Selection bias in working with the top genes in supervised classification of tissue samples

X. Zhu [a,b], C. Ambroise [c], G.J. McLachlan [a,b,d,*]

[a] *ARC Centre for Bioinformatics, Institute for Molecular Bioscience, University of Queensland (UQ), St. Lucia, Brisbane, Australia 4072*

[b] *Department of Mathematics, UQ, St. Lucia, Brisbane, Australia 4072*

[c] *U.T.C U.M.R.C. N.R.S. 6599 Heudiasyc, Centre de Recherches de Royallieu, B.P. 20529, F-60205 Compiegne Cedex*

[d] *ARC Special Research Centre for Functional and Applied Genomics, UQ*

Abstract

Currently there is much interest in using microarray gene-expression data to form prediction rules for the diagnosis of patient outcomes. A process of gene selection is usually carried out first to find those genes that are most useful according to some criterion for distinguishing between the given classes of tissue samples. However, there is a bias (selection bias) introduced in the estimate of the final version of a prediction rule that has been formed from a smaller subset of the genes that have been selected according to some optimality criterion. In this paper, we focus on the bias that arises when a full data set is not available in the first instance and the prediction rule is formed subsequently by working with the top-ranked genes from the full set. We demonstrate how large the subset of top genes must be before this selection bias is not of practical consequence.

Key words: gene selection, support vector machine, error rates, cross-validation, selection bias

1 Introduction

High-density DNA microarray technology allows researchers to monitor the interactions among thousands of gene transcripts in an organism on a single

* Corresponding author.
Email addresses: j.zhu@imb.uq.edu.au (X. Zhu), ambroise@utc.fr (C. Ambroise), gjm@maths.uq.edu.au (G.J. McLachlan).

experimental medium, which is often a glass microscope slide or nylon membrane. Prior to the computerization and miniaturization of this technology, researchers were limited to examinations of much smaller numbers of genetic units per experiment and were only able to assess interactions among genes under changing conditions on a much smaller scale. Given the availability of microarray data, there is increasing interest in changing the emphasis of tumor classification from morphologic to molecular. In this context, the problem is to construct a discriminant (prediction) rule $r(\boldsymbol{y};\, \boldsymbol{t})$ that can accurately predict the class of origin of a tumor tissue with feature vector $\boldsymbol{y}$, which is unclassified with respect to a known number $g\,(\geq 2)$ of distinct tissue classes, denoted here by $C_1, \ldots, C_g$. The vector

$$\boldsymbol{t} = (\boldsymbol{y}_1^T, \boldsymbol{z}_1^T, \ldots, \boldsymbol{y}_n^T, \boldsymbol{z}_n^T)^T, \tag{1}$$

denotes the training data, where

$$\boldsymbol{z}_j = (z_{1j}, \ldots, z_{gj})^T$$

is the class-indicator vector, and z_{ij} is one or zero according as $\boldsymbol{y}_j$ comes from the ith class C_i or not $(i = 1, \ldots, g;\ j = 1, \ldots, n)$.

Here the feature vector $\boldsymbol{y}$ contains the expression levels on a very large number N of genes (features); that is, $\boldsymbol{y}$ is the expression signature vector of a tissue. In applications concerned with the diagnosis of cancer, one class C_1 may correspond to cancer and the other (C_2) to benign tumors. In applications concerned with patient survival following treatment for cancer, one class (C_1) may correspond to the good-prognosis group and the other C_2 to the poor-prognosis group. Also, there is interest in the identification of "marker" genes that characterize the different tissue classes. This is the feature selection problem.

The data set at hand as described above consists of N genes, but usually it is a subset of a much larger set containing the expression levels of N_T genes over the tissue samples. For example, N_T might be of the order of tens of thousands, while N will be of the order of thousands. The subset of N genes is usually obtained by applying some ad hoc filtering process to the N_T genes. Having reduced the full data set down to N genes, typically some finer form of gene selection is employed to reduce this subset down further to a much smaller number N_o for the formation of a discriminant rule. A consequence of basing the final form of the discriminant rule on a small subset of the genes selected in some optimal way is that a selection bias has to be allowed for in the estimation of the generalization error of the rule based on the optimal subset of N_o genes. Otherwise, a false overoptimistic impression will be obtained for the discriminatory power of the rule. This bias has often been overlooked in the bioinformatics literature [2]. Also, this bias arises in an unsupervised context

(cluster analysis) with tests and plots on the number of clusters.

Typically in practice, N is not very small relative to the total number of genes N_T, so the only selection bias we have to worry about is that incurred in going from N to N_o genes. However, if N is very small relative to N_T, then there will be a selection bias in working with the N genes and not the total number N_T or a much larger number than N if the total number N_T is not used. If we have access to the full set of N_T genes and know how it was reduced to N genes in the first instance, then we can correct for the bias in working with only the N genes. But if it happens that the available information is limited to only the N genes, then we will be unable to correct for any possible bias is not working with the full set of genes N.

In this paper, we demonstrate the bias in working with only N genes and not the full set of N_T genes, where the N genes are the top N ranked genes among the total number N_T of genes for varying sizes of N. It will be seen that this bias is of practical significance when N is very small relative to N_T.

Before we proceed to these examples, we shall describe some methods of gene selection and how the selection bias can be corrected for via cross-validation if given the full data set.

2 Need for Gene Selection

In a standard discriminant analysis, the number of training observations n is usually much larger than the number of feature variables p. But in the present context of microarray data, the number of tissue samples n is typically between 10 and 100, and the number of genes ($p = N$) is in the thousands. This presents a number of problems. Firstly, the discriminant rule $r(\boldsymbol{y}; \boldsymbol{t})$ may not be able to be formed using all the available genes. For example, the pooled within-class sample covariance matrix $\boldsymbol{S}$ required to form Fisher's linear discriminant function is singular if $n < g + p$, where g is the number of classes. Secondly, even if all the genes can be used as, say, with the nearest-centroid rule or a support vector machine (SVM), the use of all the genes may allow the noise associated with genes of little or no discriminatory power, to inhibit and degrade the performance of the rule $r(\boldsymbol{y}; \boldsymbol{t})$ in its application to unclassified data. That is, although the apparent error rate $A(\boldsymbol{t})$ (the proportion of the training tissues misallocated by $r(\boldsymbol{y}; \boldsymbol{t})$) will decrease as it is formed from more and more genes, its error rate in classifying tissues outside of the training set will eventually increase. That is, the generalization error of $r(\boldsymbol{y}; \boldsymbol{t})$ will be increased if it is formed from a sufficiently large number of genes. Hence, in practice, consideration has to be given to implementing some procedure for reducing the dimension of the feature vector of genes to be used in constructing

the rule $r(\boldsymbol{y}; \boldsymbol{t})$.

2.1 Some Methods

A common approach is to carry out a principal component analysis (PCA) and work with the leading components. The disadvantages of this approach are that the PCA does not take into account the class structure of the genes, and genes that show a large variation across the tissues may not be differentially expressed. Also, as the principal components are linear combinations of the original number of genes, biological interpretation of the components is not straightforward. One method that does take into account the class structure of the tissue samples in reducing the dimension of the feature space is partial least squares. However, it still suffers from the same interpretation difficulties as with principal components, as the components are linear combinations of all the genes. Nguyen and Rocke [11] demonstrated in their study that if the top genes for discrimination purposes were selected before performing the principal component analysis, then it would give similar results to partial least squares.

One common way of approaching the gene selection problem is to perform a preliminary ranking of genes on the basis of a fast computable criterion and then arbitrarily select a number of the best-ranked genes. Then either a discriminant rule is formed on the basis of these selected genes or further selection is undertaken before constructing the rule.

A commonly used criterion for ranking the individual genes $y_v = (\boldsymbol{y})_v$ ($v = 1, \ldots, p$) is the ratio of the between-class sum of squares to the within-class sum of squares,

$$F_v = (\boldsymbol{B})_{vv}/(\boldsymbol{W})_{vv}, \tag{2}$$

where $\boldsymbol{B}$ and $\boldsymbol{W}$ are the between and within sums of squares and products matrices, respectively. Under the null hypothesis that the vth gene has the same variance in each class, the statistic F_v has an F-distribution with $g-1$ and $n-g$ degrees of freedom. The use of (2) is equivalent to the likelihood ratio statistic $-2\log\lambda$ for the test of no differences between the means of the classes under the assumption of the homoscedastic model for the class-covariance matrices. Also, in the case of $g = 2$ classes, it is equivalent to the usual two-sample (pooled) Studentized t-statistic.

A further criterion is to rank the genes on the basis of the absolute values of their coefficients in the linear form of $r(\boldsymbol{y}; \boldsymbol{t})$ for an SVM formed with linear kernel. This is to be discussed further in the next section. There are also rules

where the ranking is being done implicitly in their construction; for example, nearest-shrunken centroids [12].

Another way to handle the problem of having to form a discriminant rule from a very large number of genes is to put the genes into groups either by some clustering method or by some supervised selection procedure that makes use of their known class labels. There is now a variety of ways proposed in the literature for the grouping of the genes. Having so grouped the genes, a discriminant rule can be formed from the genes (metagenes) selected to represent each group; see, for example, [9] and [10, Chapter 7].

3 Error-Rate Estimation

It is the conditional or actual error rates of $r(\boldsymbol{y};\, \boldsymbol{t})$ that are of central interest once the training data $\boldsymbol{t}$ have been obtained. We let $ec(\boldsymbol{t})$ denote the overall conditional error rate of $r(\boldsymbol{y};\, \boldsymbol{t})$ in its application to a new observation $\boldsymbol{y}$ subsequent to the training data $\boldsymbol{t}$. This error rate, which is conditional on the training data $\boldsymbol{t}$, also depends on the class-conditional distributions. But this dependence is suppressed here for simplicity of notation.

3.1 Apparent Error Rate

An obvious and easily computed nonparametric estimator of the conditional error rate $ec(\boldsymbol{t})$ of $r(\boldsymbol{y};\, \boldsymbol{t})$ is the apparent error rate $A(\boldsymbol{t})$ of $r(\boldsymbol{y};\, \boldsymbol{t})$ in its application to the observations in $\boldsymbol{t}$. That is, $A(\boldsymbol{t})$ is the proportion of the observations in $\boldsymbol{t}$ misallocated by $r(\boldsymbol{y};\, \boldsymbol{t})$. Thus we can write

$$A(\boldsymbol{t}) = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} Q[i, r(\boldsymbol{y}_j;\, \boldsymbol{t})], \tag{3}$$

where for any u and v, $Q[u, v] = 0$ for $u = v$ and 1 for $u \neq v$.

As the apparent rate is obtained by applying the rule to the same data from which it has been formed, it provides an optimistic assessment of the true conditional error rates. In particular, for complicated discriminant rules, overfitting is a real danger, resulting in a grossly optimistic apparent error. Although the optimism of the apparent error rate declines as n increases, it usually is of practical concern.

One way of avoiding the bias in the apparent error rate as a consequence of the rule being tested on the same data from which it has been formed (trained), is to use a holdout method as considered by Highleyman [8], among others. The available data are split into disjoint training and test subsets. The discriminant rule is formed from the training subset and then assessed on the test subset. Clearly, this method is inefficient in its use of the data. Indeed, it is not practical in the present context where the number of tissue samples (n) is so small relative to the number of genes. There are, however, methods of estimation, such as cross-validation, the Quenouille–Tukey jackknife, and the bootstrap of Efron [3], that obviate the need for a separate test sample. An excellent account of these three methods has been given by Efron [4], who has exhibited the close theoretical relationship between them.

The optimism arising from the use of the apparent error rate may be almost eliminated using cross-validation. The (leave-one-out) cross-validated estimate is given by

$$A^{(CV)}(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} Q[i, r(\boldsymbol{y}_j;\, \boldsymbol{t}_{(j)})], \tag{4}$$

where $\boldsymbol{t}_{(j)}$ denotes $\boldsymbol{t}$ with the point $(\boldsymbol{y}_j^T, \boldsymbol{z}_j^T)^T$ deleted ($j = 1, \ldots, n$). Hence before the sample rule is applied at $\boldsymbol{y}_j$, it is deleted from the training set and the rule recalculated on the basis of $\boldsymbol{t}_{(j)}$. This procedure at each stage can be viewed as the extreme version of the holdout method where the size of the test set is reduced to a single entity.

As remarked by Efron [5], cross-validation is often carried out, removing large blocks of observations at a time. Suppose, for example, that the training set is divided into, say q blocks, each consisting of m data points where, thus, $n = qm$ ($m \geq 1$). Let now

$$\boldsymbol{t}_{(k)} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_{(k-1)m}^T, \boldsymbol{y}_{km+1}^T, \ldots, \boldsymbol{y}_n^T)^T;,$$

that is, the training set after the deletion of the kth block of m observations. Then the q-fold cross-validated error rate is given by

$$A^{(CVq)}(\boldsymbol{t}) = \sum_{i=1}^{g}\sum_{j=1}^{m}\sum_{k=1}^{q} z_{ij} Q[i,\, r(\boldsymbol{y}_{(k-1)m+j}; \boldsymbol{t}_{(k)})]/n, \tag{5}$$

which requires only q recomputations of the rule. The choice of $q = n$ (leave-one-out) does not perturb the data enough and results in higher variance. In

the present context, this variance can be quite high. The values $q = 5$ or 10 are a good compromise.

3.3 External Cross-Validation

Caution has to be exercised in estimating the error rate of a discriminant rule formed by optimally selecting a small number of variables (genes) from a large set. This is because there will be a selection bias associated with choosing the optimal of a large number of possible subsets, regardless of the criterion used. We let $\boldsymbol{y}^{(s)}$ denote the subvector of $\boldsymbol{y}$ formed from the subset s of the full set of p variables, and let $r(\boldsymbol{y}^{(s)}; \boldsymbol{t}^{(s)})$ denote some arbitrary sample discriminant rule formed from the classified training data $\boldsymbol{t}^{(s)}$ on the subvector $\boldsymbol{y}^{(s)}$. Suppose that s_o defines the subset of feature variables of some specified size p_{s_o} that minimizes some criterion, say, $A^{(CV)}(\boldsymbol{t}^{(s)})$, over all possible $\binom{p}{p_{s_o}}$ distinct subsets s of size p_{s_o}. Although $A^{(CV)}(\boldsymbol{t}^{(s)})$ may be an (almost) unbiased estimator of the overall conditional error rate of the rule $r(\boldsymbol{y}^{(s)}; \boldsymbol{t}^{(s)})$, $A^{(CV)}(\boldsymbol{t}^{(s_o)})$ is obviously not providing an almost unbiased estimate of the error rate of $r(\boldsymbol{y}^{(s_o)}; \boldsymbol{t}^{(s_o)})$, as it is obtained by taking the smallest of the estimated error rates after they have been ordered according to their size. Here the (leave-one-out) cross-validated estimate is given by

$$A^{(CV)}(\boldsymbol{t}^{(s_o)}) = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} Q[i, r(\boldsymbol{y}_j^{(s_o)}; \boldsymbol{t}_{(j)}^{(s_o)})], \tag{6}$$

where $\boldsymbol{t}_{(j)}^{(s_o)}$ denotes the training data $\boldsymbol{t}^{(s_o)}$ with $(\boldsymbol{y}_j^{(s_o)^T}, \boldsymbol{z}_j^T)^T$ deleted.

In order to reduce the selection bias which is still present in the estimate (6), an external cross-validation should be performed whereby the selection process is undertaken for each deletion of a feature vector from the training set. This external cross-validated estimate of the overall error rate of $r(\boldsymbol{y}^{(s_o)}; \boldsymbol{t}^{(s_o)})$ is given by

$$A^{(CVE)}(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} Q[i, r(\boldsymbol{y}_j^{(s_{oj})}; \boldsymbol{t}_{(j)}^{(s_{oj})})], \tag{7}$$

where s_{oj} denotes the optimal subset, according to the adopted selection criterion applied to the training data $\boldsymbol{t}_{(j)}$ without $(\boldsymbol{y}_j^T, \boldsymbol{z}_j^T)^T$.

As the notation implies, the selected subset s_{oj} for the allocation of the jth entity may be different for each j ($j = 1, \ldots, n$). We can make use of this fact to identify potential marker genes. We can note the number of times a gene is chosen in the selected subset on each split of the training data during the external cross-validation [10, Chapter 7].

An illustration of this selection bias is to be given for the supervised classification of microarray data. But we first consider the support vector machine as it is the discriminant rule to be adopted in the sequel.

4 Support Vector Machine with Recursive Feature Elimination

Support vector machines are becoming increasingly popular classifiers for microarray data. Advantages of a support vector machine (SVM) in the present context, where the number of feature variables (genes) p is so large relative to the sample size n, are that it is able to be fitted to all the genes and that its performance appears not to be too affected by using the full set of genes. However, in practice, some form of gene selection would generally be contemplated. Another advantage of the SVM (with a linear kernel) is that gene selection can be undertaken fairly simply using the vector of weights as the criterion.

For an SVM with linear kernel, the rule $r(\boldsymbol{y}; \boldsymbol{t})$ can be written as

$$r(\boldsymbol{y}; \boldsymbol{t}) = \text{sign}\,(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{y}), \tag{8}$$

where $\hat{\beta}_v = (\hat{\boldsymbol{\beta}})_v$ denotes the coefficient of the expression level y_v for gene v.

As shown by Guyon et al. [7], a good guide to the relative importance of the genes in this SVM is given by the relative size of the absolute values of their fitted coefficients $\hat{\beta}_v$ (that is, the weights). Hence a ranking of the discriminatory power of the genes can be given by ranking the genes from top to bottom on the basis of the absolute values of the weights $\hat{\beta}_v$.

We consider here the selection procedure of Guyon et al. [7], who used a backward selection procedure, which they termed recursive feature elimination (RFE). It considers initially all the available genes, which are ranked according to their weights and the bottom-ranked genes discarded. The SVM is then refitted to the remaining genes, which are then reranked according to their new weights. Again, the bottom-ranked genes are discarded, and so on.

In the applications to follow on microarray data, we proceeded as in [7] and first discarded enough bottom-ranked genes so that the number retained was the greatest power of 2 (less than the original number of genes). We then

proceeded sequentially to discard half the current number of genes on each subsequent step. The error rate at any stage can be assessed by undertaking an external cross-validation as described above.

5 Example of Selection Bias Starting with All the Genes

Ambroise and McLachlan [2] investigated the magnitude of the selection bias and its correction for an SVM (with linear kernel) and Fisher's linear discriminant function in their application to two cancer data sets. We give in Figure 1 their results for the SVM applied to the colon data of Alon et al. [1]. They used Affymetrix oligonucleotide arrays to monitor absolute measurements on expressions of over 6,500 human genes in 40 tumor and 22 normal colon tissue samples. These samples were taken from 40 different patients, so that 22 patients supplied both a tumor and a normal tissue sample. Alon et al. [1] focused on the 2000 genes with highest minimal intensity across the samples.

For this illustration, we thus have $N = 2000$ and $N_T > 6500$. For these relative values of N and N_T, there would be little bias in working with the $N = 2000$ genes and not the full set of over 6500 genes. Thus we focus in this example on the selection bias incurred when the ordinary (internal) stratified cross-validated estimate (6) as used in Guyon et al. [7] is adopted instead of the external version (7) when a support vector machine with recursive feature elimination is applied.

To illustrate the size of the selection bias for the colon data set, Ambroise and McLachlan [2] split it into a training set and a test set, each of size 31, by stratied sampling without replacement from the 40 tumor and 22 normal tissues separately, so that each set contained 20 tumor and 11 normal tissues. The training set is used to carry out gene selection and to form the apparent error rate $A(\boldsymbol{t})$, the (leave-one-out) cross-validated error rate $A^{(CV)}(\boldsymbol{t})$ using just internal validation, and the external ten-fold cross-validated rate $A^{(CV10E)}(\boldsymbol{t})$ for a selected subset of genes. An unbiased error-rate estimate is given by the test error equal to the proportion of tissues in the test set misallocated by the rule. They calculated these quantities for 50 such splits of the colon data into training and test sets. The average values of the error-rate estimates are plotted in Figure 1. The error bars on the test error refer to the 95% confidence limits. The 0.632+ bootstrap error estimate, $B^{(0.632+)}$, was formed using $K = 30$ bootstrap replications for each of the 50 splits of a full training set. The latter estimate was proposed by Efron and Tibshirani [6] and first applied in the context of microarray data by Ambroise and McLachlan [2].

In Figure 1, the apparent error $A(\boldsymbol{t})$, the (leave-one-out) cross-validated error $A^{(CV)}(\boldsymbol{t})$, the external ten-fold cross-validated error $A^{(CV10E)}(\boldsymbol{t})$, the 0.632+

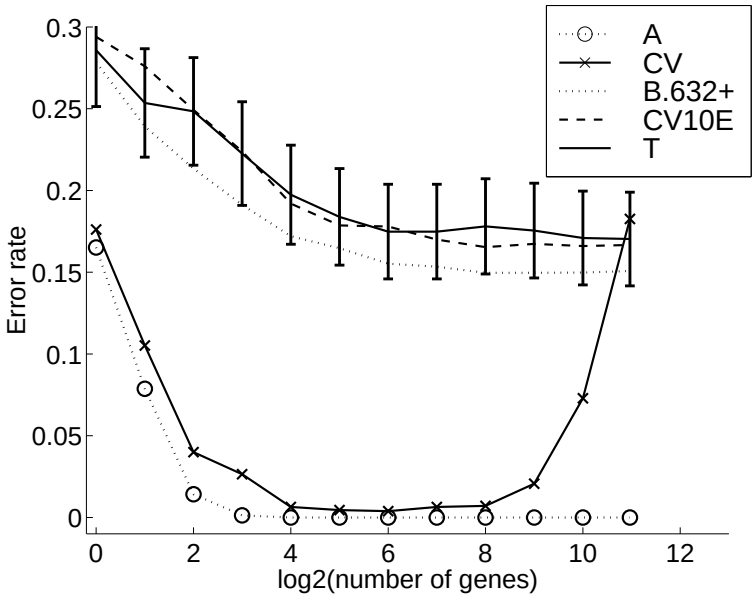

Fig. 1. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 62 colon tissue samples into training and test subsets of 31 samples each.

bootstrap error estimate $B^{(0.632+)}$, and the test error are denoted by A, CV, CV10E, B.632+, and T, respectively. It can be seen from Figure 1 that the true prediction error rate as estimated by T is not negligible, being above 15% for all selected subsets. The lowest value of 17.5% occurs for a subset of 2^6 genes at which the internal cross-validated rate (which ignores the selection bias) is zero.

6 Selection Bias in Not Working with the Full Set of Genes

6.1 Breast Cancer Data Set

We now demonstrate the selection bias that can occur when we work with only a small subset of the total number of genes. The data set considered concerns the breast cancer study of van 't Veer et al. [14]. They used inkjet synthesized oligonucleotide arrays to measure the expressions of 24881 genes in 98 primary breast cancers acquired from three groups of patients: 44 representing a good-prognosis group (that is, those who remained metastasis free after a period of more than 5 years), 34 from a poor-prognosis group (those who developed distant metastases within 5 years), and 20 representing a hereditary form of cancer, due to a BRCA1 (18 tumors) or BRCA2 (2 tumors) germline mutation. The 78 sporadic (non-BRCA) breast cancer patients were chosen specifically on the basis of their clinical outcome. van 't Veer et al. [14] applied a filter in

which only genes with both a P-value of less than 0.01 and at least a two fold difference in more than five out of the ninety-eight tissues for the gene were retained. This filter effectively reduced the initial set of genes to 4869. They subsequently were able to identify a set of 70 genes with expression profiles associated with the risk of early metastasis. This selection was carried out on the basis of the correlation between the gene expression profile and the class label, which is equivalent to using the (pooled) two-sample t-statistics; that is, using (2) in the case of $g = 2$. They called these 70 genes the prognostic marker genes.

We illustrate first the selection bias when we work with just these top 70 genes; that is, $N = 70$. We applied the same filtering process of van 't Veer et al. [14] to the 24881 genes but now just to the 78 tissue samples for the sporadic breast cancer tumours, which resulted in 5422 genes being retained. That is, we take this set of $N_T = 5422$ genes on the $n = 78$ sporadic breast cancer tumours to be our full data set and demonstrate the bias when we work with only the top $n = 70$ genes according to the criterion (2). We ignore here the bias in reducing the actual full set of 24881 genes by filtering to 5422 genes, but this bias is negligible which we did confirm.

What motivated us to examine the bias incurred in working with only the top 70 genes is that van de Vijver et al. [13] studied a larger series of breast cancer patients which consisted of 61 of the sporadic breast cancer 78 patients in the study of van 't Veer et al. [14], along with an additional 234 patients. But the gene expressions in their 295 tumour samples were made available only for the top 70 genes as defined above. Thus, if one wanted to work with the data set of van de Vijver et al. [13], there would be no option but to work with this very small reduced set of $N = 70$ genes, thereby incurring a selection bias that could not be corrected for since the expression levels on the full set of 25000 or so genes (or indeed any set other than the 70 genes) was not available.

6.2 Application of SVM with RFE to Breast Cancer Data

We applied the SVM with recursive feature elimination to the 78 tumour samples from $g = 2$ classes using just the $N = 70$ genes. At each stage of the feature elimination process, the overall error rate was estimated using ten-fold cross-validation. We performed the latter, using external cross-validation, but limited to correcting for the selection bias in choosing optimally a subset with fewer than 70 genes. That is, the top 70 genes were fixed during the validation process, and so it ignores the selection bias in working with the top 70 genes from the set of $N_T = 5422$ genes. In addition, we estimated the error rate where the external cross-validation is extended to correct also for the bias in working with the top 70 genes. This latter bias is corrected for by going back

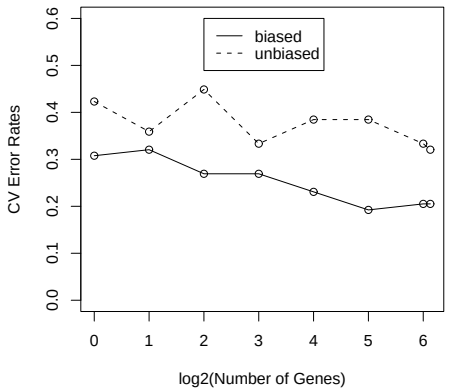

Fig. 2. The solid line is the ten-fold cross-validated error rate of the SVM with RFE applied to the top 70 genes in the 78 tissue samples in [14], calculated without correction for selection bias due to using the top 70 genes. The dashed line is the corresponding rate with correction for this bias.

to the full set of 5422 genes and selecting the top 70 genes on the training subset at each stage of cross-validation. Then the SVM with RFE is applied to this selected set of top 70 genes, which may have little in common with the original set of top 70 genes. The results for these two cross-validated error rates are listed in Table 1 and plotted in Figure 2. It can be seen from Table 1 and Figure 2 that the (estimated) selection bias in ignoring the fact that the SVM is being applied to the top $N = 70$ genes from a total of $N_T = 5422$ genes can be high as 14%, depending on the size of the final subset of genes.

We have also listed in Table 1 the external cross-validated error for the SVM with RFE, starting with the full set of $N_T = 5422$ genes. It can be seen that it is similar to that of the external cross-validated error rate of the rule starting with the top 70 genes. But to provide this latter estimate in practice one would need to have access to the full data set.

The bias arising from using just the top N genes of the 5422 genes in the study of van 't Veer et al. [14] will decrease in magnitude as N approaches the size $N_T = 5422$ of the full set. To investigate how large N must be before this bias is not of practical significance, we applied the same process to the top N genes for N taken (as a multiple of two) to be equal to 64, 128, 256, 512, 1024, 2048, and 4096. The biased (internal cross-validated) and unbiased (external cross-validated) error rates for each scenario are listed in Figure 3. The process is repeated with the nearest centroid (NC) classifier modified so that each gene is weighted by its sample-specific standard deviation rather than a class-specific standard deviation common for all genes. The results for this classifier are displayed in Figure 4.

From Figure 3 and Figure 4, we see that the difference between the error rates

Table 1. The number of Genes and Error Rates with and without Correction for Selection Bias.

Number of Genes	Error Rate for Top 70 Genes (without Correction for Selection Bias as Top 70)	Error Rate for Top 70 Genes (with Correction for Selection Bias as Top 70)	Error Rate for 5422 Genes (with Correction for Selection Bias)
1	0.31	0.42	0.44
2	0.32	0.36	0.42
4	0.27	0.45	0.35
8	0.27	0.33	0.31
16	0.23	0.38	0.33
32	0.19	0.38	0.33
64	0.21	0.33	0.37
70	0.21	0.32	–
128	–	–	0.44
256	–	–	0.45
512	–	–	0.44
1024	–	–	0.41
2048	–	–	0.44
4096	–	–	0.42
5422	–	–	0.45

starts to decrease as N increases; that is, the selection bias due to working with just the top N genes is shrinking. When the top $N = 4096$ genes are used, which include almost the entire data set, this bias is very small.

7 Discussion

In classifying a microarray gene-expression data with N genes, it is customary to reduce the number of genes N by some selection method and to base the final version of the discriminant rule (prediction rule) on a reduced set N_o, where N_o may be much smaller than N. In estimating the error rate of the rule based on the N_o selected genes, care must be taken that the selection bias

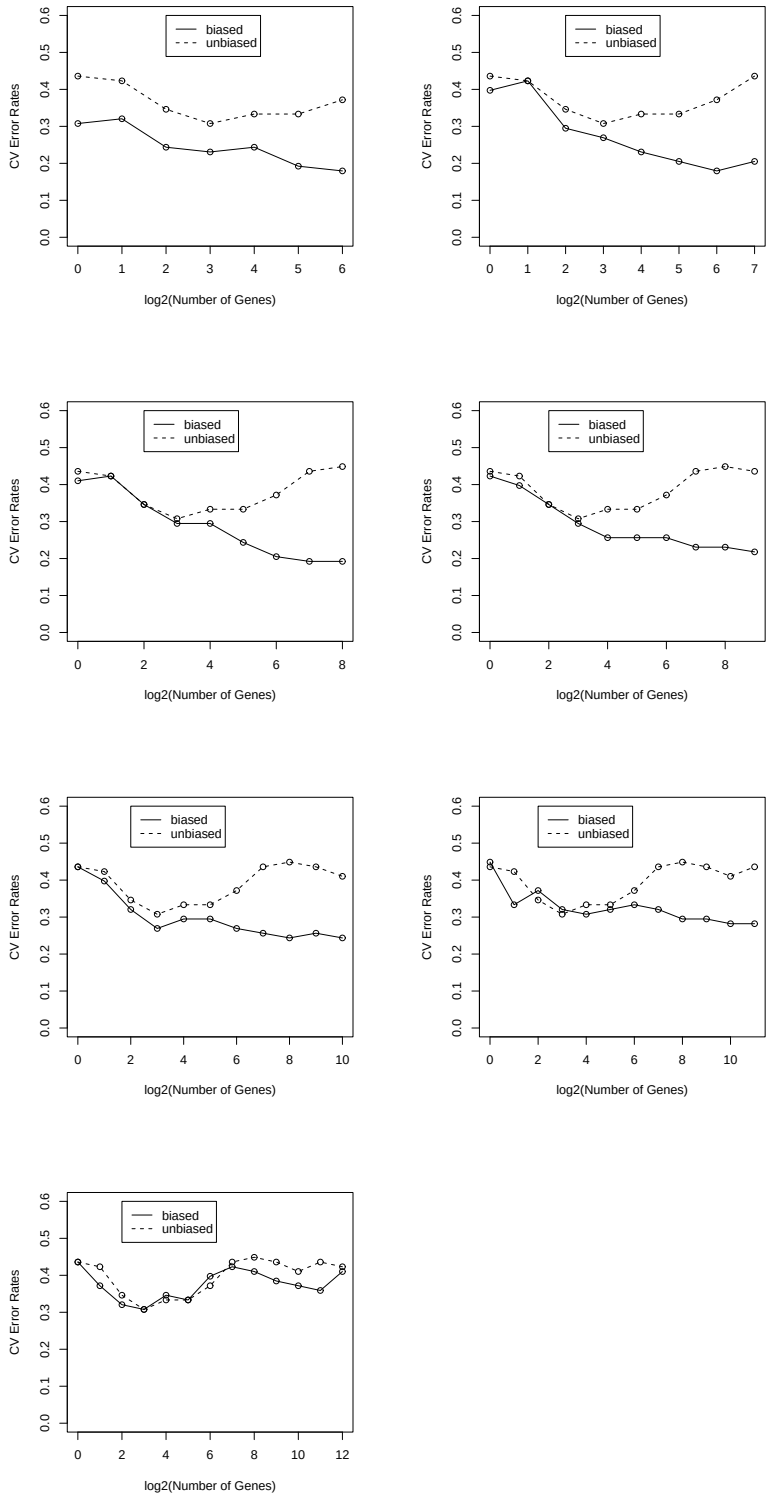

Fig. 3. The solid line is the ten-fold cross-validated error rate of the SVM with RFE applied to the top N genes ($N = 64, 128, 256, 512, 1024, 2048, 4096$) in the 78 tissue samples [14], calculated without correction for the selection bias due to using the top N genes. The dashed line is the corresponding rate with correction for this bias.

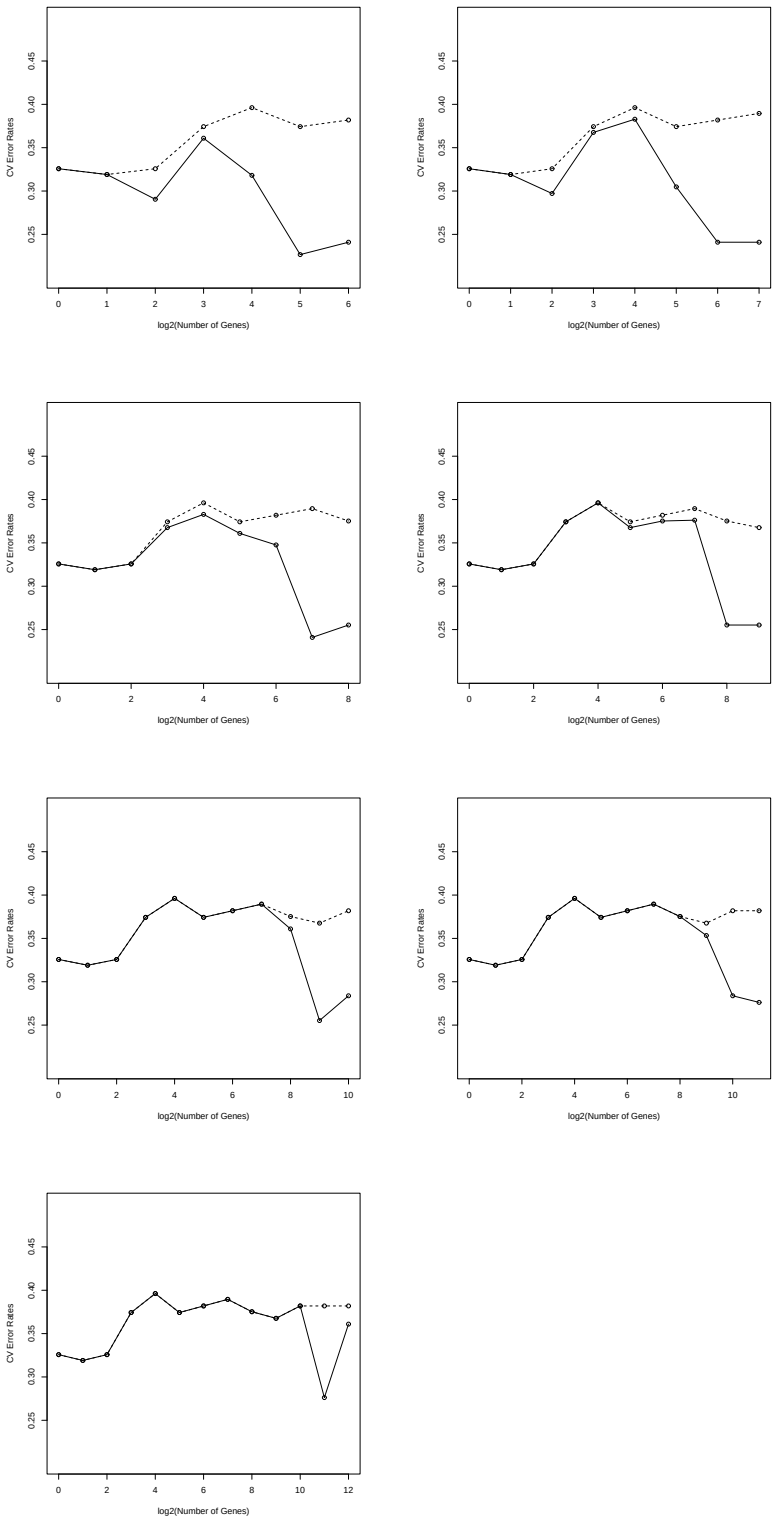

Fig. 4. The solid line is the ten-fold cross-validated error rate of the NC with RFE applied to the top N genes ($N = 64, 128, 256, 512, 1024, 2048, 4096$) in the 78 tissue samples [14], calculated without correction for the selection bias due to using the top N genes. The dashed line is the corresponding rate with correction for this bias.

has been corrected for as it can be quite appreciable, as illustrated in Figure 1.

In this paper, we caution that there is another source of selection bias that arises when the set of N genes on which gene selection has been performed is actually a subset of a much larger set of N_T genes. In practice, the N_T genes are usually reduced in size to N genes using some filtering process before a more sophisticated gene selection method is applied to the N genes. Now this will induce a bias as the N retained genes are not randomly chosen, but have been obtained by some filtering process designed in part to eliminate genes that appear not to be differentially expressed between the classes of tissue samples. Typically, N is still sufficiently large relative to the total number of genes N_T that the magnitude of this bias is not of practical importance. However, as demonstrated in an example involving a breast cancer data set, this bias is of concern if the set of N genes represents the top genes in some sense in the full set of N_T genes and N is relatively small. This situation can occur when an investigator having analysed a data set on a large number of genes, only makes available the expression levels on the tissue samples studied for what he/she has found to be the top N genes, say $N = 100$. This was almost the situation with the study of van de Vijver et al. [13]. Their study was on some 25000 genes on 295 breast cancer tumours where, in the reporting of their results, they have made available only the gene expression levels for the "top" 70 genes. These 70 genes were the top ranked genes according to the criterion (2) on the basis of some 78 tumour samples from the study of van 't Veer et al. [14] of which 61 are included in their larger data set of 295 tumours. Thus there will be bias in the estimate of a discriminant rule formed from the expression levels of these 70 genes over the 295 tumours, although it will not be as high as if the 70 genes had been ranked on the basis of all 295 tumours rather than a subset of 61 tumours.

The example we have given also serves to make the point that care must be exercised in comparing the error rates of two discriminant rules formed from the same tissue samples of different sets of genes. For example, one rule r_1 may be formed from a training set of $n = M$ tissue samples of $p = N$ genes, while another rule r_2 might be formed using a subset of these N genes, say, the top 100 genes. If a fair comparison is to be made between the error rates of these two rules, then the error rate of the second rule r_2 should not be estimated by just working with the top 100 genes during the cross-validation. Rather, one should start initially with the full set of N genes and select the top 100 genes on each stage of the training of r_2 in the cross-validation trials.

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences USA* **96** (1999) 6745–6750.

[2] Ambroise, C. and McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences USA* **99** (2002) 6562–6566.

[3] Efron, B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7** (1979) 1–26.

[4] Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM, 1982.

[5] Efron, B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78** (1983) 316–331.

[6] Efron, B. and Tibshirani, R. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* **92** (1997) 548–560.

[7] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **46** (2002) 389–422.

[8] Highleyman, W.H. The design and analysis of pattern recognition experiments. *Bell Systems Technical Journal* **41** (1962) 723–744.

[9] McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley, 1992

[10] McLachlan, G.J., Do, K.-A., and Ambroise, C. *Analyzing Microarray Gene-Expression Data.* Hoboken, New Jersey: Wiley, 2004

[11] Nguyen, D.V. and Rocke, D.M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, (2002) 39–50.

[12] Tibshirani, R.J., Hastie, T., Narasimhan, B., and Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences USA* **99** (2002) 6567–6572.

[13] van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., and Bernards, R. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347** (2002) 1999–2009.

[14] van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., 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. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** (2002) 530–536.