.00
Relief-F	95.80	93.00
SVM-RFE	98.33	98.60
mRMR	95.83	95.83
GA	98.61	84.72

Table 4.7: LOOCV accuracy achieved by gene selection using optimal # of genes and classification methods on Leukemia data.

Chapter 5

Concluding Remarks

5.1 Conclusion and Discussion

We presented a comparative study of state-of-the-art gene selection methods and classification methods based on gene expression data. The efficiency of two different classification methods including: SVM and KNN, and eight different gene selection methods, including t-test, chi square test, information gain, relief-F, SVM-RFE, mRMR, genetic algorithm, and differential evaluation was compared. These methods were applied to two publicly available gene expression data sets Colon Tumor and Leukemia. Hold out and LOOCV were used to evaluate the classification performance.

Different experiments have been applied to compare the performance of the classification methods with and without performing feature selection. Results revealed the importance of feature selection in classifying gene expression data. By performing feature selection, the classification accuracy can be significantly boosted by using a small number of genes. We found that there is no any exact winner among used gene selection methods. With different data, with different classification methods, the best gene selection methods changes but we can say overall Relief-F, Information Gain, SVM-RFE and GA performed better as compared to others.

5.2 **Future Work**

In the future, we will try to study more gene selection methods and more classification methods. In this research, we have used only two datasets. It may be possible that if we apply these gene selection methods with more datasets, we can perform better evaluation.

Another thing we can do is, we can select those genes which are selected by most of gene selection methods and using those genes, we can analyze the results.
Another direction of future research is to combine the information from different data sets together. It is a commonly seen scenario that there are a number of biological data sets, that share the same features but are collected by different groups under different experimental conditions. Thus, they may have different underlying distributions. Yet they share highly relevant information. Each data set may be small and not sufficient to learn a good classifier. In such cases, transfer learning is a possible way to borrow information between the data sets. For instance, if we combine the two data sets Freije and Phillips together, we will end up with 172 samples (74 from Freije and 98 from Phillips).

Bibliography

[1] F. Crick. The biological replication of macromolecules. Symposium of the Society of Experimental Biology, (1958), 12:138-163.

[2] T. Paul, and H. Iba, Extraction of informative genes from microarray data. In Proceedings of the Genetic and Evolutionary Computation Conference, (2005):453–460.

[3] Shah, S., Kusiak, A.: Cancer gene search with data mining and genetic algorithms. Computers in Biology and Medicine 37(2), 251–261 (2007)].

[4] Liu, H., Liu, L., Zhang, H.: Ensemble gene selection for cancer classifi-cation. Pattern Recognition 43(8), 2763–2772 (2010) ISSN 0031-3203, 10.1016/j.patcog.2010.02.008.

[5] Park, C., Cho, S.-B.: Evolutionary ensemble classifier for lymphoma and colon cancer classification. In: The 2003 Congress on Evolutionary Computation, CEC 2003, December 8-12, vol. 4, pp. 2378–2385 (2003).

[6] Mohamad, M.S., Omatu, S., Yoshioka, M., Deris, S.: An Approach Using Hybrid Methods to Select Informative Genes from Microarray Data for Cancer Classification. In: Second Asia International Conference on Modeling & Simulation, AICMS 2008, May 13-15, pp. 603–608 (2008).

[7] Mohamad, M.S., Omatu, S., Deris, S., Hashim, S.Z.M.: A Model for Gene Selection and Classification of Gene Expression Data. International Journal of Artificial Life & Robotics 11(2), 219–222 (2007).

[8] T. Golub, D. Slonim , P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, and M. Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, (1999), 286:7-531.

[9] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classication with gene expression profiles. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, (2000): 583-598.

[10] M. Chow, I. Moler, and M. Ejand. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. Physiol Genomics, (2001), 5:99-111.

[11] R. Blanco, P. Larranaga, I. Inza, and B. Sierra. Gene selection for cancer classification using wrapper approaches. International Journal of Pattern Recognition and Artificial Intelligence, (2004), 18(8):1373-1390.

[12] L. Ein-Dor, I. Kela, G.Getz, D. Givol, E. Domany. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics, (2004), 12:171-178.

[13] A. Statnikov, C. Aliferis, I. Tsamardinos, D.Hardin, and S. Levy. A comprehensive evaluation of multi category classification methods for microarray gene expression cancer diagnosis. Bioinformatics, (2005), 21(5):631-643.

[14] J. Zhang, and H. Deng. Gene selection for classification of microarray data based on the Bayes error. *BMC* Bioinformatic, (2007), 8:370.

[15] R. Kohavi, and G. John. Wrappers for feature subset selection. Artificial Intelligence, (1997), 97(1-2):273-324.

[16] E. Xing, M. Jordan, and R. Karp. Feature selection for high dimensional genomic microarray data. In Proceedings of the 18th International Conference on Machine Learning, (2001):601-608.

[17] Osareh, A., Shadgar, B.: Microarray data analysis for cancer classification. In: 2010 5th International Symposium on Health Informatics and Bioinformatics (HIBIT), April 20-22, pp. 125–132 (2010).

[18] Nurminen, J.K.: Using software complexity measures to analyze algorithms—an experiment with the shortest-paths algorithms. Computers & Operations Research 30(8), 1121– 1134 (2003) ISSN 0305-0548, 10.1016/S0305-0548(02)00060-6.

[19] M. Pirooznia, J. Yang, M. Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics,(2008), 9(S1):S13.

[20] H. Liu, J. Li, and L. Wong. A Comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informatics (2002), 13:51–60.

[21] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, (2004), 20(15):2429-2437.

[22] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, (2004), 20(15):2429-2437.

[23] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, (2004), 20(15):2429-2437.

[24] E.P. Xing, M.I. Jordan, and R.M. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," Proc. 18th Int'l Conf. Machine Learning, (2001).

[25] Mundra PA, Rajapakse JC. Svm-rfe with mrmr filter for gene selection. IEEE Trans Nanobiosci (2010);9.

[26] J. Jaeger, R. Sengupta, and W.L. Ruzzo, "Improved Gene Selection for Classification of Microarrays," Proc. Pacific Symp. Biocomputing, pp. 53-64, (2003).

[27] D. Karaboga, B. Akay, A comparative study of artificial bee colony algorithm, Applied Mathematics and Computation 214 (2009) 108–132.

[28] H. Alshamlan, G. Badr, Y. Alohali mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling.Journal of BioMed Research (2015).

[29] P. Dunn-Rankin, G. Knezek, S. Wallace, and S. Zhang. Scaling methods. Lawrence Erlbaum, (2004), Mahwah, NJ.

[30] B. Bolstad, R. Irizarry, M. Astrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, (2003);19(2):185-193.

[31] F. Ahmad, S. Deris, N. Norwawi, and N. Othman. A review of feature selection techniques via gene expression profiles, IEEE, (2008), 2:1-7.

[32] T. Golub and D. Slonim *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, (1999).

[33] S. Cho, and H. Won. Machine learning in DNA microarray analysis for cancer classification. In Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics, (2003): 189-198.

[34] Kononenko I. Estimating features: analysis and extension of RELIEF. In: Proc. 6th European conf. onmachine learning; (1994). p. 171–82.

[35] Wall, M. E.; Rechtsteiner, A.; and Rocha, L. M. A Practical Approach to Microarray Data Analysis. Norwell, MA: Kluwel.(2003) chapter 5, 91–109.

[36] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, (2002).

[37] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol (2005);3:185–205.

[38] Peng H et al. Feature selection based on mutual information: criteria of maxdependency max relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell (2005);27:1226 38.

[39] Leping lie, Clarice R. Weinberg (2001), "Gene Selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method" Bioinformatics(2001)-Li-1131-42.

[40] Storn, R. and Price, K., 'Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces', Journal of Global Optimization, (1997) 11, pp. 341–359.

[41] X. Wu, V. Kumar, J. Quinlan , J. Ghosh , Q. Yang , H. Motoda , G. McLachlan, B. Liu , P. Yu, Z. Zhou, M. Steinbach, D. Hand, and D. Steinberg. Top 10 algorithms in data mining. Knowledge and Information Systems,(2008), 14(1):1-37.

[42] Vapnik, V.. Statistical Learning Theory: Wiley-Interscience, NY, USA(1998).

[43] N. Cristianini, C. Campbell, and J. Shawe-Taylor. An Introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, (2000): 204 –210.

[44] B. Schlkopf, K. Tsuda, and J. Vert. Kernel methods in computational biology. MIT Press series on Computational Molecular Biology, (2004):131-154.

[45] Fix, E. & Hodges, J.L., "Discriminatory analysis-nonparametric discription: Consistency properties (No. 4). " Randolf Field, Texas: USAF School of Aviation Medicine (1951).

[46] S. Bay. Combining nearest neighbor classifiers through multiple feature subsets. International Conference on Machine Learning, (1998), 37–45.

[47] T. Cover, and P. Hart. Neighbor pattern Classification. IEEE Transactions on Information Theory, (1967), 13:21-27.

[48] E. Dougherty, S. Chao, J. Hua, B. Hanczar, and U. Braga-Neto. Performance of error estimators for classification. Bioinformatics,(2010), 5(1):53.