Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 374–386
journal of statistical planning and inference
www.elsevier.com/locate/jspi

On selection biases with prediction rules formed from gene expression data

J.X. Zhu[a, b], G.J. McLachlan[a, b, c, *], L. Ben-Tovim Jones[a, b, c], I.A. Wood[d]

[a] *Department of Mathematics, University of Queensland, St. Lucia, Brisbane 4072, Australia*
[b] *ARC Centre for Bioinformatics, University of Queensland, St. Lucia, Brisbane 4072, Australia*
[c] *Institute for Molecular Biosciences, University of Queensland, St Lucia, Brisbane 4072, Australia*
[d] *School of Mathematical Sciences, QUT-Gardens Point, Brisbane QLD 4001, Australia*

Available online 12 June 2007

Abstract

There has been ever increasing interest in the use of microarray experiments as a basis for the provision of prediction (discriminant) rules for improved diagnosis of cancer and other diseases. Typically, the microarray cancer studies provide only a limited number of tissue samples from the specified classes of tumours or patients, whereas each tissue sample may contain the expression levels of thousands of genes. Thus researchers are faced with the problem of forming a prediction rule on the basis of a small number of classified tissue samples, which are of very high dimension. Usually, some form of feature (gene) selection is adopted in the formation of the prediction rule. As the subset of genes used in the final form of the rule have not been randomly selected but rather chosen according to some criterion designed to reflect the predictive power of the rule, there will be a selection bias inherent in estimates of the error rates of the rules if care is not taken. We shall present various situations where selection bias arises in the formation of a prediction rule and where there is a consequent need for the correction of this bias. We describe the design of cross-validation schemes that are able to correct for the various selection biases.

Keywords: Gene expression data; Selection bias; Discriminant analysis; Support vector machine; Cross-validation

1. Introduction

Finding a diagnostic test for cancer has been an important goal of cancer diagnosis research for three decades (Ransohoff, 2005b). Recently, there has been increasing interest in changing the emphasis of tumour classification from morphologic to molecular. To this end, attention has focussed on the use of data from microarrays to find genes whose expression profiles are indicative of survival or the type of cancers. For example, accurately differentiating breast cancers with an inherently bad prognosis from those with a good prognosis could have a major effect on the provision of tailored and appropriate therapy. A clinically useful diagnostic tool would be based on some smaller set of genes (marker genes) whose expression profile is related to outcome, chosen from the much larger set of genes measured in the microarray experiment.

* Corresponding author. Department of Mathematics, University of Queensland, St. Lucia, Brisbane 4072, Australia. Tel.: +61 7 3365 2150; fax: +61 7 3365 1477.

E-mail addresses: j.zhu@imb.uq.edu.au (J.X. Zhu), gjm@maths.uq.edu.au (G.J. McLachlan), liatj@maths.uq.edu.au (L. Ben-Tovim Jones), i.wood@qut.edu.au (I.A. Wood).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.06.003

Microarrays allow the routine assessment of mRNA transcript levels on a genome-wide scale, and represent a major technological advance in molecular biology. However, the data-rich and highly dimensional nature of microarray data means that there are a number of issues to be addressed in using these data to form diagnostic tests. In particular, there is a need to be able to deal with the curse of dimensionality before microarrays can be used to provide reliable predictions in medical diagnosis and prognosis applications. Typically, there are only a limited number of tissue samples from the specified classes of tissues (tumours), whereas each tissue sample may contain the expression levels of thousands of genes. Thus researchers are faced with the problem of forming a prediction rule on the basis of a small number of classified tissue samples, which are of very high dimension. Usually, the final form of the prediction rule adopted will use only a small subset of the originally available genes. As the subset of genes used in the final form of the rule have not been randomly selected but rather chosen according to some criterion designed to reflect the predictive power of the rule, there will be a selection bias inherent in estimates of the error rates of the rules if care is not taken.

Indeed, the potential to overlook serious selection biases in the reporting of results on prediction rules formed from microarray data has led to claims that this may threaten the validity of cancer molecular–marker research (Ransohoff, 2005a). In addition, there have been questions raised in the literature over the validity of marker genes found in several recent microarray cancer studies. It has been suggested that often the list of genes identified as predictors of prognosis are highly unstable, and depend on the selection of patients in the training sets (Michiels et al., 2005).

We consider in depth the breast cancer study of van’t Veer et al. (2002), which has generated much discussion in the literature. We also consider a further study from this group of researchers, which includes a larger set of patients (van de Vijver et al., 2002). We shall describe various situations where selection bias arises in the formation of a prediction rule and where there is a consequent need for the correction of this bias.

In presenting the results for selection bias, we shall focus exclusively on the support vector machine (SVM) with linear kernel as the prediction rule, except in the last example where we also consider Fisher’s linear discriminant rule. The SVM (with linear kernel) is comparable if not better than any other rule in the case of training data consisting of a limited number of tissue samples containing the expression levels for thousands of genes (McLachlan et al., 2004, Chapter 6). In any event, the choice of prediction rule is not crucial in our results on selection bias, as similar biases will occur for any prediction rule formed on subsets of genes selected in a similar manner to that adopted in our applications of the SVM.

2. Notation

Before we proceed to discuss the selection bias problem in estimating the accuracy of a prediction rule, we need to introduce some notation. Also, we need to define formally the prediction rules to be considered and the methods of error-rate estimation to be adopted.

In the present context, the problem is to construct a prediction (discriminant) rule $r(\boldsymbol{y})$ that can accurately predict the class of origin of a tumour tissue with feature vector $\boldsymbol{y}$, which is unclassified with respect to a known number $g(\geqslant 2)$ of distinct tissue types, denoted here by $C_1, \ldots, C_g$. Here the feature vector $\boldsymbol{y}$ contains the expression levels of a very large number p of genes (features). If $r(\boldsymbol{y})=i$, then it implies that $\boldsymbol{y}$ should be assigned to the ith class C_i $(i=1,\ldots,g)$. In applications concerned with the diagnosis of cancer, one class C_1 may correspond to cancer and the other (C_2) to benign tumours. In applications concerned with patient survival following treatment for cancer, one class (C_1) may correspond to the good-prognosis class and the other C_2 to the poor-prognosis class. Also, there is interest in the identification of “marker” genes that characterize the different tissue classes. This is the feature selection problem. In the situation where the intention is limited to making an outright assignment to one of the possible classes, it is perhaps more appropriate to use the term prediction rather than discriminant to describe the rule. However, we shall use either nomenclature regardless of the underlying situation. In the pattern recognition jargon, such a rule is referred to as a classifier.

In order to train the prediction rule, there are available training data $\boldsymbol{t}$ consisting of n tissue samples of known classification. These data are obtained from n microarrays, where the jth microarray experiment gives the expression levels of the p genes in the jth tissue sample $\boldsymbol{y}_j$ of the training set. The vector

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

denotes the training data, where

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

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 $(i = 1, \ldots, g;\ j = 1, \ldots, n)$. We shall write the sample rule formed from the training data $\boldsymbol{t}$ as $r(\boldsymbol{y};\ \boldsymbol{t})$ to show its dependence on the training data $\boldsymbol{t}$.

2.1. Different types of error rates

For a given realization $\boldsymbol{t}$ of the training data $\boldsymbol{T}$, it is the conditional or actual allocation rates of a sample prediction rule $r(\boldsymbol{y};\ \boldsymbol{t})$ that are of central interest. They are given by

$$ec_{ij} = \mathrm{pr}\{r(\boldsymbol{Y}; \boldsymbol{t}) = j \mid \boldsymbol{Y} \in C_i, \boldsymbol{t}\} \quad (i, j = 1, \ldots, g). \tag{2}$$

That is, ec_{ij} is the probability, conditional on $\boldsymbol{t}$, that a randomly chosen observation from C_i is assigned to C_j by $r(\boldsymbol{y};\ \boldsymbol{t})$.

The unconditional or expected allocation rates of $r(\boldsymbol{y};\ \boldsymbol{t})$ are given by

$$\begin{aligned} eu_{ij} &= \mathrm{pr}\{r(\boldsymbol{Y};\ \boldsymbol{T}) = j \mid \boldsymbol{Y} \in C_i\} \\ &= E\{ec_{ij}\} \quad (i, j = 1, \ldots, g). \end{aligned}$$

The unconditional rates are useful in providing a guide to the performance of the rule before it is actually formed from the training data.

Concerning the error rates specific to a class, the conditional probability of misallocating a randomly chosen member from C_i is

$$ec_i = \sum_{j \neq i}^{g} ec_{ij} \quad (i = 1, \ldots, g).$$

The overall conditional error rate for an entity drawn randomly from a mixture G of $C_1, \ldots, C_g$ in proportions $\pi_1, \ldots, \pi_g$, respectively, is

$$ec = \sum_{i=1}^{g} \pi_i ec_i.$$

The individual class and overall unconditional error rates, eu_i and eu, are defined similarly.

If $r(\boldsymbol{y};\ \boldsymbol{t})$ is constructed from $\boldsymbol{t}$ in a consistent manner with respect to the Bayes rule $r_{\mathrm{o}}(\boldsymbol{y})$, then

$$\lim_{n \to \infty} eu = e_{\mathrm{o}},$$

where e_{o} denotes the optimal error rate. Interest in the optimal error rate in practice is limited to the extent that it represents the error of the best obtainable version of the sample-based rule $r(\boldsymbol{y};\ \boldsymbol{t})$. The Bayes or optimal rule $r_{\mathrm{o}}(\boldsymbol{y})$ is defined by

$$r_{\mathrm{o}}(\boldsymbol{y}) = \arg\max_{i} \tau_i(\boldsymbol{y}), \tag{3}$$

where

$$\begin{aligned} \tau_i(\boldsymbol{y}) &= \mathrm{pr}\{\boldsymbol{Y} \in C_i \mid \boldsymbol{y}\} \\ &= \pi_i f_i(\boldsymbol{y})/f(\boldsymbol{y}) \end{aligned} \tag{4}$$

is the posterior probability that $\boldsymbol{y}$ belongs to the ith class $C_i\,(i = 1, \ldots, g)$. Here $f_i(\boldsymbol{y})$ denotes the density of $\boldsymbol{y}$ in class C_i and

$$f(\boldsymbol{y}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}) \tag{5}$$

is the unconditional density of $\boldsymbol{y}$.

3. Prediction rule

3.1. Fisher's linear rule

In presenting the results on selection bias in this paper in the case of $g = 2$ classes, we shall focus on the use of the SVM, but we also use Fisher's linear rule in the last example.

For Fisher's linear rule, $r(\boldsymbol{y};\ \boldsymbol{t}) = 1$ or 2 according as

$$\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{y} \tag{6}$$

is greater or less than zero, where

$$\beta_0 = -\tfrac{1}{2}(\bar{\boldsymbol{y}}_1 + \bar{\boldsymbol{y}}_2)^{\mathrm{T}}\boldsymbol{S}^{-1}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2) \tag{7}$$

and

$$\boldsymbol{\beta} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2). \tag{8}$$

In (7) and (8), $\bar{\boldsymbol{y}}_i$ denotes the sample mean of the n_i observations from class $C_i\,(i = 1, 2)$, and

$$\boldsymbol{S} = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}(\boldsymbol{y}_j - \bar{\boldsymbol{y}}_i)(\boldsymbol{y}_j - \bar{\boldsymbol{y}}_i)^{\mathrm{T}}/(n-g) \tag{9}$$

is the pooled within-class sample covariance matrix.

3.2. Support vector machine (SVM)

For an SVM with linear kernel, the rule $r(\boldsymbol{y};\ \boldsymbol{t})$ is 1 or 2 according as the sign of (6) is positive or negative, where β_0 and the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ are chosen to maximize the distance between the separating hyperplane and the nearest point in any of the classes. When the data are not separable, there is no separating hyperplane; in this case, the aim is still to try to maximize the margin but allow some classification errors subject to the constraint that the total error (distance from the hyperplane on the wrong side) is less than a constant (Vapnik, 1998).

4. Feature reduction

As the number of genes is much greater than the number of tissues, consideration is usually given initially to reducing the number of genes. Frequently, this is undertaken in some ad hoc manner before a more formal method of feature selection is adopted in conjunction with the choice of prediction rule. Two formal methods of feature selection are considered in this paper in presenting the results on selection bias. One is based on the ratio of the between-class sum of squares to the within-class sum of squares, while the other is based on the size of the weights of the SVM.

For a given feature variable (gene), the ratio of the between-class sum of squares to the within-class sum of squares on their degrees of freedom can be expressed as the following:

$$F_v = (\boldsymbol{B})_{vv}/(\boldsymbol{S})_{vv}, \tag{10}$$

where

$$\boldsymbol{B} = \frac{1}{g-1}\sum_{i=1}^{g} n_i(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})^{\mathrm{T}} \tag{11}$$

and $\boldsymbol{S}$ is defined by (9). When $g = 2$, F_v is equal to the square of the pooled two-sample Studentized t-statistic.

The other method used here to rank the genes with applications of the SVM is based on the size of the magnitude of the coefficient β_v of the expression level of the vth gene in forming the linear combination (6). This is called a wrapping method, as the gene reduction method is embedded in the prediction rule.

Guyon et al. (2002) have proposed that the feature variables be selected according to a backward elimination process, which they call recursive feature elimination (RFE). It starts by considering all available features. Each step consists in ranking the features according to the order of magnitude of their associated coefficients (weights) β_i, and then discarding the bottom-ranked variables. 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 examples to be presented here, we follow the practice adopted by Guyon et al. (2002) and first discard enough bottom-ranked genes so that the number retained is the greatest power of 2 (less than the original number of genes). We then sequentially proceeded to discard half the current number of genes on each subsequent step.

As noted by Guyon et al. (2002), the process can be refined at the expense of greater computational time by deleting fewer variables on each step. One suggested procedure is to apply RFE by removing chunks of features in the first few iterations and then removing one feature at a time once the feature set size reaches a few hundred.

5. Error-rate estimation

We consider now the estimation of the error rates associated with a discriminant rule $r(\boldsymbol{y};\, \boldsymbol{t})$ formed from some realized training data $\boldsymbol{t}$. 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 $r(\boldsymbol{y};\, \boldsymbol{t})$. 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.

5.1. Cross-validation

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 of $r(\boldsymbol{y};\, \boldsymbol{t})$ in its application to the observations in $\boldsymbol{t}$. That is, A is the proportion of the observations in $\boldsymbol{t}$ misallocated by $r(\boldsymbol{y};\, \boldsymbol{t})$. Thus we can write

$$A = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\, Q[i, r(\boldsymbol{y}_j;\, \boldsymbol{t})], \tag{12}$$

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

One way of almost eliminating the bias in the apparent error rate is through the leave-one-out (LOO) technique as described by Lachenbruch and Mickey (1968) or cross-validation (CV) as discussed in a wider context by Stone (1974) and Geisser (1975). For the estimation of $ec(\boldsymbol{t})$ by the apparent error rate A, the LOO cross-validated estimate is given by

$$A^{(\mathrm{CV})} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\, Q[i, r(\boldsymbol{y}_j;\, \boldsymbol{t}_{(j)})], \tag{13}$$

where $\boldsymbol{t}_{(j)}$ denotes $\boldsymbol{t}$ with the point $\boldsymbol{y}_j$ 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.

5.2. q-Fold CV

CV 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, $n = qm (m \geqslant 1)$. Let now

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

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

$$A^{(qCV)} = \sum_{i=1}^{g}\sum_{j=1}^{m}\sum_{k=1}^{q} z_{ij}\, Q[i,\ r(\boldsymbol{y}_{(k-1)m+j}; \boldsymbol{t}_{(k)})]/n, \tag{14}$$

which requires only q recomputations of the rule.

The choice of $q = n$ (LOO) does not perturb the data enough and results in higher variance. With $q = 2$, the training sets are too small relative to the full training sets. The values $q = 5$ or 10 are a good compromise.

For q-fold CV, there is an easily computed estimate of the standard errors of the estimated error rates $A^{(q\text{CV})}$. It is given by the standard error of the q error rates that are obtained when the discriminant rule is applied to the q validation subsamples during the CV. This standard error of $A^{(q\text{CV})}$ is

$$\text{SE}(A^{(q\text{CV})}) = \left[\frac{1}{q(q-1)}\sum_{h=1}^{q}(A_h^{(q\text{CV})} - A^{(q\text{CV})})^2\right]^{\frac{1}{2}}, \tag{15}$$

where $A_h^{(q\text{CV})}$ is the error rate when the rule formed from the hth training subsample is applied to the hth validation subsample. Similarly, the standard errors of the individual q-fold cross-validated error rates can be formed.

6. Error-rate estimation with selection bias

Caution has to be exercised in selecting a small number of variables from a large set, as there will be a selection bias associated with choosing the optimal of a large number of possible subsets, regardless of the criterion. In such situations, it is important to distinguish between what we shall call an ordinary or internal CV of the prediction rule and a so-called external CV. We shall see that it is by an appropriate external CV that we are able to correct for the various types of selection bias in forming rules from a very large number of feature variables on the basis of limited training data points of known classification.

6.1. External CV

We now describe external CV for a rule formed from a subset of the available feature variables. 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 prediction 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, $A^{(\text{CV})}(\boldsymbol{t}^{(s)})$, over all possible $\binom{p}{p_{s_o}}$ distinct subsets s of size p_{s_o}. Although $A^{(\text{CV})}(\boldsymbol{t}^{(s)})$ may be an (almost) unbiased estimator of the overall conditional error rate of the rule $r(\boldsymbol{y}^{(s)};\ \boldsymbol{t}^{(s)})$, $A^{(\text{CV})}(\boldsymbol{t}^{(s_o)})$ 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 (LOO) cross-validated estimate is given by

$$A^{(\text{CV})}(\boldsymbol{t}^{(s_o)}) = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\, Q[i, r(\boldsymbol{y}_j^{(s_o)}; \boldsymbol{t}_{(j)}^{(s_o)})], \tag{16}$$

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

In order to reduce the selection bias which is still present in the estimate (16), an external CV should be performed whereby the selection process is undertaken for each deletion of a feature vector from the training set. This externally cross-validated estimate of the overall error rate of $r(\boldsymbol{y}^{(s_o)}; \boldsymbol{t}^{(s_o)})$ is given by

$$A^{(\text{CVE})}(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\, Q[i, r(\boldsymbol{y}_j^{(s_{oj})}; \boldsymbol{t}_{(j)}^{(s_{oj})})], \tag{17}$$

where s_{oj} denotes the optimal subset, according to the adopted selection criterion applied to the training data $\boldsymbol{t}_{(j)}$ without $(\boldsymbol{y}_j^{\mathrm{T}}, \boldsymbol{z}_j^{\mathrm{T}})^{\mathrm{T}}$. As the notation implies, the selected subset s_{oj} for the allocation of the jth observation may be different for each j $(j = 1, \ldots, n)$.

It is this reselection of a new subset s_{oj} from the training subsample on the validation trial for the jth tissue that is crucial in eliminating the selection bias, which would be present in the estimate if the subset s_o from the full training set were used on each trial.

7. Selection bias: optimal subset of fixed size

We firstly demonstrate the selection bias that can occur when we estimate the error rate of a prediction rule based on a subset of the available genes via CV without taking into account that the subset has been selected in some optimal way. We consider a subset of the data set in van't Veer et al. (2002). This data set contains the expression levels of 24,881 genes in 98 primary breast cancers acquired from three classes of patients: 44 representing a good-prognosis class (that is, those who remained metastasis free after a period of more than five years), 34 from a poor-prognosis class (those who developed distant metastases within five years), and 20 representing a hereditary form of cancer, due to a BRCA1 (18 tumours) or BRCA2 (two tumours) germline mutation. The 78 sporadic (non-BRCA) breast cancer patients were chosen specifically on the basis of their clinical outcome and they were used to form the two classes (good- and poor-prognosis classes) in this example. van't Veer et al. (2002) identified a set of 70 genes with expression profiles associated with the risk of early metastasis (the poor-prognosis class) for the 78 patients with sporadic breast cancer.

After adopting the same filtering procedure used in van't Veer et al. (2002) for these 78 sporadic breast cancer patients, the data set to be considered here consisted of the expression levels of some 5422 genes. We applied a SVM with RFE to these data, and the (10-fold) cross-validated error rates at each stage of the selection procedure are displayed in Table 1. The first column gives the error rate $A^{(10\mathrm{CV})}$ without an external CV being implemented, while the second column gives the increased estimate $A^{(10\mathrm{CVE})}$ using an external validation. More specifically, consider the entries of 0.15 and 0.29 for $A^{(10\mathrm{CV})}$ and $A^{(10\mathrm{CVE})}$, respectively, for the SVM formed from the remaining eight genes during the RFE process. With the former, the subset of eight genes as obtained by RFE applied to the full data set is used on each of the 10 validation trials. But with the latter, the selection procedure RFE is run on each of the ten validation trials, starting with all the genes, to obtain a possibly new reduced subset of eight genes. That is, it uses the 10-fold version of (17). As noted by Ambroise and McLachlan (2002), this selection bias is often overlooked in the bioinformatics literature.

The fact that we do not always obtain the same eight genes on each on the 10 validation trials can be used to identify potential marker genes. To look at this for the current example, we have listed in Table 2 the genes obtained on each of

Table 1
Cross-validated error rates of SVM with RFE applied to 5422 genes on $n_1 = 44$ good-prognosis patients (C_1) and $n_2 = 34$ poor-prognosis patients (C_2)

Number of genes	Internal CV error rate	External CV error rate
1	0.28	0.40
2	0.19	0.40
4	0.21	0.42
8	0.15	0.29
16	0.18	0.38
32	0.12	0.38
64	0.10	0.33
128	0.12	0.32
256	0.17	0.29
512	0.15	0.31
1024	0.19	0.32
2048	0.22	0.35
4096	0.31	0.37
5422	0.37	0.37

Table 2
Selection frequency of genes in external cross-validation of error rate for SVM based on eight genes retained using RFE

Frequency of retention in eight-gene subset	Gene labels
8	**2510**
7	**1531**
6	**2316**
5	902 **4750**
4	1747 2226
3	1416 3386
2	4401 4953
1	355 416 1055 **1269** 1360 1609 1680 1732 2191 2207 2287 2469 2719 2876 3027 3112 3362 3389 3518 3925 4156 4474 4624 4660 4861 4972 **5001** 5060 5115 5134 5334

Table 3
Cross-validated error rates of SVM with RFE applied to the top 70 genes on $n_1 = 44$ good-prognosis patients (C_1) and $n_2 = 34$ poor-prognosis patients (C_2) without bias correction for starting with a top subset of the genes

Number of genes	Internal CV error rate	External CV error rate
1	0.28	0.40
2	0.19	0.31
4	0.17	0.24
8	0.17	0.23
16	0.18	0.22
32	0.19	0.19
64	0.23	0.23
70	0.23	0.23

the 10 validation trials during CV. The gene labels simply refer to the position of that gene in the set of 5422 genes and, for the genes discussed below, the gene identifier or gene name where known is given in parentheses. The eight-gene set obtained for SVM with RFE on the full data set consists of the genes labelled 3979 (Contig50129_RC), 902 (U72507), 1747 (WISP1), 1269 (LOC57110), 1531 (ALDH4), 4750 (OXCT), 2226 (Contig48328_RC), and 2510 (AL080059). Except for the first gene, all of these are included in the van't Veer subset of 70 genes.

It can be seen from Table 2 that there is one gene (2510, AL080059) that is selected on eight of 10 validation trials. Gene 1531 (ALDH4) is selected on seven of the 10 validation trials, while gene 2316 (AA555029_RC) is selected on six of the 10 trials. The genes 2510, 1531 and 2316 are ranked 1, 6 and 5, respectively, in the 70 genes of van't Veer et al. (2002). In Table 2, we have listed in boldface those genes (six in number) that are among this top 70 subset.

8. Additional selection bias from preliminary screening

The selection of the 70 genes by van't Veer et al. (2002) 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 (10) in the case of $g = 2$. They called these 70 genes the prognostic marker genes. In Table 3, we list the ordinary 10-fold cross-validated error rate $A^{(10\mathrm{CV})}$ and the 10-fold externally cross-validated rate $A^{(10\mathrm{CVE})}$ for the SVM with RFE applied to these 70 "top" genes. These error rates are plotted in Fig. 1.

On comparing the values for the externally cross-validated error rate $A^{(10\mathrm{CVE})}$ for the SVM based on less than 70 genes with those obtained from starting with the full set of 5422 genes as in Table 3, it would appear that it is beneficial for predictive purposes to first select a top set of genes according to the t-statistic and then run SVM with RFE on this top subset rather than starting with all the genes. This is contrary, however, to the widely held view that the performance of the SVM is not unduly affected by forming it on the basis of the set of all available genes, even though the latter

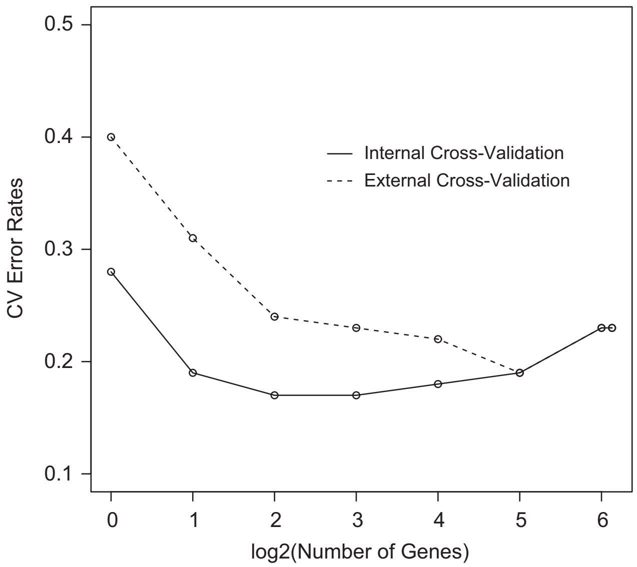

Fig. 1. Plot of internally and externally cross-validated error rates for SVM applied with RFE to the top 70 genes.

Table 4
Cross-validated error rates of SVM with RFE applied to the top 70 genes on $n_1 = 44$ good-prognosis patients (C_1) and $n_2 = 34$ poor-prognosis patients (C_2) with bias correction for starting with a top subset of the genes

Number of genes	External CV error rate using top 70 genes with bias correction for using just this subset	External CV error rate using 5422 genes
1	0.40	0.40
2	0.37	0.40
4	0.40	0.42
8	0.32	0.29
16	0.33	0.38
32	0.35	0.38
64	0.32	0.33
70	0.32	
128		0.32
256		0.29
512		0.31
1024		0.32
2048		0.35
4096		0.37
5422		0.37

may be very large in magnitude. This apparent conflict is resolved on realizing as in Zhu et al. (2005) that there is a selection bias due to working with a relatively small subset of the available genes selected in some “optimal” way.

To demonstrate this selection bias for this data set, we list in Table 4 the values of $A^{(10\mathrm{CVE})}$ for the SVM with RFE applied to the 70 top genes, but where now the external CV is implemented to remove the bias due to starting with an optimally selected subset. That is, on each of the 10 validation trials, the t-statistics for the 5422 genes were calculated for the training subsample and the top 70 genes selected for the subsequent application of RFE to them.

It could be argued that there is still a residual selection bias in the externally cross-validated estimates in Table 4, since the 5422 genes are a subset of the full set of 24,481 genes, which have not been randomly selected but have been chosen to be useful in some sense. This has been investigated in Zhu et al. (2005), who concluded that the initial set of genes has to be relatively small relative to the total number of genes available for this bias to be of practical significance.

Table 5
Cross-validated error rates starting with 70 genes versus all available genes with missing values in the latter handled by either substituting zero or k-NN estimate

Number of genes	External CV error rate using 70 genes (without correction for bias from using just these 70 genes)	External CV error rate using original 24,481 genes (missing values set to 0)	External CV error rate using original 24,481 genes (using k-NN for missing values, $k = 10$)
1	0.29	0.40	0.42
2	0.17	0.39	0.38
4	0.20	0.37	0.38
8	0.13	0.30	0.31
16	0.11	0.23	0.23
32	0.09	0.20	0.22
64	0.10	0.20	0.19
70	0.10		
128		0.16	0.16
256		0.15	0.15
512		0.14	0.14
1024		0.13	0.15
2048		0.15	0.14
4096		0.15	0.14
8192		0.16	0.16
16,384		0.17	0.17
24,481		0.17	0.18

But this example does serve to make the point that in studies concerning classification results for microarray gene expression data it is important that the expression levels of a sufficient number of the full set of genes be made available. Otherwise, it would not be possible to confirm that the selection bias due to working with a top subset of genes is not of practical significance.

It can be seen from Table 1 that the data in van't Veer et al. (2002) consisting of the sporadic 78 breast cancers are of limited discriminatory value. Subsequently, van't Veer and colleagues (van de Vijver et al., 2002) considered a larger set of 295 breast cancer tissue samples, which consisted of 61 of the 78 patients in van't Veer et al. (2002) who were lymph-node negative. The good- and poor-prognosis training samples were not based on outcomes but were actually obtained by assigning them to the two classes on the basis of the gene-expression signatures for these 61 patients. More precisely, each of the 234 new tissues was assigned to class C_1 or C_2 according to whether the correlation between its gene-signature vector and the mean of the good-prognosis class C_1 was greater or less than 0.4. Each of the 61 original tissues was reassigned to C_1 to C_2 according to whether the (cross-validated) correlation between its gene-signature vector and the mean of C_1 was greater or less than 0.55 (a threshold that resulted in a 10% rate of false negatives in the study of van't Veer et al. (2002)).

For their expanded data set of 295 breast cancers, van de Vijver et al. (2002) listed only the expression levels for the top 70 genes as found in the previous study of van't Veer et al. (2002). In Table 5, we list in the first column the externally cross-validated error rates at each stage of the RFE procedure applied to the 70 genes. However, as the process is being started from a small subset of a much larger number (24,481) of available genes, there will be a selection bias unaccounted for. But this bias would not be as large if the 70 genes being used were the top 70 for the t-statistic based on all the 295 tissue samples rather than on 61 of them. The expression levels of all 24,481 genes were later made available at http://microarray-pubs.stanford.edu/wound_NKI/ in the study by van't Veer and van de Vijver (Chang et al., 2005). Thus it is now possible to correct for this bias. The estimated error rates corrected for this bias are listed in the last two columns of Table 5 for two methods of allowing for many missing values in the full data set. It can be seen that this bias due to starting with the 70 genes is around 0.05 since the smallest error using them is 0.084 in contrast to 0.136 by starting with all the genes (and using 10-NN neighbour estimates for missing values). Given that the majority of the training samples in this data set of van de Vijver et al. (2002) are not based on outcomes and are not random samples from the two classes, there will be a bias due to this, but we are unable to correct for it.

Table 6
Ten-fold cross-validated error rates for SVM based on subset with minimum error rate in RFE process with and without bias correction for optimization over subsets of varying sizes

Data set	Error rate without bias correction for varying size of optimal subset	Error rate with bias correction for varying size of optimal subset
van’t Veer (5422 genes)	0.29	0.37
van’t Veer (70 genes)	0.093	0.11
Sharma	0.156	0.18
Alon	0.108	0.12

9. Selection bias: optimal subset of unrestricted size

In the previous examples, we have focussed on the need to correct for selection bias in estimating via CV the error rate of a prediction rule based on subsets of a specified size (that is, specified *a priori*). We now consider the case where the size of the selected subset is not specified *a priori*, but where the final form of the rule is based on the subset of genes yielding the lowest estimated error rate over the subsets of different sizes considered. Given this minimization over subsets of varying sizes, there is a further selection bias that needs to be corrected for. For example, in Table 1, the externally cross-validated error rate has its lowest value of 0.29 for subsets with 8 and 256 genes. In this situation in practice, the final form of the SVM would be based on the eight-gene subset. Hence if we are to use the SVM based on the subset of genes that minimizes the externally cross-validated error rate over the retained subsets of genes of powers of 2, we need to correct also for this bias. Thus on each of the 10 validation trials, we have to determine the subset of genes that has smallest externally cross-validated error rate. That is, we have to carry out an external CV of the error rate of the SVM for each training subsample in order to determine the subset of genes on which to base the SVM for its application to the test subsample. There are thus two layers of CV being performed. The optimal subset of retained genes may be of a different size on each validation trial.

On undertaking these two layers of CV, we find that the cross-validated error rate of the SVM based on the subset with smallest error rate is 0.37, which is larger than the uncorrected value of 0.29. This result is given in Table 6 where we also have listed the corresponding result for the van’t Veer et al. (2002) data set using only 70 genes and for two other data sets from Alon et al. (1999) and Sharma et al. (2005). The former consists of the expression levels of 2000 genes in 62 tissue samples for 40 colon cancer and 22 normal patients. The latter consists of the expression levels of 1368 genes in 102 tissue samples for 40 breast cancer and 62 normal patients. It can be seen that this bias is of some practical significance.

10. Selection bias: use of test set in selection

One way of avoiding the bias in the 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. The available data are split into disjoint training and test subsamples. The prediction rule is formed from the training subsample and then assessed on the test subsample. Clearly, this method is inefficient in its use of the data, particularly in the context of microarray data, where the number of tissue samples is usually limited.

The holdout method has been used in feature selection. The prediction rule is formed for various subsets from the training subset and applied to the test subsample to provide estimates of the error rate for each subset of variables (genes) under consideration. The subset having the lowest error rate is then chosen as the optimal subset and the final version of the prediction rule is based on it. However, the error rate for this rule is not estimated unbiasedly by its error rate on the holdout subsample. In order for the holdout method to provide an unbiased estimate, the test subsample should play no role in the formation of the prediction rule. In the aforementioned selection method, the test subsample plays a role in leading to the choice of the final form of the prediction rule. However, this is frequently overlooked in the bioinformatics literature (Somorjai et al., 2003). It often leads to claims that a prediction rule can be formed from only a few genes that has almost zero error rate.

To demonstrate this selection bias in using the test subsample to play a role in the choice of the genes, we first perform a simple simulation experiment. The training subsample was taken to consist of $n_1 = n_2 = 10$ observations from each

Table 7
Comparison of error rates with and without bias correction for selection bias for univariate version of Fisher's linear rule

p	Error rate (with bias correction)	Error rate (without bias correction)
50	0.32	0.10
100	0.32	0.08
250	0.29	0.06
500	0.30	0.04

Table 8
Comparison of error rates with and without bias correction for selection bias for univariate version of SVM

p	Error rate (with bias correction)	Error rate (without bias correction)
50	0.33	0.10
100	0.33	0.09
250	0.31	0.06
500	0.29	0.04

of two classes in which the feature vector was taken to be multivariate normal with common covariance matrix equal to the identity matrix and with mean zero in class C_1 and mean $(1, 1, \ldots, 1)^{\mathrm{T}}$ in class C_2. The dimension p of the feature vector was taken to be equal to 50, 100, 250, and 500, and 200 simulations were performed for each value of p. On each simulation, a univariate prediction rule was formed for each feature variable and its error rate estimated by its application to the test subsample. The final form of the prediction rule was taken to be the version for that variable leading to the lowest error rate on the test subsample. The error rate so obtained over the 200 simulations is given for Fisher's linear rule in Table 7 and for the SVM in Table 8. We also list in these tables an unbiased estimate of the error rate obtained by an external LOO CV. With the latter, on each validation trial, the prediction rule is formed for the gene that leads to the lowest error rate on the test subsample after deletion of an observation from the training subsample.

It can be seen from Tables 7 and 8 that, as one would anticipate, the selection bias becomes more marked as the number of available genes for selection of the singleton subset grows in size. Even for a set of 100 genes, the selection bias is large, as an external CV shows that the error rate for Fisher's rule or the SVM rule formed from the optimal selection of a single gene is approximately 33%, whereas the error rate without bias correction is 8%.

In this situation the selection bias is a serious issue since the separation between the classes is poor. The bias will be of less importance if the classes are widely separated. To demonstrate this, we attempted to find the gene that minimizes the error rate of the SVM applied to the test data of the leukaemia data set of Golub et al. (1999). They considered the use of microarray data for differentiating between two different human acute leukaemias, namely acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). These data consisted of the expression levels of some 7129 genes for 38 training tissues (27 from AML and 11 from AML) and 34 test tissues (20 from AML and 14 from ALL). Using the SVM with linear kernel, it was found that three genes (1685, 2288, and 6855) gave the lowest error rate of 0.029 among all singleton gene subsets in their application to the test data set. On correction for bias, this error rate estimate was increased to 0.059.

This leukaemia data set of Golub et al. (1999) has been widely studied, and we know that AML and ALL classes are easily distinguished using expression levels, due to the nature of these cancers. Thus the selection bias is less of an issue here. However, in reality, the problem of using prediction rules to separate known classes of cancer is rarely of interest; rather the goal is to classify patients according to outcome. In addition, in the case of solid tumours, such as breast cancers, the tissues are of a heterogeneous nature, making the gene expression profiles more difficult to interpret. In these cases the subclasses are not so clear cut, and so the selection bias is an important issue in reporting the results of prediction rules.

11. Discussion

We have illustrated here the selection biases that can arise in practice when a prediction rule is formed from a subset of a very large number of genes. Attention has concentrated on the use of the SVM as the prediction rule, but the choice of this rule is incidental to the existence of the selection bias. The focus has been on the design of CV schemes to correct for these selection biases. We have demonstrated how it is possible to correct for selection bias by using an appropriate form of CV (an external CV) where the choice of subset is reconsidered on each validation trial of the CV process. A situation where it is not possible to completely correct for the selection bias occurs when the user-supplied data set gives only the expression levels of a subset of the leading genes according to some criterion and not the full set of genes. It can been seen from the size of the biases exhibited in the examples that selection bias is a serious issue. If it is overlooked with applications of prediction rules for diagnosis and prognosis, it can lead to a grossly overoptimistic assessment of their accuracy and hence of their usefulness in the detection and management of diseases.

Acknowledgements

The authors would like to thank Professor Peter Visscher (Queensland Institute of Medical Research) for helpful communications on this topic. Dr. Ian Wood wishes to acknowledge support from the ARC Center for Complex Dynamic Systems and Control CEO10348165 and NHMRC Grant 389892. The authors also would like to thank Dr. Richard Bean (Institute for Molecular Bioscience, University of Queensland) in the writing of this paper.

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In: Proceeding of the National Academy of Sciences USA, vol 96, pp. 6745–6750.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. In: Proceedings of the National Academy of Sciences USA, vol. 99, pp. 6562–6566.

Chang, H.Y., Nuyten, D.S.A., Sneddon, J.B., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H., He, Y.D., van't Veer, L.J., Bartelink, H., van de Rijn, M., Brown, P.O., van de Vijver, M.J., 2005. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. In: Proceedings of the National Academy of Sciences USA, vol. 102, pp. 3738–3743.

Geisser, S., 1975. The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70, 320–328.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeck, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learning 46, 389–422.

Lachenbruch, P.A., Mickey, M.R., 1968. Estimation of error rates in discriminant analysis. Technometrics 10, 1–11.

McLachlan, G.J., Do, K.-A., Ambroise, C., 2004. Analyzing Microarray Gene-Expression Data. Wiley, Hoboken, NJ.

Michiels, S., Koscielny, S., Hill, C., 2005. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365, 488–492.

Ransohoff, D.F., 2005a. Bias as a threat to the validity of cancer molecular–marker research. Nature 5, 142–149.

Ransohoff, D.F., 2005b. Lessons from controversy: ovarian cancer screening and serum proteomics. J. Nat. Cancer Inst. 97, 315–319.

Sharma, P., Sahni, N.S., Tibshirani, R., Skaane, P., Urdal, P., Berhagen, H., Jensen, M., Kristiansen, L., Moen, C., Sharma, P., Zaka, A., Arnes, J., Sauer, T., Akslen, L.A., Schlichting, E., Børresen-Dale, A.-L., Lönneborg, A., 2005. Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. Breast Cancer Res. 7, R634–R644.

Somorjai, R.L., Dolenko, B., Baumgartner, R., 2003. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 19, 1484–1491

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions (with discussion). J. Roy. Statist. Soc. B 36, 111–147.

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., Bernards, R., 2002. A gene-expression signature as a predictor of survival in breast cancer. New England J. Medicine 347, 1999–2009.

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., Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

Zhu, J.X., Ambroise, C., McLachlan, G.J., 2005. Selection bias in working with the top genes in supervised classification of tissue samples. Statist. Methodology 3, 29–41.