Chapter 5

Relevant Gene Subset Selection using Ranking and Statistical Redundancy Reduction Methods

5.1 Introduction

In many studies on cancer classification using microarray data, gene ranking approaches have been widely investigated (Devijver and Kittler, 1982; Tibsrani et al., 2002; Guyon et al., 2003). In gene ranking approach, genes are evaluated using statistical characteristics of data and topmost $k$ genes are selected for building a classifier. Different gene ranking methods measure different characteristics of data. Therefore, the informative genes selected by different ranking methods may not be same. If one particular gene ranking method is used then there is a possibility of missing some informative genes. Another disadvantage associated with gene ranking methods is that they ignore correlation among genes because of their univariate approach. Hence, the selected gene subset may have low discriminatory capacity and higher redundancy.

In this chapter, a novel two-stage ensemble approach to determine a subset of relevant genes for accurate cancer classification is proposed. In first stage, a pool of important genes is created by taking union of topmost $k$ genes selected by different gene ranking methods. This reduces the chance of missing informative genes. However, this set of informative genes may contain redundant genes as ranking methods do not take into account correlation between different genes. In general, redundant genes degrade the performance of classifier and also increase computational cost. Hence, there is a need to remove redundancy. In second stage, a sequential forward feature selection method is used with a measure that reduces redundancy by considering correlation between genes. Three different measures Chernoff Distance (JC), Kullback Divergence (JD) and Linear Regression for redundancy reduction have been investigated.

Chernoff Distance (JC) and Kullback Diversion (JD) are estimated under the assumption that the class probability density follows multivariate normal density. Under this assumption closed form expressions for JC and JD are known. Chernoff Distance and Kullback Divergence involve computation of mean vector, determinant and inverse
of covariance matrix. As sequential forward search is used, genes are added one by one. This requires computation of mean vector, determinant and inverse incrementally. We have proposed a recursive definition to compute redundancy measures JC and JD. The recursive formulation of Bayesian classifier, which is used as one of the classifiers, is also proposed. This helps in reducing the time complexity to compute JC and JD. The training time and testing time of Bayesian classifier is also reduced using incremental framework.

5.2 Gene Ranking Methods

Most of the studies on gene selection adopt gene ranking methods because of their computational efficiency. Gene ranking methods aim to retain a certain number of genes, especially by ranking threshold, with scores determined according to a measure of relevance, discriminatory capability and information content or quality index. Some of the commonly used ranking methods for gene selection are Pearson’s Coefficient (PC), Spearman Coefficient (SC), Signal-to-noise ratio (SN), t-statistics (TS) and Kruskal-Wallis (KW). A brief description of these measures is given below.

5.2.1 Pearson's Correlation Coefficient

Pearson correlation coefficient (Dowdy and Wearden, 1983) is the most familiar measure for finding linear dependency between two quantities. Pearson correlation coefficient \( \rho_{X,Y} \) between two random variables \( X \) and \( Y \) with standard deviations \( \sigma_X \) and \( \sigma_Y \) is defined as:

\[
\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}
\]  

(5.1)

Pearson correlation is defined only if both of the standard deviations are finite and nonzero. It is a corollary of the Cauchy–Schwarz inequality that the correlation between \( X \) and \( Y \) cannot exceed 1 in absolute value. Correlation coefficient is symmetric i.e. \( \rho_{X,Y} = \rho_{Y,X} \). It is +1 in the case of a perfect positive (increasing) linear relationship, −1 in the case of a perfect decreasing (negative) linear relationship and between −1 and 1 in all other cases. As it approaches zero, there is less correlation between the variables. The closer the coefficient is to either −1 or 1, the stronger is the
correlation between the variables. To rank genes, Pearson correlation is calculated between each gene and class labels.

5.2.2 Spearman Coefficient

Spearman Coefficient (Dowdy and Wearden, 1983) is a non-parametric measure of statistical dependence between two variables. It is named after Charles Spearman. It assesses how well the relationship between two variables can be described using a monotonic function. A perfect Spearman correlation of +1 or −1 occurs when one variable is a perfect monotonic function of other variable. The $n$ raw scores $X_i, Y_i$ are converted to ranks $x_i, y_i$, and differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are calculated. Spearman coefficient $\rho$ is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$  (5.2)

For ranking genes, spearman coefficient between each gene and class labels is calculated.

5.2.3 Signal-to-Noise Ratio

This method is used for ranking genes of two-class problem. It is first used by Golub et al. (1999) for ranking genes. This method ranks the genes based on their ability to separate the two classes. It is given by

$$SN(g) = \frac{|\mu_1 - \mu_2|}{|\sigma_1 + \sigma_2|}$$  (5.3)

where $\mu_1$ and $\sigma_1$ are the mean and variance of the samples belonging to one class and $\mu_2$ and $\sigma_2$ are the mean and variance of samples belonging to other class. The gene with higher signal to noise ratio is able to distinguish the class better and hence is ranked higher.

5.2.4 T-statistics

It is widely used in literature to rank genes. It was proposed by Welch (1947). It measures the difference between the distributions of two groups of samples. Assuming
that there are two classes of samples in the gene expression dataset, the t-value for gene g is given by:

\[
TS(g) = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}
\]  (5.4)

where \(\mu_1\) and \(\sigma_1\) are the mean and variance of samples belonging to one class and \(\mu_2\) and \(\sigma_2\) are the mean and variance of samples belonging to other class. \(n_1\) is the number to sample belonging to one class and \(n_2\) is the number of samples belonging to other class. The gene that shows larger distinction between two groups is more important for classification. Hence, the top genes ranked by t-statistics are selected for classification.

### 5.2.5 Kruskal Wallis

Kruskal-Wallis test (Gibbons, 1985) is a nonparametric version of the classical one-way ANOVA. This method compares medians of different genes and returns p-value for the null hypothesis that all samples are drawn from the same population. The Kruskal Wallis test starts by substituting rank in the overall dataset for each measurement value. The sum of ranks is calculated for each group and then the test statistics H is calculated using the following equation

\[
H = \frac{(n-1)\sum_{i=1}^{k} n_i (r_i - \bar{r})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2}
\]  (5.5)

where \(r_{ij}\) is the rank of sample \(j\) from class \(c_i\), \(n_i\) is the number of samples of class \(c_i\), \(r_i = \frac{\sum r_{ij}}{n_i}\), \(\bar{r} = \frac{1}{2}(n + 1)\)

and \(n\) is total number of samples.

All the methods discussed above are widely used for ranking different genes. Once the genes are ranked, topmost \(k\) genes are selected and classifier is built using these genes. These methods are simple to implement and takes less computation time. The time complexity of ranking algorithm is \(O(d \log d)\) where \(d\) is the number of genes.
in original set. However, a common drawback with gene ranking methods is that they assume genes are independent to each other and only detect relationship between a gene and class label. The mutual information among genes is discarded. In microarray datasets, there is lot of redundancy and if one gene is ranked higher, then correlated gene will also be ranked higher. Hence the selected gene subset may contain redundant genes which need to be removed.

5.3 Redundancy Reduction Method

In literature, few methods have been suggested which reduce redundancy by considering relationship between genes. Some of the methods suggested in literature for removing redundancy for two class problems are Linear Regression (Han-Saem et al., 2007), Chernoff Distance measure and Kullback Divergence measure (Devijver and Kittler, 1982). A brief explanation of each is given below:

5.3.1 Regression Analysis

It is an approach to model the relationship between a target variable $C$ and one or more variables denoted as $X$. In regression analysis, the partial correlations between the target variable and the variables that explain the target well are analyzed. It can analyze the impact of one or more variables on another variable. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable, by fitting a linear equation to observed data. While using regression analysis for gene expression data, a multiple regression model is considered as there can be many genes which may affect the presence or absence of cancer. A multiple regression model with a target variable $C$ and multiple variables $x_1, x_2$ and $x_3$ is given by (Han-Saem, et al., 2007):

$$c_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \xi_i, \quad i = 1, 2, ..., n \tag{5.6}$$

where $\beta_0, \beta_1, \beta_2$ are constants estimated by observed values of genes $x_1, x_2, x_3$ and class label $C$, and $\xi_i$ is estimated by normal distribution having mean zero and variance $\sigma^2$. After the values of constants are estimated, the value of class label $c_i$ is calculated using eqn. (5.6). The sum of squared error (SSE) is given by:
\[ SSE = \sum_{i=1}^{n} \left( c_i - \text{predicted } c_i \right)^2 \]  

(5.7)

A large value of SSE means that the regression line is predicted poorly. The total sum of squares (SSTO) is given by:

\[ SSTO = \sum_{i=1}^{n} \left( c_i - \bar{c} \right)^2 \]  

(5.8)

where \( \bar{c} \) is the average of \( c_i \).

In a regression model, the choice of genes which best explains the class label depends on the value of \( R^2 \) given by:

\[ R^2 = 1 - \frac{SSE}{SSTO} \]  

(5.9)

The value of \( R^2 \) varies between 0 and 1. For each variable which is not selected, a regression model is created. The variables are ranked based on the value of \( R^2 \) and at each stage the variable with maximum \( R^2 \) value is selected. When maximum \( R^2 \) becomes zero, the process is terminated.

### 5.3.2 Chernoff Distance Measure

It is a probabilistic distance measure based on Bayes error (Devijver and Kittler, 1982). It provides an upper bound on the classification error. It can be used to select those genes which minimize Bayes error. The general form of Chernoff Distance measure (JC) is given by:

\[ JC = -\log(\int p(X | c_i) p(X | c_j) dx) \]  

(5.10)

For two class problem, when \( p(X | c_i) \) follows multivariate normal distribution i.e. \( p(X | c_i) = N(\mu_i, \Sigma_i) \), \( i = 1, 2 \), JC takes the following closed form expression (Duda et al., 2002):

\[ JC = \frac{1}{2} \beta (1 - \beta) (\mu_2 - \mu_1) \left[ (1 - \beta) \Sigma_2 + \beta \Sigma_1 \right]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \log \left| (1 - \beta) \Sigma_1 + \beta \Sigma_2 \right| \left( \Sigma_1 \right)^{-\beta} \left| \Sigma_2 \right|^\beta \]  

(5.11)
Chernoff Distance can be used as a criterion for gene selection in sequential forward gene selection. At each stage, the gene which maximizes Chernoff Distance should be selected.

5.3.3 Kullback Divergence Measure

This measure is another probabilistic distance measure used for gene selection based on the interclass separability measure (Devijver and Kittler, 1982). It selects the gene which increases interclass separability. The general form of Kullback Divergence (JD) measure is given by:

\[
JD = \int \left( p(X|c_i) - p(X|c_2) \right)^* \log \left( \frac{p(X|c_i)}{p(X|c_2)} \right) dx
\]  

(5.12)

For a two class problem, when \( p(X|c_i) \) follows multivariate normal distribution i.e. \( p(X|c_i) = N(\mu_i, \Sigma_i), i = 1, 2 \), JD takes the following closed form expression.

\[
JD = \frac{1}{2} (\mu_2 - \mu_1)' \left( (\Sigma_1)^{-1} + (\Sigma_2)^{-1} \right) (\mu_2 - \mu_1) + \frac{1}{2} \text{tr} \left( (\Sigma_1)^{-1} \Sigma_2 + (\Sigma_2)^{-1} \Sigma_1 \right)
\]

(5.13)

For gene selection, Kullback Divergence (JD) can be used as a criterion in sequential forward feature selection. At each stage, the gene which maximizes Kullback Divergence measure should be selected.

5.4 Incremental Framework

The common characteristic of all search algorithms involved in gene selection is that the gene subsets to be evaluated at the \((k+1)^{th}\) stage of an algorithm are constructed from an appropriate gene set, \( F_k \), obtained at the \( k^{th} \) step of algorithm by adding (or subtracting) a small number of genes to the set. If we already know the value of the function at \( k^{th} \) stage, it would be recommendable to use this value to calculate the value of the function at \((k+1)^{th}\) stage. From eqn. (5.13) it is observed that for computing JD, we need to calculate the covariance matrix and inverse of covariance matrix. Similarly, we need to calculate covariance matrix, its determinant and inverse of covariance matrix to compute JC. The computation of determinant and inverse of a matrix are computationally expensive. In this section, we have proposed a method to incrementally update Kullback Divergence and Chernoff Distance measures. An incremental method
for updating the probability density in the training phase of Bayesian classifier, when features following multivariate normal distribution are considered one by one, is also proposed. This incremental formulation significantly improves the computation time of both gene subset selection and Bayesian Classifier.

5.4.1 Incremental JD and JC Measures

As a new gene is considered in addition to existing k genes, the new gene vector is represented by $X_{k+1} = [x_1, x_2, ..., x_k, x_{k+1}]^T$ i.e. $X_{k+1} = [x_k^T x_{k+1}]^T$. For $X_{k+1}$, JD measure is given by:

$$J_{D,k+1} = \frac{1}{2} (\mu_{k+1}^2 - \mu_{k+1}^1)^T \left( \Sigma_{k+1}^{-2} + \left( \Sigma_{k+1}^{-1} \right)^{-1} \right) \left( \mu_{k+1}^2 - \mu_{k+1}^1 \right) +$$

$$\frac{1}{2} \text{tr} \left( \left( \Sigma_{k+1}^{-1} \right)^{-1} \Sigma_{k+1}^{2} + \left( \Sigma_{k+1}^{-1} \right)^{-1} \Sigma_{k+1}^{1} \right)$$

(5.14)

where $\mu_{k+1}^i$ is a mean vector and $\Sigma_{k+1}^i$ is a covariance matrix of the multivariate normal distribution for class $c_i$. Generally, the computation of $J_{D,k+1}$ is carried out in batch mode which involves computation of $\mu_{k+1}^i, \Sigma_{k+1}^i$ and $\left( \Sigma_{k+1}^{-1} \right)^{-1}, i = 1, 2$ for every new addition of gene to existing set of k-genes. The time complexity of computing mean vector is $O(k)$ while the time complexity of computing inverse of covariance matrix is $O(k^3)$ where k is the number of genes used for classification. Here, it is assumed that inverse of a matrix is solved using Gaussian method. The new mean vector $\mu_{k+1}^i$ and the new covariance matrix $\Sigma_{k+1}^i$ can be represented in terms of $\mu_{k}^i$ and $\Sigma_{k}^i$ as follows:

$$\mu_{k+1}^i = \begin{bmatrix} (\mu_{k}^i)^T \\ \mu_{k+1}^{i} \end{bmatrix}^T$$

(5.15)

where $\mu_{k+1}^i$ is the mean of $(k+1)^{th}$ gene of class $c_i$ and

$$\Sigma_{k+1}^i = \begin{bmatrix} \Sigma_{k}^i & B^i \\ (B^i)^T & D^i \end{bmatrix}$$

(5.16)

where $B^i$ is a $(k \times 1)$ matrix and its elements represent covariance $\sigma_{k+1,i}, i = 1, 2, ..., k$ and $D^i$ is a scalar covariance element $\sigma_{k+1,k+1}$ of gene $k+1$ of class $c_i$. 

45
Using classical results in matrix algebra (Golub et al., 1996), $|\Sigma_{k+1}^i|$ can be written in terms of $|\Sigma_k^i|$ and $(\Sigma_k^i)^{-1}$ in terms of $(\Sigma_k^i)^{-1}$ as follows:

$$|\Sigma_{k+1}^i| = |P_i^i|*|\Sigma_k^i| \quad (5.17)$$

where

$$P_i^i = \left| D^i - (B^i)^t(\Sigma_k^i)^{-1}(B^i) \right|$$

and

$$\left(\Sigma_{k+1}^i\right)^{-1} = \left[ \begin{array}{cc} (S_A^i)^{-1} & -\left(\Sigma_k^i\right)^{-1}B^i(S_D^i)^{-1} \\ -\left(D^i\right)^{-1}(B^i)^t(S_A^i)^{-1} & (S_D^i)^{-1} \end{array} \right] \quad (5.18)$$

where

$$S_A^i = \Sigma_k^i - B^i\left(D^i\right)^{-1}(B^i)^t$$

$$S_D^i = D^i - (B^i)^t(\Sigma_k^i)^{-1}B^i$$

provided the sub matrices $\Sigma_k^i$ and $D^i$ are nonsingular.

Using Woodbury Formula (Golub et al., 1996) in eqn. (5.18), we have

$$\left(\Sigma_{k+1}^i\right)^{-1} = \left[ \begin{array}{cc} (\Sigma_k^i)^{-1} + \mathbf{Y}_{11}^i & \mathbf{Y}_{12}^i \\ \mathbf{Y}_{21}^i & \mathbf{Y}_{22}^i \end{array} \right] \quad (5.19)$$

where

$$\mathbf{Y}_{11}^i = -\left(\Sigma_k^i\right)^{-1}(-B^i)(\mathbf{I} + (D^i)^{-1}(B^i)^t(\Sigma_k^i)^{-1}(-B^i))^{-1}(D^i)^{-1}(B^i)^t(\Sigma_k^i)^{-1}$$

$$\mathbf{Y}_{12}^i = -\left(\Sigma_k^i\right)^{-1}B^i(S_D^i)^{-1}$$

$$\mathbf{Y}_{21}^i = -(D^i)^{-1}(B^i)^t((\Sigma_k^i)^{-1} - \mathbf{Y}_{11}^i)$$

$$\mathbf{Y}_{22}^i = (S_D^i)^{-1}$$

Using eqn. (5.14), eqn. (5.15), eqn. (5.16) and eqn. (5.19), $JD_{k+1}^i$ can be given by

$$JD_{k+1}^i = JD_k^i + H^i + \frac{1}{2} trace \left[ \begin{array}{cc} \mathbf{M}_1 - 2\mathbf{I} & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{array} \right] \quad (5.20)$$
where,

\[
H^D = \frac{1}{2} \left[ (\mu_k^2 - \mu_k^1)G_1 (\mu_k^2 - \mu_k^1)' + (\mu_{k+1}^2 - \mu_{k+1}^1)G_3 (\mu_k^2 - \mu_k^1)' + (\mu_k^2 - \mu_k^1)G_2 (\mu_{k+1}^2 - \mu_{k+1}^1)' + (\mu_{k+1}^2 - \mu_{k+1}^1)G_4 (\mu_k^2 - \mu_k^1)' \right]
\]

\[
G_1 = Y_{11}^1 + Y_{11}^2 \quad G_2 = Y_{12}^1 + Y_{12}^2
\]
\[
G_3 = Y_{21}^1 + Y_{21}^2 \quad G_4 = Y_{22}^1 + Y_{22}^2
\]
\[
M_1 = Y_{11}^1 \Sigma_k^2 + Y_{12}^1 (B^2)' + Y_{12}^2 \Sigma_k^1 + Y_{12}^2 (B^1)'
\]
\[
M_2 = (\Sigma_k^1)^{-1} + Y_{11}^1 B^2 + Y_{12}^1 D^2 + (\Sigma_k^2)^{-1} + Y_{11}^1 B^1 + Y_{12}^2 D^1
\]
\[
M_3 = (Y_{21}^1 \Sigma_k^2 + Y_{22}^1 (B^2)')^2 + (Y_{21}^2 \Sigma_k^1 + Y_{22}^2 (B^1)')
\]
\[
M_4 = (Y_{21}^2 B^2 + Y_{22}^2 D^2) + (Y_{21}^2 B^1 + Y_{22}^2 D^1)
\]

Using eqn. (5.11), JC measure for \(X_{k+1}\) is given by

\[
JC_{k+1} = \frac{1}{2} \beta (1 - \beta) (\mu_{k+1}^2 - \mu_{k+1}^1)' (1 - \beta) \Sigma_{k+1}^1 + \beta \Sigma_{k+1}^2 \right)^{-1} (\mu_{k+1}^2 - \mu_{k+1}^1) + \frac{1}{2} \log \left| \frac{(1 - \beta) \Sigma_{k+1}^1 + \beta \Sigma_{k+1}^2}{(1 - \beta) \Sigma_{k+1}^1 + \beta \Sigma_{k+1}^2} \right|^\beta
\]  
(5.21)

Using eqn. (5.16) and eqn. (5.17), term \( (1 - \beta) \Sigma_{k+1}^1 + \beta \Sigma_{k+1}^2 \) involved in eqn. (5.21) can rewritten as

\[
(1 - \beta) \Sigma_{k+1}^1 + \beta \Sigma_{k+1}^2 = |F| (1 - \beta) \Sigma_k^1 + \beta \Sigma_k^2
\]  
(5.22)

where

\[
|F| = \left| \frac{(1 - \beta) D^1 + \beta D^2} - (1 - \beta) B^1 + \beta B^2 \right| \left( (1 - \beta) \Sigma_k^1 + \beta \Sigma_k^2 \right)^{-1}
\]

Using eqn. (5.16) and eqn. (5.19), term \( (1 - \beta) \Sigma_{k+1}^1 + \beta \Sigma_{k+1}^2 \)^{-1} can be given by

\[
(1 - \beta) \Sigma_{k+1}^1 + \beta \Sigma_{k+1}^2 \left[ Q_2 \begin{bmatrix} Q_1 & Q_3 \\ Q_2 & Q_4 \end{bmatrix} \right]
\]  
(5.23)
where

\[
Q_1 = -\left( (1 - \beta)\Sigma^1_k + \beta \Sigma^2_k \right)^{-1} \left( (1 - \beta)B^1 + \beta B^2 \right)
\]

\[
\left[ I + \left( (1 - \beta)D^1 + \beta D^2 \right)^{-1} \left( (1 - \beta)B^1 + \beta B^2 \right) \right]
\]

\[
\left[ \left( (1 - \beta)\Sigma^1_k + \beta \Sigma^2_k \right)^{-1} \left( - \left( (1 - \beta)B^1 + \beta B^2 \right) \right) \right]
\]

\[
\left( \left( (1 - \beta)\Sigma^1_k + \beta \Sigma^2_k \right)^{-1} \left( (1 - \beta)B^1 + \beta B^2 \right) \right)
\]

\[
\left( (1 - \beta)\Sigma^1_k + \beta \Sigma^2_k \right)^{-1}
\]

\[
Q_2 = -\left( (1 - \beta)D^1 + \beta D^2 \right)^{-1} \left( (1 - \beta)B^1 + \beta B^2 \right)
\]

\[
\left( \left( (1 - \beta)\Sigma^1_k + \beta \Sigma^2_k \right)^{-1} - Q_1 \right)
\]

\[
Q_3 = -\left( (1 - \beta)\Sigma^1_k + \beta \Sigma^2_k \right)^{-1} \left( (1 - \beta)B^1 + \beta B^2 \right)
\]

\[
\left( S_D \right)^{-1}
\]

\[
Q_4 = \left( (1 - \beta)D^1 + \beta D^2 \right)^{-1} - \left( (1 - \beta)B^1 + \beta B^2 \right)
\]

\[
\left( (1 - \beta)\Sigma^1_k + \beta \Sigma^2_k \right)^{-1} \left( (1 - \beta)B^1 + \beta B^2 \right)
\]

Therefore, substituting eqn. (5.16), eqn. (5.22) and eqn. (5.23) in eqn. (5.21), the incremental updation of Chernoff Distance measure is given by:

\[
JC_{k+1} = JC_k + \frac{1}{2} \beta (1 - \beta)H^c_k + \frac{1}{2} \left[ \frac{|F|}{|P_1| - \beta |P_2|^\beta} \right] \quad (5.24)
\]

where

\[
H^c_k = \left( \mu^2_k - \mu^1_k \right) Q_1 \left( \mu^2_k - \mu^1_k \right) + \left( \mu^2_{k+1} - \mu^1_{k+1} \right) Q_2 \left( \mu^2_k - \mu^1_k \right)
\]

\[
+ \left( \mu^2_{k+1} - \mu^1_{k+1} \right) Q_3 \left( \mu^2_{k+1} - \mu^1_{k+1} \right) + \left( \mu^2_{k+1} - \mu^1_{k+1} \right) Q_4^{-1} \left( \mu^2_{k+1} - \mu^1_{k+1} \right)
\]

In this way, the knowledge of incremental updation of mean vector, determinant and inverse of covariance matrix can be utilized for incremental updation of JD and JC measures.
5.4.2 Incremental Bayesian Classifier

As discussed in chapter 3, the general multivariate normal probability density of k-dimensional sample \( X_k = [x_1, x_2, \ldots, x_k]^T \) for a given class \( c_i \) is given by

\[
p(X_k | c_i) = \frac{1}{(2\pi)^{k/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2} (X_k - \mu_i')\Sigma_i^{-1}(X_k - \mu_i')^T\right)
\]  
(5.25)

where \( \mu_i' \) is a mean vector and \( \Sigma_i \) is a \( k \times k \) covariance matrix of the normal distribution for class \( c_i \).

As a new gene is considered in addition to existing k genes, the new gene vector is represented by \( X_{k+1} = [x_1, x_2, \ldots, x_k, x_{k+1}]^T \). The probability density of \( X_{k+1} \) for a class \( c_i \) is given by

\[
p(X_{k+1} | c_i) = \frac{1}{(2\pi)^{(k+1)/2} |\Sigma_{k+1}|^{1/2}} \exp\left(-\frac{1}{2} (X_{k+1} - \mu_{i+1}')\Sigma_{k+1}^{-1}(X_{k+1} - \mu_{i+1}')^T\right)
\]  
(5.26)

where \( \mu_{i+1}' \) is a mean vector and \( \Sigma_{k+1} \) is a covariance matrix of the normal distribution for class \( c_i \).

In the proposed incremental Bayesian classifier, during training phase, whenever a new gene arrives, pre-computed \( |\Sigma_i| \) and \( (\Sigma_i)^{-1} \) is used for computation of \( |\Sigma_{k+1}| \) using eqn. (5.17) and \( (\Sigma_{k+1})^{-1} \) will be computed using eqn. (5.18). Using (5.17) \( |\Sigma_{k+1}| \) can be calculated in \( O(k^2) \). Also computation of \( \Sigma_{k+1}^{-1} \) using (5.18) requires \( O(k^3) \) time. Hence, incremental Bayesian classifier is computationally more efficient in comparison to batch Bayesian classifier.

5.5 Proposed Method

Linear regression, Chernoff Distance measure and Kullback Divergence measure cannot be used directly on high dimensional microarray datasets for determining relevant gene set as it involves large computation time. Hence, we propose a two stage approach for gene selection. In the first stage, gene ranking method is used to select top \( m \) genes. In literature, it has been observed that different gene ranking techniques rank genes differently. So if one gene ranking method is chosen, there are chances of losing some
informative genes. Therefore, to take into account the variation due to different ranking methods, union of informative genes selected by different gene ranking methods is considered. This reduces the chance of missing informative genes. To account for variation due to different training data, for each method, a pool of relevant genes is selected using 5-fold cross validation and union of top 25 genes from each ranked list is selected. However, ranking methods do not take into account the partial correlation among the selected genes as they only calculate the similarity between the class label and the gene itself on a one-to-one basis. Partial correlations of the selected genes are measured using one of the methods viz Kullback Divergence measure (JD), Chernoff Distance measure (JC) and linear regression. In the second phase, a sequential forward feature selection method with one of the above measures is used to remove redundancy. To select the smallest set of genes, classification accuracy is calculated using topmost \( i \) genes. The value of \( i \) for which classification accuracy is maximum is selected and top \( i \) genes are selected. The outline of the proposed two stage gene selection method is given below and shown in Figure 5.1:

1. \( f \) in [PC, SC, SN, TS, KW]
2. \( R(f) = \{g_{i1}, g_{i2}, ..., g_{im}\} \) // topmost \( m \) genes selected by using gene ranking method \( f \) using 5-fold cross validation
3. End
4. \( R = \text{Union}(R(\text{PC}), R(\text{SC}), R(\text{SN}), R(\text{TS}), R(\text{KW})) \)
5. Apply a Sequential Forward Selection method on \( R \) to select a set of relevant and non redundant genes using one of the measures such as JC, JD and linear regression.
6. Return the final subset of genes.
Gene ranking using M1  Gene ranking using M2  Gene ranking using M3  Gene ranking using M4  Gene ranking using M5

Union of top genes returned by each ranking method

Applying Redundancy Reduction Methods on the selected genes

Ranked list of genes

Calculate the accuracy using top i gene and return genes giving maximum accuracy

Figure 5.1: Outline of the Proposed Algorithm

5.6 Experimental Setup and Results

To evaluate our proposed algorithm, we have applied it on six well-known publicly available datasets. All these datasets are 2-class problems except SRBCT which has been converted into 2-class problem. In pre-processing, datasets are normalized using Z-score. Each dataset is divided into 5 parts and gene ranking method is applied 5 times. Each time ranking is done using four parts leaving one part. Thus, we got five different ranked lists. From each list, 25 top ranked genes are selected and their union is taken to get the final gene list. This procedure is repeated for different ranking methods and their union is taken to get the set of important genes. Different gene ranking methods used in our experiment are Spearman Coefficient, Pearson Coefficient, Signal-to-Noise ratio, t-statistics and Kruskal Wallis. After first stage, the number of genes is reduced considerably. Table 5.1 shows the number of genes after the first stage.
Table 5.1: Reduced Number of genes after stage 1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of genes</th>
<th>Number of genes after Stage 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>2000</td>
<td>56</td>
</tr>
<tr>
<td>SRBCT</td>
<td>2308</td>
<td>46</td>
</tr>
<tr>
<td>Prostate</td>
<td>5967</td>
<td>79</td>
</tr>
<tr>
<td>Leukemia</td>
<td>7129</td>
<td>54</td>
</tr>
<tr>
<td>Lungcancer</td>
<td>12534</td>
<td>159</td>
</tr>
<tr>
<td>Ovary</td>
<td>15154</td>
<td>108</td>
</tr>
</tbody>
</table>

In second stage, to remove redundancy from the selected pool of genes, three methods are investigated and compared: Chernoff Distance measure, Kullback Divergence measure and Linear Regression. Three classifiers namely SVM, KNN and Bayesian are used to evaluate the performance of genes selected by these methods. To calculate accuracy of classifier, leave-one-out cross validation (LOOCV) method is used. All the simulations are done using matlab. Linear kernel is used in SVM. For KNN, matlab function knnc is used in which the value of $k$ is optimized with respect to the leave-one-out error.

For Chernoff Distance measure, genes are selected using 5 different values of $\beta$ ranging from 0.1 to 0.9 with an increment of 0.2. To select a set of relevant and non-redundant genes in the second stage, classification accuracy is calculated as genes are added one by one. The maximum classification accuracy achieved for different datasets using different values of $\beta$ are shown in Table 5.2. It is observed from Table 5.2 that for different datasets maximum accuracy is achieved with different values of $\beta$. But in general, values of $\beta$ greater than 0.5 gives better classification accuracy. It is also observed from Table 5.2 that better classification accuracy is achieved with less number of genes. Gene subset is also determined using other redundancy methods: Kullback Divergence measure and Linear Regression. As genes are added one-by-one, their classification accuracy is calculated. Figures 5.2-5.7 show the variation in classification accuracy as the genes are added one-by-one for all the three redundancy reduction methods. For Chernoff distance the value of $\beta$ that gives best classification accuracy is considered.
<table>
<thead>
<tr>
<th></th>
<th>ColonCancer</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>SVM</td>
<td>90.32(11)</td>
<td>90.32(11)</td>
<td>93.54(16)</td>
<td>91.93(9)</td>
<td>91.93(8)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>88.70(8)</td>
<td>88.70(13)</td>
<td>91.93(15)</td>
<td>91.93(7)</td>
<td>91.93(11)</td>
</tr>
<tr>
<td>KNN</td>
<td>91.93(22)</td>
<td>91.93(10)</td>
<td>93.54(40)</td>
<td>91.93(9)</td>
<td>91.93(37)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Lungcancer</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>SVM</td>
<td>100(60)</td>
<td>100(27)</td>
<td>100(4)</td>
<td>100(2)</td>
<td>100(2)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>98.89(6)</td>
<td>100(15)</td>
<td>100(4)</td>
<td>100(16)</td>
<td>100(19)</td>
</tr>
<tr>
<td>KNN</td>
<td>100(30)</td>
<td>100(15)</td>
<td>100(4)</td>
<td>100(2)</td>
<td>100(2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SRBCT</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>SVM</td>
<td>100(14)</td>
<td>100(9)</td>
<td>100(9)</td>
<td>100(7)</td>
<td>100(7)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>98.79(11)</td>
<td>100(6)</td>
<td>100(6)</td>
<td>100(12)</td>
<td>100(8)</td>
</tr>
<tr>
<td>KNN</td>
<td>100(11)</td>
<td>100(4)</td>
<td>100(7)</td>
<td>100(8)</td>
<td>100(11)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Leukemia</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>SVM</td>
<td>98.61(19)</td>
<td>98.61(11)</td>
<td>100(12)</td>
<td>100(9)</td>
<td>100(16)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>100(18)</td>
<td>100(13)</td>
<td>100(10)</td>
<td>100(13)</td>
<td>100(14)</td>
</tr>
<tr>
<td>KNN</td>
<td>98.61(36)</td>
<td>98.61(19)</td>
<td>98.61(17)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Ovary</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>SVM</td>
<td>100(3)</td>
<td>100(3)</td>
<td>100(3)</td>
<td>100(3)</td>
<td>100(3)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>99.60(3)</td>
<td>100(3)</td>
<td>100(3)</td>
<td>100(3)</td>
<td>100(3)</td>
</tr>
<tr>
<td>KNN</td>
<td>100(3)</td>
<td>100(2)</td>
<td>100(2)</td>
<td>100(3)</td>
<td>100(3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Prostate</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>SVM</td>
<td>98.04(6)</td>
<td>98.04(6)</td>
<td>99.02(11)</td>
<td>98.04(4)</td>
<td>95.09(5)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>95.09(6)</td>
<td>95.09(8)</td>
<td>95.09(4)</td>
<td>95.09(4)</td>
<td>96.08(5)</td>
</tr>
<tr>
<td>KNN</td>
<td>97.06(6)</td>
<td>99.02(6)</td>
<td>98.04(4)</td>
<td>98.04(4)</td>
<td>96.08(4)</td>
</tr>
</tbody>
</table>
Figure 5.2: Variation of Classification Accuracy with number of genes using different redundancy reduction methods for Colon Cancer datasets
Figure 5.3: Variation of Classification Accuracy with number of genes using different redundancy reduction methods for SRBCT dataset
Figure 5.4: Variation of Classification Accuracy with number of genes using different redundancy reduction methods for Prostate Cancer dataset.
Figure 5.5: Variation of Classification Accuracy with number of genes using different redundancy reduction methods for Leukemia dataset
Figure 5.6: Variation of Classification Accuracy with number of genes using different redundancy reduction methods for Lungcancer dataset
Figure 5.7: Variation of Classification Accuracy with number of genes using different redundancy reduction methods for Ovary dataset.
Table 5.3 shows the best classification accuracy achieved with different classifiers along with the number of genes for different measures. We observe the following from Table 5.3:

1. For Lung Cancer, 100% classification accuracy is achieved for all the three classifiers. 100% classification accuracy is achieved with 2 genes, 12 genes and 5 genes using JC, JD and Linear Regression respectively. Best result is achieved by JC in combination with KNN and SVM classifier.

2. For Leukemia, 100% classification accuracy is achieved for all the three classifiers. Minimum number of genes is selected by linear regression method in conjunction with KNN. It is also observed that the genes selected by JD are not able to give 100% accuracy. Maximum classification accuracy achieved by the genes selected by JD is 98.61%. 100% classification accuracy is achieved using JC with 9 genes, 9 genes and 10 genes for KNN, SVM and Bayesian classifier respectively.

3. For SRBCT dataset, classification accuracy of 100% is achieved for all the classifiers. It is observed that 100% classification accuracy is achieved with only 4 genes using KNN and Bayesian classifier and with 7 genes using SVM classifier. Minimum number of genes is selected by JC using KNN and SVM and JD in case of Bayesian classifier.

4. For Colon dataset, classification accuracy of 93.54% is achieved for all the classifiers. For KNN and Bayesian, this classification accuracy is achieved with the genes selected by JD whereas for SVM, this classification accuracy is achieved with the genes selected with all the three methods. Maximum classification accuracy is achieved with 8 genes, 16 genes and 9 genes using KNN, SVM and Bayesian classifier respectively.

5. For Ovary dataset, classification accuracy of 100% is achieved with all the classifiers. For all the classifiers performance of JC and Linear Regression is same.

6. For Prostate dataset, classification accuracy of 99.02% is achieved for KNN and SVM. This accuracy is achieved with 6 genes and 11 genes in KNN and SVM respectively. Maximum classification accuracy achieved for Bayesian classifier is 96.08% using JC method.
7. It is also observed that maximum classification accuracy is achieved with SVM and KNN in conjunction with JC for all datasets.
8. None of the methods at second stage seems to be a clear winner. But in most of the datasets, it is observed that JC outperforms other measures. In general, maximum accuracy is achieved with the genes selected by JC. Also it is observed that performance of KNN is better than other classifiers in terms of number of genes required for achieving maximum accuracy.

Table 5.3: Maximum Accuracy with number of genes for different redundancy reduction methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>K-Nearest Neighbour</th>
<th>SVM</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JC</td>
<td>JD</td>
<td>Linear Reg</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>100 (2) (β=0.9)</td>
<td>100 (12)</td>
<td>100 (5)</td>
</tr>
<tr>
<td></td>
<td>Leukemia</td>
<td>100 (9) (β=0.7)</td>
<td>98.61 (9)</td>
</tr>
<tr>
<td></td>
<td>SRBCT</td>
<td>100 (4) (β=0.3)</td>
<td>100 (27)</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>93.54 (40) (β=0.5)</td>
<td>93.54 (8)</td>
<td>91.93 (7)</td>
</tr>
<tr>
<td></td>
<td>Ovary</td>
<td>100 (2) (β=0.3)</td>
<td>100 (3)</td>
</tr>
<tr>
<td>Prostate</td>
<td>99.02 (6) (β=0.3)</td>
<td>92.16 (4)</td>
<td>98.04 (4)</td>
</tr>
</tbody>
</table>

For each of the classifiers, the smallest set of genes giving maximum accuracy using different redundancy reduction technique has been listed in Table 5.4. It is observed from Table 5.4 that gene subset depends on the choice of classifier. The gene subset which is good for one classifier might not give good classification accuracy with other classifier. It can also be observed from Table 5.4 that some genes are common (shown in bold) with all the classifiers where as some genes are different.
Table 5.4: Smallest subset of non-redundant genes selected for different classifiers using the proposed method

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Maximum Accuracy</th>
<th>Selected Gene subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>KNN</td>
<td>93.54(8)</td>
<td>{765, 1042, 138, 1263, 72, 1325, 295, 1221}</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>93.54(16)</td>
<td>{765, 1042, 513, 1208, 1900, 1221, 377, 493, 1771, 1263, 391, 1411, 652, 1648, 822, 1839}</td>
</tr>
<tr>
<td></td>
<td>Bayesian</td>
<td>93.54(9)</td>
<td>{765, 1042, 138, 1263, 72, 1325, 295, 1221, 66}</td>
</tr>
<tr>
<td>SRBCT</td>
<td>KNN</td>
<td>100(4)</td>
<td>{1389, 365, 1203, 545}</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>100(7)</td>
<td>{1389, 365, 1203, 545, 1613, 1003, 384} {1319, 1389, 1708, 2050, 566, 1613, 36}</td>
</tr>
<tr>
<td></td>
<td>Bayesian</td>
<td>100(4)</td>
<td>{1708, 2050, 1319, 1954}</td>
</tr>
<tr>
<td>Prostate</td>
<td>KNN</td>
<td>99.02(6)</td>
<td>{5575, 4284, 591, 4843, 3972, 2586}</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>99.02(11)</td>
<td>{5575, 2586, 3972, 4843, 1816, 3381, 4238, 2992, 591, 5077, 4978}</td>
</tr>
<tr>
<td></td>
<td>Bayesian</td>
<td>96.08(5)</td>
<td>{591, 2711, 3381, 5077, 2586}</td>
</tr>
<tr>
<td>Leukemia</td>
<td>KNN</td>
<td>100(3)</td>
<td>{4847, 4951, 6281}</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>100(9)</td>
<td>{758, 4680, 1685, 2833, 2642, 2288, 1909, 760, 6855}</td>
</tr>
<tr>
<td></td>
<td>Bayesian</td>
<td>100(10)</td>
<td>{758, 2288, 4680, 760, 1779, 2128, 1882, 1834, 6539, 6855}</td>
</tr>
<tr>
<td>Ovary</td>
<td>KNN</td>
<td>100(2)</td>
<td>{1680, 181}</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>100(3)</td>
<td>{1680, 181, 2237}</td>
</tr>
<tr>
<td></td>
<td>Bayesian</td>
<td>100(3)</td>
<td>{1680, 181, 2237}</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>KNN</td>
<td>100(2)</td>
<td>{7200, 4336}</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>100(2)</td>
<td>{7200, 4336}</td>
</tr>
<tr>
<td></td>
<td>Bayesian</td>
<td>100(4)</td>
<td>{3764, 7200, 6139, 4336}</td>
</tr>
</tbody>
</table>
To check the relevance of the selected genes subset, we performed 10 fold cross validation using the selected genes for all the datasets. Experiment is repeated 10 times. The average classification accuracy of 10 runs along with standard deviation is given in Table 5.5. It can be observed from Table 5.5 that the 10 fold cross validation accuracy does not deviate much from LOOCV accuracy. This shows that the gene set selected is not over fitted.

### Table 5.5: Classification accuracy achieved by the genes selected by proposed method for different classifier using 10 fold cross-validation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>KNN</th>
<th>SVM</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>KNN</td>
<td>91.29±2.45</td>
<td>91.94±1.61</td>
<td>92.58±0.88</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>100±0</td>
<td>100±0</td>
<td>99.52±1.08</td>
</tr>
<tr>
<td>SRBCT</td>
<td>Kullback</td>
<td>100±0</td>
<td>100±0</td>
<td>98.99±0.62</td>
</tr>
<tr>
<td>Leukemia</td>
<td>Chernoff</td>
<td>98.94±1.16</td>
<td>99.72±0.62</td>
<td>98.99±0.62</td>
</tr>
<tr>
<td>Prostate</td>
<td>Linear</td>
<td>96.86±0.82</td>
<td>98.24±1.07</td>
<td>95.88±1.07</td>
</tr>
<tr>
<td>Ovary</td>
<td>Regression</td>
<td>99.76±0.35</td>
<td>100±0</td>
<td>100±0</td>
</tr>
<tr>
<td>Lungcancer</td>
<td></td>
<td>99.45±0</td>
<td>100±0</td>
<td>100±0</td>
</tr>
</tbody>
</table>

### 5.6 Summary

In this chapter, a novel two stage approach for gene selection has been discussed. Since different gene ranking methods may provide diverse subsets of informative genes, in first stage union of informative genes selected by different gene ranking methods is considered. This reduces chance of missing informative genes. Using union of different gene ranking methods at first stage gives a better set of informative genes as compared to using a single gene ranking method. But, this may contain redundant genes as ranking methods do not take into account the relationship between different genes. In second stage, three different measures (Kullback Divergence measure, Chernoff Distance measure and Linear Regression) are studied to remove redundancy. From experimental results on six microarray datasets, it is observed that none of the redundancy reduction methods seem to be clear winner. However, JC outperforms other
in most of the cases. It is also observed that maximum classification accuracy is achieved with SVM and KNN classifiers in conjunction with JC for all datasets. It is also observed that performance of KNN is better than all other classifiers in terms of number of genes required to achieve maximum accuracy. To decrease the computation time a recursive definition to compute redundancy measures: Chernoff Distance and Kullback Diversion is also proposed.