Peridocity, Change Detection and Prediction in Microarrays

Chapter 1

Improved Class Prediction in Gene Expression Microarray Data

1.1 INTRODUCTION

The advancement of cDNA microarrays and high-density oligoneucleotide chips in biotechnology has drawn much interest of statistical analysis in cancer research. One of the primary areas of focus is the classification of tumors using the gene expression data. A better understanding of the molecular variation among the tumors can be studied thanks to the possibility of simultaneously analyzing thousands of gene expression profiles. However, the fruitful endeavor for this understanding depends on the selection of proper statistical approach.

Dudoit *et al.* (2002) provided an extensive comparison of different classification methods. For the implementation of most of the methods, there needs to be an initial gene selection to make the number of genes to be less than the number of samples. Although their analyses show that the diagonal *linear discriminant analysis* (DLDA) maintains one of the top-ranking classifiers, the implementation of this method to other data sets does not seem to be appealing (Fort and Lambert-Lacroix, 2005). The big challenge of dealing with the microarray data is that the number of covariates is in thousands whereas the number of samples is usually not more than one hundred. Similar to regression method, the traditional discriminant analysis methods are not efficient in such situation. The method with principal components, partial least squares, ridge regression with their penalized forms are discussed in several articles as ways to solve the problem of classification (Ghosh, 2003; Fort and Lambert-Lacroix, 2005).

Nearest neighbour algorithm is one of the most frequently used techniques in classification problem. This algorithm is also known as instance-based learning. Holmes and Adams (2003) proposed a method which takes into account multiple nearest neighbors as a set of covariates in contrast to traditional method where only single nearest neighbor is selected on the basis of cross-validation error rate. The authors also proposed that the optimization of k can be done by maximum pseudolikelihood instead of using cross-validation for misclassification rate. In a logistic regression setting, the theory is flexible as it can take the original covariates as well as multiple nearest neighbor covariates (NNC). Original covariates capture the linear effects and multiple NNC capture nonlinear effects present at different scales within the data. The presence of thousands of genes as covariates will lead to a problem in variable selection in their method. This is because the traditional step-wise regression will no longer be feasible in such circumstances. Although the procedure can be reformed in terms of tens of genes selected by some procedure, the performance of classification method depends on initial gene selection process (Lee *et al.*, 2005). Also, may researchers feel it is best to include as many genes as possible and are reluctant to use subset approaches (Guo *et al.*, 2007).

Fort and Lambert-Lacroix (2005) put their suggestion against using k-nearest neighbor method for some of the data sets due to many occurrences of indecision. Still the analysis shows that the performance of this method is much better than many other methods (Dudoit *et al.*, 2002). The presence of high positive correlation of the gene expression observations within the same group and high negative correlation between different groups brings about the nearest neighbor classifier to perform as a good classifier in several data sets.

The estimation of regularized parameters involved in any model can be performed in several ways. *Bayesian Information Criterion* (BIC) is one of the popular criteria in selecting best model. Subset selection is highly variable as it is a discrete process, which either takes a variable or discard it (Hastie *et al.*, 2001). Tibshirani (1996) proposed *Least Absolute Shrinkage and Selection Operator* (LASSO), that shrinks some regression coefficients and sets other to zero, and thus works as a variable selection method. The $L1$ lasso penalty can be used in logistic regression framework when we have quite a large number of covariates. However, we found that the misclassification rate gets higher when all the nearest neighbor covariates are included in variable selection stage.

Nguyen and Rocke (2002) used partial least squares method for the purpose of classification in gene expression data. Recent methods include *Support Vector Machine* (SVM) and *Shrunken Centroid Regularized Discriminant Analysis* (SCRDA) (Hastie *et al.*, 2001; Guo *et al.*, 2007). SVM works in classification by producing linear boundary in the feature space and thus refers to non-linear boundary in input space. SCRDA is an extension of Fisher Linear Discriminant Analysis. This solves the non-singularity problem and provides a gene selection during the process.

Including NNC prior to running any of the method gives an augmented form. This provides some extra information to the classifier. First NNC can lead to capture non-linear relationship which might be ignored otherwise. Thus an improved version of many sophisticated methods can be achieved using this kind of augmentation.

1.2 METHODS

Suppose that we have expression levels for p genes over a size of n samples. The data matrix is given by $X = (x_{ij})$, a matrix of dimension $n \times p$. The value x_{ij}

refers to the expression level for j-th gene in i-th sample. The response variable is a categorical variable taking values as $y_i = \{A_1, A_2, \ldots, A_g\}$, where g is the number of classes. In the present work we discuss only two-class prediction problem. Hence we can express y_i as taking values $\{-1, 1\}$. Predictions are built on the training set and the performances are evaluated using the test set. In an one-leave-out validation process, successively all but one observations are considered as training set and the error rate is measured.

1.2.1 K-Nearest Neighbor

The nearest neighbor method is based on the distance function; for example, correlation or Euclidian distance for pairs of observations. In a k-nearest neighbor method, predictions of new observations are made through the training set $\{y_i, x_i\}$ for $i = 1, 2, \ldots, n$. For a new observation, we find the k closest observations in the training set and then predict the class to be the one where the majority of the k-neighbours belong to. The process is run for each specified values of k and then the selection of k is done using cross-validation. However, Holmes and Adams (2003) proposed a new method for finding optimum value of k. Instead of using cross-validation method, optimum value of k is derived by maximizing pseudolikelihood from a logistic regression.

After the initial selection of a number of genes, say P, we have our set of variables as $\{x_1, x_2, \ldots, x_p\}$. Corresponding to the i-th observation, k-nearest neighbor autocovariate is defined as:

$$\nu_{i(k)}(A_1) = \frac{1}{k} \sum_{j \sim i} [I(y_j = A_1) - I(y_j = A_0)] \tag{1.1}$$

The indicator variable $I(x = \omega)$ takes the value 1 if $x = \omega$ and 0 otherwise; $\sum_{j \sim i}$ denotes that the summation is over the k-nearest neighbors of x_i in the set $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$. The autocovariate $\nu_{i(k)}$ refers to the proportion of class A_1's to class A_0's within the k nearest neighbors of x_i. Therefore, if all the k nearest neighbors of x_i are in A_1 then the autocovariate $\nu_{i(k)}$ is 1; if all the k nearest neighbors of x_i are in A_0 then the autocovariate $\nu_{i(k)}$ is 0. Then a logistic regression model containing the covariates $\nu_{i(k)}$ can be written as

$$\Pr(y_i = A_1) = \eta_i = \frac{\exp(\alpha_k \nu_i(k))}{1 + \exp(\alpha_k \nu_i(k))} \tag{1.2}$$

The pseudolikelihood function is therefore,

$$L(\alpha_k; \nu_{(k)}) = \prod_{i=1}^{n} \eta_i^{\tilde{y_i}} (1-\eta)^{1-\tilde{y_i}} \tag{1.3}$$

where

$$\tilde{y_i} = \begin{cases} 0, & \text{if } y_i = A_0 \\ 1, & \text{if } y_i = A_1 \end{cases} \tag{1.4}$$

Optimal value of k is selected by maximizing the likelihood function. That is,

$$\hat{k} = \text{argmax}\,_k L(\alpha_k; \nu_{(k)}) \tag{1.5}$$

1.2.2 DLDA and DQDA

Let $f_k(x)$ be the conditional density of $\mathtt{x}$ in class $y = A_k$ and assume that this follows multivariate normal distribution of the form:

$$\mathtt{x}|y = A_k \sim \text{MVN}\left(\mu_k, \Sigma_k\right)$$

Let π_k be the prior probability of class k. Then the discriminant function is expressed as

$$\text{Ł}_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log \pi_k \tag{1.6}$$

This is called quadratic discriminant function as it does not assume equal covariances throughout the classes. In a two class setting, the decision boundary between two classes can be given by a quadratic equation $\{x : \text{Ł}_1(x) = \text{Ł}_2(x)\}$. If the class density has diagonal covariance matrix of the form $\Sigma_k = \text{diag}\left(\sigma_{k1}^2, \sigma_{k2}^2, \ldots, \sigma_{kP}^2\right)$, then the discrimination rule is called *diagonal quadratic discriminant analysis* (DQDA).

If we assume that class density has same covariance matrix for all the classes; that is if $\hat{\Sigma}_k = \hat{\Sigma}$, this leads to *linear discriminant analysis* (LDA). When covariance matrix in LDA is diagonal of the form $\Sigma = \text{diag}\,(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)$, then this is called *diagonal linear discriminant analysis* (DLDA).

We predict an observed value x_0 to a class which maximizes the discriminant function in Equation 1.6; that is, $y(x) = \text{argmax}_{\,k}\text{Ł}_k(x)$.

1.2.3 Shrunken Centroid RDA

Shrunken Centroid Regularized Discriminant Analysis (SCRDA) was introduced by Guo *et al.* (2007). This is a modified version of LDA. After estimating the parameters, we can write the discriminant function from equation 1.6 as:

$$\text{Ł}_k(x) = x^T \hat{\Sigma}^{-1} \bar{x}_k - \frac{1}{2} \bar{x}_k^T \hat{\Sigma}^{-1} \bar{x}_k + \log \pi_k \tag{1.7}$$

where $\bar{x}_k$ represents the mean vector in k-th class. In high-dimensional setting, the estimates in LDA will be unstable and therefore cannot provide optimal results (Guo *et al.*, 2007). In order to overcome the singularity problem in such situation, the authors proposed using regularized form of the covariance estimate:

$$\tilde{\Sigma} = \alpha \hat{\Sigma} + (1 - \alpha) I_P \tag{1.8}$$

where α is a non-negative value in the range $0 \leq \alpha \leq 1$. Using the equation 1.8, we can redefine the discriminant function as:

$$\tilde{\text{Ł}}_k(x) = x^T \tilde{\Sigma}^{-1} \bar{x}_k - \frac{1}{2} \bar{x}_k^T \tilde{\Sigma}^{-1} \bar{x}_k + \log \pi_k \tag{1.9}$$

Then the SCRDA can be constructed as classifying an observation x in a group that minimizes:

$$(x - \bar{x}'_{k'})^T \tilde{\Sigma}^{-1} (x - \bar{x}'_{k'}) - \log \pi_{k'} \tag{1.10}$$

where $\bar{x}'_{k'}$ is the vector of shrunken centroid for group k. A shrunken centroid $\bar{x}'$ is defined as

$$\bar{x}' = \text{sgn}\,(\bar{x})(|\bar{x} - \Delta)_+ \tag{1.11}$$

Instead of shrinking centroid $\bar{x}$, one can shrink $\tilde{\Sigma}^{-1}\bar{x}$.

Two methods were proposed to estimate the tuning parameter pair (α, Δ); however, we use the ”Min-Min” rule in the analysis. The first step is to find all the pairs that yield minimum cross-validation error in training set. Finally, optimum pair of (α, Δ) refers to that values which correspond to minimum number of selected genes.

1.2.4 Support Vector Machine

The space that $\mathtt{x} = \{x_1, x_2, \ldots, x_p\}$ takes is called input space. A space obtained after transforming $\mathtt{x}$ to $\tau(\mathtt{x})$ is called feature space. An SVM is a technique that separates classes through non-linear boundary by creating linear boundary in transformed feature space.

Let $u'x+a = 0$ is the separating hyperplane between the groups. There exists two other bounds - the distance between which is sought to be maximum for separating the classes. This distance is called margin and denoted as $m = \frac{1}{||u||}$. We can define the decision boundary through the optimization problem

$$\begin{gathered}\text{minimize } \tfrac{1}{2}||u||^2 \\ \text{subject to } y_i(u'x+a) \geq 1 \text{ or } 1 - y_i(u'x+a) \leq 0\end{gathered}$$

The Lagrangian is

$$L = \frac{1}{2}u'u + \sum_{i=1}^{n} \alpha_i(1 - y_i(u'x_i + a)) \tag{1.12}$$

where α_i is Lagrange multiplier. Setting gradient of L w.r.t u and a to zero and then substituting $u = \sum_{i=1}^{n} \alpha_i y_i x_i$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$, we get

$$L = -\frac{1}{2}\sum\sum \alpha_i \alpha_j y_i y_j x_i' x_j + \sum \alpha_i \tag{1.13}$$

Therefore, the optimization problem becomes

$$\begin{gathered}u(\alpha) = -\tfrac{1}{2}\sum\sum \alpha_i \alpha_j y_i y_j x_i' x_j + \sum \alpha_i \\ \text{subject to } \alpha_i \geq 0 \text{ and } \sum \alpha_i y_i = 0\end{gathered}$$

However, if the classes overlap in feature space, there will arise some non-negative slack variables $\xi = \{\xi_1, \xi_2, \ldots, \xi_n\}$. Thus, we get modified optimization problem as

$$\text{minimize } \tfrac{1}{2}||u||^2 + c\textstyle\sum_{i=1}^{n} \xi_i$$
$$\text{subject to } y_i(u'x_i + a) \geq 1 - \xi_i,\ \xi_i \geq 0$$

where c is tradeoff parameter between error and margin. This corresponds to

$$u(\alpha) = -\tfrac{1}{2} \textstyle\sum\sum \alpha_i\alpha_j y_i y_j x_i' x_j + \sum \alpha_i$$
$$\text{subject to } c \geq \alpha_i \geq 0,\ \textstyle\sum \alpha_i y_i = 0$$

As mentioned before, linear operation in the feature space is equivalent to non-linear operation in input space. Thus we reach to another SVM optimization problem through substituting the inner product $x_i'x_j$ by

$$K(x_i, x_j) = \tau(x_i)'\tau(x_j) \qquad (1.14)$$

There are different types of kernals for SVM optimization; however the popular ones (Hastie *et al.*, 2001) are:

- Radial basis function kernel with width σ:
  $K(x_i, x_j) = \exp(-||x_i - x_j||^2/2\sigma^2)$
- Polynomial kernal with degree l: $K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^l$
- Neural network: $K(x_i, x_j) = \tanh(\kappa_1\langle x_i, x_j\rangle + \kappa_2)$

1.3 IMPLEMENTATION

For each nearest neighbour $\{K_1, K_2, \ldots, K_l\}$, we can obtain the covariates as $\nu_{(K_1)}, \nu_{(K_2)}, \ldots, \nu_{(K_l)}$. Then the augmented set of inputs would be $\{x_1,\ x_2,\ \ldots,\ x_p,\ \nu_{K_1},\ \nu_{K_2},\ \ldots,\ \nu_{K_l}\}$. After inclusion of unit column vector, the design matrix is of the form $D = (1, X, V)$. We found that implementation of all the covariates in V lead to high misclassification rate. In such case, any method picks some unnecessary covariates that deters the optimization of the classification rate. Practical implementation reveals that 1-NN can provide good result in bioinformatic applications. In present work, we investigate the performance of four methods; namely DLDA, DQDA, SVM and SCRDA when first NNC is added to the original set of inputs.

1.3.1 Assessing Prediction Accuracy

Cross-validation is a simple but widely used method for assessing prediction accuracy. In a K-fold cross-validation, we randomly divide the data into K segments. We leave one part out, say j-th part, and fit the model for the remaining parts. Then estimate the error rate for that j-th part. We repeat the process for each of K segments, and finally find the overall misclassification error. In our analysis, we use $K = N$ which leads to *leave-one-out* (LOO) cross validation. We also perform re-randomization analysis. In this case we randomly divide the data into learning and validation part. The size of the validation part is considered as one fifth of the total sample size. A model is tuned from the learning part and prediction error is estimated from validation part. We repeat the process for 300 times and find overall error rate.

1.3.2 Computation

We use Beowulf cluster computing environment with 58 nodes for doing all the analyses. Yu (2002) developed the package `Rmpi`, which is an interface to *Message Passing Interface* (MPI). This package allows to implement R codes cooperatively in parallel across multiple machines. Some of the microarray data sets are very large and so running the leave-one-out or re-sampling procedure demands lots of computation time. We enjoy very good computational savings using this Beowulf cluster computing facility.

1.4 SIMULATION RESULT

To discuss the motivation of proposed method, we use a simulated data set. The concept of this simulation is similar to what was discussed by Guo *et al.* (2007) as two-group dependent structure. We assume that the conditional densities of $\mathtt{x}$ in two classes are $\text{MVN}\,(\mu_1, \Sigma_1)$ and $\text{MVN}\,(\mu_1, \Sigma_1)$. There are $P = 2000$ input variables. The mean, μ_1, for first group is $P \times 1$ vector of elements 0. The mean vector in another group has first 100 elements as 0.5 and rest 1900 as 0.

The covariance for both groups is block diagonal but with different block sizes. Both the densities have covariance structure as:

$$\begin{pmatrix} \Sigma_\rho & 0 & 0 & \cdots & \cdots & \cdots \\ 0 & \Sigma_{-\rho} & 0 & 0 & \cdots & \vdots \\ 0 & 0 & \Sigma_\rho & 0 & \cdots & \vdots \\ \vdots & 0 & 0 & \Sigma_{-\rho} & 0 & \vdots \\ \vdots & \vdots & \vdots & 0 & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \tag{1.15}$$

Each block in the covariance matrix has autoregressive form. If ρ is autocorrelation between successive genes, and the block size is B then Σ_ρ can be written as:

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{B-1} \\ \rho & 1 & \rho & \cdots & \rho^{B-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{B-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{B-1} & \rho^{B-2} & \cdots & \cdots & 1 \end{pmatrix} \tag{1.16}$$

We consider same autocorrelation, $\rho = 0.9$, for both groups; but take different block sizes $B = 40$ and $B = 100$. Training set contains 100 observations from each class. To evaluate the performance, 500 test samples are generated from each group using same procedure.

We see from Table 1.1 that all methods are hugely improved through the use of NNC. SCRDA gained the most improvement as the error rate decreased from 25.8% to only 11.1%. This augmentation turned SCRDA to be the best performing method in this data set. The gain in SVM is minimum.

Performance of different methods with and without augmented covariates

Method	k=0	k=1
DLDA	27.8	21.5
DQDA	28.9	24.7
SVM	19.3	16.7
SCRDA	25.8	11.1

Table 1.1: Misclassification rate for different methods in simulated data. All the rates are measured in percentage. Here $k = 0$ refers to original set of covariates, and $k = 1$ refers to one NNC augmented to the original set. A total of 200 training samples and 1000 test samples, measuring $P = 1000$ variables, are generated.

1.5 MICROARRAY DATA SETS

We assess the proposed method using four publicly available data sets. The data sets are (i) colon cancer data (Alon *et al.*, 1999), (ii) acute leukemia data (Golub *et al.*, 1999), (iii) prostate cancer data (Singh *et al.*, 2002) and (iv) breast cancer data (vant Veer it et al., 2002). An overview of the data sets is given in Table 1.2. All of the data sets were either originally divided into groups of training and test sets, or by the aforementioned authors. However, for an extensive comparison we merge all the training and test samples and thereafter find leave-one-out as well as re-sampling error rates.

Summary of the microarray data sets used in the analysis.

Name	Description	P	n_1	n_2
Alon	Colon cancer	2000	40	22
Golub	Acute leukemia	7129	47	25
Singh	Prostate cancer	12600	59	77
Veer	Breast cancer	24188	51	46

Table 1.2: Summary table of four data sets that we analyze to evaluate the performance of proposed method. P refers to the number of genes in corresponding data. n_1 and n_2 are the number of samples available for class 1 and 2 respectively.

1.5.1 Colon Cancer Data

This data set contains 62 tissue samples with 40 tumor and 22 normal samples (Alon *et al.*, 1999). An Affymetrix oligonucleotide array complementary to more than $6,500$ human genes was used to analyze expression levels for these samples. Finally 2000 genes are finally included in the data, which are not readily preprocessed. We follow the pre-processing steps mentioned by Dudoit *et al.* (2002):

- thresholding at floor of 100 and ceiling of 16000,
- filtering to exclude the genes with $\max / \min \leq 5$ and $(\max - \min) \leq 500$
- transformation using logarithm of base 10.

1.5.2 Acute Leukemia Data

Acute leukemia data set contains 72 bone marrow samples obtained from adults with acute leukemia (Golub *et al.*, 1999). Expression levels for 7129 genes are measured using Affymetrix high-density oligonucleotide arrays. There are 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of myeloid leukemia (AML). The data is not preprocessed and so same procedure as that of Colon data is applied here.

1.5.3 Prostate Cancer Data

In this data set total of 12600 gene expression levels are measured for 136 tissue samples (Singh *et al.*, 2002). Expression profiles were derived from 77 prostate tumors and 57 nontumor prostate samples from patients undergoing surgery. The objective here is to separate tumor tissues from normal tissues. The pre-processing steps mentioned by Singh *et al.* (2002) are applied to the data set (Fort and Lambert-Lacroix, 2005):

- thresholding at floor of 10 and ceiling of 16000,
- filtering to exclude the genes with $\max / \min \leq 5$ and $(\max - \min) \leq 50$.

- transformation using logarithm of base 10 is used.

1.5.4 Breast Cancer Data

The data contains 24188 expression profiles for 97 breast cancer patients. They are divided into two groups - (i) who developed metastases within 5 years and (ii) who remained disease-free within 5 years (vant Veer it et al., 2002). 46 patients developed distant metastases and 51 did not. The objective is to predict the presence of subclinical metastases in order to provide a strategy to select patients who would benefit from adjuvant therapy. The data set is preprocessed and so no further preprocessing step is applied.

1.6 Gene Selection

Generally, selecting a subset of best differential genes provides better classification result (Fort and Lambert-Lacroix, 2005) for different methods. Lee *et al.* (2005) compared different classification methods for three different types of initial gene selection. It was showed that process of initial gene selection makes difference in the performance. We use a criterion that is based on ratio of between to within group sum of squares of the genes (Dudoit *et al.*, 2002). The ratio for gene j is

$$\frac{\mathrm{BSS}\,(j)}{\mathrm{WSS}\,(j)} = \frac{\sum_i \sum_l I(y_i = l)(x_{lj} - \bar{x}_{.j})^2}{\sum_i \sum_l I(y_i = l)(x_{ij} - \bar{x}_{lj})^2}$$

where $\bar{x}_{.j}$ is the average expression level of gene j across all samples and $\bar{x}_{1j}$ is the average expression level of gene j across samples in class l. A selection of P genes are made by considering the genes having largest BSS /WSS ratios. Although SCRDA can automatically select the genes during the process, we use BSS /WSS criterion to select primarily 1000 genes for comparison with other methods.

Performance of different methods with and without augmented covariates

		LOO		OS	
	Data set	$k=0$	$k=1$	$k=0$	$k=1$
DLDA	Alon	12.90	12.90	13.72	13.66
	Golub	4.17	1.39	2.81	2.28
	Singh	29.41	28.68	28.43	27.83
	Veer	32.99	32.99	32.14	32.07
DQDA	Alon	12.90	12.90	14.25	14.13
	Golub	1.39	1.39	1.95	1.90
	Singh	36.76	36.76	36.17	36.03
	Veer	31.96	30.92	29.04	28.56
SVM	Alon	12.90	12.90	14.72	14.72
	Golub	1.39	1.39	1.59	1.59
	Singh	5.88	5.88	7.22	7.07
	Veer	30.93	30.93	31.17	31.14
SCRDA	Alon	12.90	9.68	13.87	13.67
	Golub	6.94	2.78	6.03	5.53
	Singh	8.38	5.15	6.01	5.87
	Veer	33.84	33.60	30.88	30.41

Table 1.3: *Leave-one-out* (LOO) and *out of sample* (OS) misclassification rates (in %) of different methods with and without the augmented nearest neighbour covariates (NNC). Here $k = 0$ refers to no NNC and $k = 1$ refers to first NNC included in the initial covariate set. A selection of best 1000 genes was made for the comparison.

1.7 CONCLUSION

We have discussed the plausibility of using a modified classification procedure to improve prediction accuracy in existing methods. Performance of the approach was evaluated through one simulated and four real data sets. The method is flexible and provides better results in most situation.

The simulation was constructed such a way that the decision boundary between two classes is non-linear. It was found that all methods got substantial improvement through the use of NNC approach. Table 1.3 demonstrates the misclassification error rate using different methods. We see from the result of leave-one-out cross-validation that the classification accuracy improves in almost all methods. SCRDA experiences greatest gain in prediction for most of the data sets. Moreover, the application of re-sampling technique shows that some systemic decrease in misclassification rate can be gained through the use of first NNC.

Investigation showed that some other dimension reduction techniques; for example, *Principal component regression* (PCR) or *partial least squares regression* (PLSR) with augmented NNC can provide very good result. This approach can be extended to any classification rule for plausible improvement.

1.8 FUTURE WORK

We will extend this approach to study the performance in multi-class problem. The procedure can take multiple number of *nearest neighbour covariates* (NNC). We will develop some adaptive selection procedure for the optimal number of NNC to be finally added in the model.

Bibliography

Alon, A., Barkai, N., Notterman,D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci.*, **96**, 6745-6750.

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, **97**, 7787.

Efron, B., Hastie, T., Johnstone, I.M. and Tibshirani, R. (2002). Least Angle Regression. *Technical report, Department of Statistics, Stanford University.*

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21** , 1104-1111.

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, **59**, 992-1000.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286 (5439)**, 531-537.

Guo, Y., Hastie, T., Tibshirani, R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics.* **8**, 1, 86-100.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* New York: Springer-Verlag.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A. and Trent, J. (2001). Gene-Expression Profiles in Hereditary Breast Cancer, *The New England Journal of Medicine*, **344**, 539-548.

Holmes, C. C. and Adams N. M. (2003). Likelihood inference in nearest-neighbour classification models. *Biometrika*, **90**, 1, 99-112.

Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., Takao, T., Tamesa, T., Tangoku, A., Tabuchi, H., Hamada, K., Nakayama, H., Ishitsuka, H., Miyamoto, T., Hirabayashi, A., Uchimura, S. and Hamamoto, Y. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *The Lancet*, **361**, 923-929.

Lee, J. W., Lee, J. B., Park, M. and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, **48**, 869-885.

Nguyen,D. and Rocke,D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18(1)**, 39-50.

Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R. and Louis, D. N. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, **63(7)**, 1602-1607.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., Zhao, Y. (2004). *Design and Analysis of DNA Microarray Investigations.* New York: Springer.

Singh, D., Febbo, P., Ross, D., Jackson, G., Manola, J., Ladd, C., Tamayo, A., Renshaw, A., DAmico, A. V., Richie, J., Lander E., Loda, M., Kantoff, P., Golub, T., Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell.* **1**, 203209.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, **58**, 1, 267-288.

vant Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy,K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*, **415**, 530-536.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. In Gani, J. (ed.), *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett.* Academic Press, London. 117-142.

Yu, H. (2002). Rmpi: Parallel Statistical Computing in R. *R News*, **2(2)**, 10-14.