BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 11 2007, pages 1363–1370
doi:10.1093/bioinformatics/btm117

Gene expression

Classification based upon gene expression data: bias and precision of error rates

Ian A. Wood[1,*], Peter M. Visscher[2] and Kerrie L. Mengersen[1]

[1]School of Mathematical Sciences, Queensland University of Technology, Gardens Point, GPO Box 2434, Brisbane, QLD 4001, Australia and [2]Queensland Institute of Medical Research, Post Office, Royal Brisbane Hospital, 300 Herston Rd., Herston, QLD 4029, Australia

Received on November 27, 2006; revised on March 12, 2007; accepted on March 15, 2007

Advance Access publication March 28, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Gene expression data offer a large number of potentially useful predictors for the classification of tissue samples into classes, such as diseased and non-diseased. The predictive error rate of classifiers can be estimated using methods such as cross-validation. We have investigated issues of interpretation and potential bias in the reporting of error rate estimates. The issues considered here are optimization and selection biases, sampling effects, measures of misclassification rate, baseline error rates, two-level external cross-validation and a novel proposal for detection of bias using the permutation mean.

Results: Reporting an optimal estimated error rate incurs an optimization bias. Downward bias of 3–5% was found in an existing study of classification based on gene expression data and may be endemic in similar studies. Using a simulated non-informative dataset and two example datasets from existing studies, we show how bias can be detected through the use of label permutations and avoided using two-level external cross-validation. Some studies avoid optimization bias by using single-level cross-validation and a test set, but error rates can be more accurately estimated via two-level cross-validation. In addition to estimating the simple overall error rate, we recommend reporting class error rates plus where possible the conditional risk incorporating prior class probabilities and a misclassification cost matrix. We also describe baseline error rates derived from three trivial classifiers which ignore the predictors.

Availability: R code which implements two-level external cross-validation with the PAMR package, experiment code, dataset details and additional figures are freely available for non-commercial use from http://www.maths.qut.edu.au/profiles/wood/permr.jsp

Contact: i.wood@qut.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Recent studies suggest that a number of complex diseases can be accurately diagnosed on the basis of measurements of gene expression levels from microarrays and similar technology. Furthermore, they often suggest lists of the genes likely to be involved in the disease. Methods used include support vector machines (SVMs) (Guyon *et al.*, 2002; McLachlan *et al.*, 2004) nearest shrunken centroids (NSC) (Sharma *et al.*, 2005; Tibshirani *et al.*, 2002) neural networks (Khan *et al.*, 2001) classification trees and mixture models (McLachlan *et al.*, 2004).

These studies typically report estimates of classifier accuracy. However, it is not always clear how these results should be interpreted. There are a number of possible sources of bias in such estimates. In this study, we examine some of these, focusing particularly on optimization bias and sampling effects. We review existing methods for estimating and reporting prediction accuracy. We then give suggestions for improvements and examples of their effectiveness in the large p (number of features or predictors), small n (number of labelled samples) case common in gene expression analyses.

2 THEORY

2.1 Classification and error rates

Given a dataset of n observations, each comprising the measurement of p predictors and an expert-based classification of each point into one of G classes, we can fit a model and use this to classify these observations and future data of the same type. Methods of this type are known as classifiers or discriminant rules. Following Efron (1983), we let the dataset be $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, drawn from a distribution F, with each $\mathbf{x}_i = (t_i, y_i)$ comprising the row vector of predictors $t_i = (t_{i1}, \ldots, t_{ip})$ and the class label $y_i \in 1, \ldots, G$. The classifier can then be described in terms of its predictive function $\eta(t, \mathbf{x})$, which allocates a class to a new predictor vector t, based on the training set $\mathbf{x}$.

We often wish to estimate the misclassification or error rate we could expect, if the classifier were asked to predict the class of a new set of predictors drawn from the same distribution as the original dataset. Let (t, y) be a new point drawn at random from F and define the zero-one loss function Q for the classifier η as follows:

$$Q[y,\eta(t,\mathbf{x})] = \begin{cases} 0 & \eta(t,\mathbf{x}) = y \\ 1 & \eta(t,\mathbf{x}) \neq y \end{cases} \quad (1)$$

*To whom correspondence should be addressed.

We define the conditional true error rate Err of η to be the expectation of Q over F given $\mathbf{x}$ (Efron, 1983).

There are many ways to measure and report the misclassification rate of a classifier when applied to a labelled test set of data. The class of each test observation is predicted based on the selected predictors, then compared against the given label. Q is the simplest type of error function which treats all errors as equally important.

It is generally informative to decompose the misclassification rate into a rate for each (true) class. This is particularly useful when the observed number of data points per class is unequal. For example, if a non-diseased class contains 80% of the data and a diseased class contains 20%, then the trivial classifier which predicts every observation to be non-diseased will have a 20% misclassification rate. Examined more closely, it will have a 0% rate of false positives and a 100% rate of false negatives. Observational studies will often produce this type of unbalanced data since some classes of response will be rarer than others in the population of interest.

The cost of errors in misclassifying observations may also vary from (true) class to class, and we may wish to report an overall estimated error rate or risk which variously weights the errors on each class. This can be done retrospectively if the error rate on each class is reported. In a more sophisticated version, we may have a matrix of misclassification costs $C = \{c_{gh}, g, h = 1, \ldots, G\}$, where c_{gh} is the cost of misclassifying a data point of class g into class h. This can be applied retrospectively if a matrix of misclassification rates is reported.

For two class problems, Wessels *et al.* (2005) suggest reporting the average of the sensitivity and the specificity so that the aforementioned trivial classifier does not appear too successful. Here, we report estimates of the simple overall misclassification rate, error rates for each class and the average of the class error rates. The latter can also be motivated by decision-theoretic considerations regarding the construction of a Bayes optimal rule (McLachlan *et al.*, 2004) (p. 188). Let $\pi = \{\pi_g, g = 1, \ldots, G\}$ be the true proportions of each class in the population of interest. These will often be known with high precision, but sometimes must be estimated from the data. Let $\hat{\pi}_g$ be the observed proportion of responses in class g. For simplicity, assume that all misclassifications of a given class have equal cost, i.e. $c_{gh} = c_g, g \neq h$ and that $c_{gg} = 0, \forall g$ (see McLachlan (1992) p. 8 for more generality). If the error rate conditional upon the true class being g is Err_g, then the expected cost per observation is $\sum_g \pi_g c_g \mathrm{Err}_g$. As argued by McLachlan *et al.* (2004), misclassification cost is often nearly inversely proportional to relative frequency, so $\pi_g c_g$ may be near constant for all g. In this case, the expected cost will be approximately a multiple of the average of the class error rates Ea, so an estimate of this is a useful summary measure.

2.2 Cross-validation

Some of the most popular methods for estimating error rates are cross-validation (Breiman *et al.*, 1984; Stone, 1974), the bootstrap (Efron, 1983), the holdout method (McLachlan, 1992) (p. 341) and the 0.632 estimator (Efron, 1983; Efron and Tibshirani, 1997). All of these rely on training the classifier based on a subset of the data, and testing it on a separate subset of the data. Here, we consider only cross-validation since it is popular and effective (Molinaro *et al.*, 2005) and uses the data efficiently and is almost unbiased when used correctly (McLachlan, 1992). However, it does have significant variance when used with small sample sizes (Braga-Neto and Dougherty, 2004) and can be subject to bias if used naively, as we show in the following sections.

Cross-validation can be formalized as follows. The dataset $\mathbf{x}$ will be split into K disjoint 'folds' $\mathbf{x}^k = (\mathbf{t}^k, \mathbf{y}^k)$, $k = 1, \ldots, K, 2 \leq K \leq n$, of approximately equal sizes n^k. One then fits the classifier and tests it K times, such that in iteration k, the classifier is fitted based on the data in the training folds $\mathbf{x}'^{\mathbf{k}} = \mathbf{x} \backslash \mathbf{x}^k$ and evaluated on the test fold $\mathbf{x}^k$. The cross-validated error rate estimates $\hat{\mathrm{Err}}$ and $\hat{\mathrm{Ea}}$ are obtained by averaging over the performance on the K test folds as follows:

$$\hat{\mathrm{Err}} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n^k} Q[y_i^k, \eta(t_i^k, \mathbf{x}'^{\mathbf{k}})]}{n} \tag{2}$$

$$\hat{\mathrm{Ea}} = \frac{1}{g} \sum_{i=1}^{g} \left(\frac{\sum_{k=1}^{K} \sum_{l=1}^{n^k} Q[y_l^k, \eta(t_l^k, \mathbf{x}'^{\mathbf{k}})] \mathbb{I}(y_l^k, i)}{\sum_{l=1}^{n} \mathbb{I}(y_l, i)} \right) \tag{3}$$

where $\mathbb{I}()$ is the identity function, being 1 when its arguments are equal and 0 otherwise.

The simplest method of forming the folds is to split the randomly ordered data into K pieces with the largest fold containing at most one element more than the smallest. The bias due to the uneven distribution of classes within folds can be reduced by attempting to balance or stratify the folds, so that the empirical distribution of classes in each fold is similar to that of the whole dataset (Breiman *et al.*, 1984).

In standard K-fold cross-validation, folds of size $\lfloor n/K \rfloor$ are created by sampling from the data without replacement and each of the remaining n mod K data points is assigned randomly to a different fold. In stratified or balanced cross-validation (Breiman *et al.*, 1984) (p. 246), the data are first ordered by the response value or class. This list is broken up into $\lfloor n/K \rfloor$ bins each containing K points with many similar response values. Any remaining points at the end of the list are assigned to an additional bin. A fold is formed by sampling one point without replacement from each of the bins. Except for the ordering of the data, this is equivalent to standard cross-validation.

For a classifier η and dataset $\mathbf{x}$, we define the bias in the estimation of the error rate using cross-validation to be: $B = E_F(\hat{\mathrm{Err}} - \mathrm{Err})$. This is typically intractable, but some contributing components in $\hat{\mathrm{Err}}$ can be described and efforts made to reduce $|B|$.

The number of folds K can be any integer between 2 and n, the number of data points. The use of cross-validation to estimate error introduces a small positive component into B since each training set has mean size $n(K-1)/K$ rather than the n used to construct the final classifier. As K is reduced, the mean training set size shrinks, and this positive bias component grows. The training sets also become more different from each other, which tends to reduce the variance of the error estimate. It is common practice (e.g. McLachlan *et al.*, 2004 p. 214) to

compromise between minimizing the bias and variance by using $K = 5$ or 10 folds.

2.3 Selection bias

Cross-validation can be used to estimate model or classifier parameters as well as perform model and variable selection. However, combining these steps with error estimation for the final classifier can lead to bias unless one is particularly careful. 'External' cross-validation (Ambroise and McLachlan, 2002) (1-external cv) leaves out a single 'test fold' of the data, selects the model, variables and parameters based on the remaining 'training folds' and then evaluates the misclassification rate on the test fold. When averaged over K folds, this should provide a nearly unbiased estimate of the true error rate of the final classifier.

'Selection bias' (McLachlan *et al.*, 2004) (p. 218) can occur when cross-validation is used 'internally'. In this case, all the available data are used to select a subset of the available predictors. This subset of predictors is then fixed and the error rate is estimated by cross-validation.

2.4 Optimization bias

When largely following the above advice, it is still easy to allow a subtler bias to emerge, which we call 'optimization bias'. This can occur if cross-validation or other methods are used to estimate the error rate for multiple values of a set of free parameters, and then the set of parameter values with the lowest (optimal) estimated error rate is chosen for use in the final classifier. The free parameters can be involved in any aspect of model selection, variable selection or model fitting and include parameters as general as the index of a model or a variable subset. This method is reasonable for choosing the final classifier, but provides a downwardly biased estimate of its error rate. Varma and Simon (2006) have independently investigated this bias and call it 'parameter selection bias'.

As an example of how one might incur optimization bias, assume we have a procedure which, given a fixed $b \geq 0$, can select b predictors and fit a classifier based on the available data. The error rate for this classifier can be estimated using cross-validation. However, we may then decide to also choose an optimal value $b^{\dagger}$; from a set of values $\{b_r, r = 1, \ldots, l\}$ via $b^{\dagger} = \arg\min \hat{\mathrm{Err}}(b_r)$.

For each r, the error estimate $\hat{\mathrm{Err}}(b_r)$ will be nearly unbiased, but the estimate $\hat{\mathrm{Err}}(b^{\dagger})$ is now slightly biased. This happens because the same data is used to both estimate the error rate and to select a parameter, namely b. For similar examples involving the use of SVMs with recursive feature elimination on gene expression data, see Zhu *et al.* (2007).

Stone (1974) (p. 115) described how to carry out cross-validatory assessment of cross-validatory choice in his seminal paper. While describing only leave-one-out (LOO) cross-validation, he made clear that while one level of cross-validation is satisfactory to optimize the set of free parameters for the final classifier, a separate two-level cross-validation is needed to estimate its error rate. Failure to do this leads to optimization bias.

Two-level external cross-validation (2-external cv) can be used to avoid both selection and optimization bias in these circumstances. By two-level external cross-validation, we mean the following. At the top level, one of K_1 folds of data is left out for the purpose of assessing the error rate of the finished classifier. At the lower level, K_2-fold cross-validation is then performed on the remaining data to select the optimal value of any free parameters. When all parameters are selected, the classifier can be tested on the left out fold at the top level. By repeating this for all K_1 folds at the top level, one can construct a cross-validatory assessment of the cross-validatory choice. The same two-level procedure can be used with any method for estimating the error rate, where there are free parameters to be chosen and an overall assessment of error rate, is desired. If using cross-validation it is easiest to choose $K_2 = K_1 - 1$, so that the same fold structure can be used for both levels. The use of two levels of cross-validation to avoid bias is also discussed by Dudoit and Fridlyand (2003), Statnikov *et al.* (2005) and Wessels *et al.* (2005). If instead the whole model selection, variable selection and parameter fitting process is performed without cross-validation, then only one level of external cross-validation is needed to estimate error rates of prediction.

Optimization bias will increase in magnitude with the variability of the error estimate and with the number of parameter values considered, especially those whose true error rate is near the minimum value (within the range of variability). In the analysis of gene expression, SNP (single nucleotide polymorphism) chip and mass spectroscopy data, the number of available predictors is large, so careful variable selection is needed to avoid optimization bias.

Sharma *et al.* (2005) built a system to classify patients into those with or without breast cancer based on gene expression levels in blood. They considered 1368 genes and used the NSC method of Tibshirani *et al.* (2002) to both select a subset of genes for classification and for the classification itself. Sharma *et al.* (2005) used 10-fold cross-validation to estimate a prediction error rate of 18% based on 102 labelled samples. However, they did this for multiple gene subset sizes controlled through a threshold parameter. The optimal value of the threshold was chosen to be that producing the lowest cross-validation error rate estimate. They reported the error rate estimate for this optimal choice and constructed the final classifier using the whole dataset.

Based on the discussion above, it seems likely that these authors have incurred optimization bias. They could have avoided this bias by choosing the threshold using cross-validation based on a subset of the data as some of the same authors did in Tibshirani *et al.* (2003). This is the holdout method of assessing cross-validatory choice, which is effectively one fold of 2-external cv with a small K_1. One could make even better use of the data by completing a two-level external cross-validation. The resulting estimates would be more accurate since they would be based on K_1 holdout estimates, with each observation being used once in a test set.

A number of authors (e.g. McLachlan *et al.*, 2004 (p. 240), Dabney, 2005) estimate and report error rates for various numbers of genes b where for each value of b, an optimal subset of b predictors is selected. The natural response to a table of estimated error rates for various values of b is to choose $b^{\dagger}$ with the minimal estimated error rate and select $b^{\dagger}$ variables in the final classifier. In the absence of other information, one is also

likely to take the reported error rate estimate for this number of genes to be indicative of that final classifier's error rate. As discussed previously, this estimate will be subject to optimization bias due to the process of choosing the optimal value for b. The study of McLachlan *et al*. (2004) was repeated using two-level cross-validation to avoid optimization bias, with the results reported in Zhu *et al*. (2007).

Varma and Simon (2006) investigated the same bias by applying the NSC and SVM to simulated datasets containing two classes of exactly 20 points each. Using 1-external cv, they estimated a bias of -0.122 with $K=10$ and an NSC and a bias of -0.083 with $K=n$ and an SVM. Using 2-external cv, they estimated biases of 0.042 and 0.033 for the NSC and SVM, respectively, which they attributed to cross-validation leaving some data out.

2.5 Sampling effects on error rate estimation

If a dataset of size n consists of a sample of two classes, each occurring with probability $\pi_1 = \pi_2 = 0.5$, then in most cases, the number of observations from each class will be different, i.e. $\hat{\pi}_1 \neq \hat{\pi}_2$. Assuming independence, the number of observations n_1 in class 1 can be described by the binomial distribution, i.e. $n_1 \sim \text{Bin}(n, \pi_1 = 0.5)$, and the size of the second class is simply $n_2 = n - n_1$. Then $E(n_1) = n\pi_1$ and $E(n_2) = n(1-\pi_1)$. For larger samples and π_1 not too close to 0 or 1, the binomial distribution $\text{Bin}(n, \pi_1)$ can be approximated by a normal distribution with mean $n\pi_1$ and variance $n\pi_1(1-\pi_1)$.

Let m_1 be the absolute difference in class size 1 from the expected size, i.e. $m_1 = |n_1 - E(n_1)|$, and similarly $m_2 = |n_2 - E(n_2)|$. Using the normal approximation to the binomial distribution, the distributions for m_1 and m_2 are both half-normal. For m_1, this is the renormalized right-half of a normal distribution with mean 0 and variance $n\pi_1(1-\pi_1)$. As a half-normal distribution, it has mean $E(m_1) = \sqrt{2n\pi_1(1-\pi_1)/\pi}$ and variance $\text{var}(m_1) = n\pi_1(1-\pi_1)(1-2/\pi)$ (Johnson *et al*., 1994).

As an example, if we have a sample size of 60 with two equiprobable classes ($\pi_1 = \pi_2 = 0.5$), the mean absolute difference in class size from the expected 30 is 3.1, with a variance of 5.45. Thus, class sizes of 33 and 27 would be typical for a random sample from this population and on average one class is 22% larger than the other. Through balanced cross-validation, one might expect the following numbers of each class in each fold (n_1, n_2): (4,2),(4,2),(4,2),(3,3),(3,3),(3,3),(3,3),(3,3),(3,3),(3,3). Hence 3 out of 10-folds would have a majority from the class which was larger in the sample; the other 7-folds would have equal numbers of each class. These types of sampling effects have an impact on the estimation of classification error rates and their interpretation.

In the above example, if the available predictors were independent of each other and of the response and the method of classification ignored the predictors, it would be likely to use the apparently different class probabilities, as estimated from the training data. Each fold of 10-fold cross-validation would contain on average 3.3 of one class and 2.7 of the other. In the combined training folds, one would expect 29.7 of the larger class and 24.3 of the smaller. Even with n-fold (or LOO) cross-validation, the excluded data point will have the same class as the larger class in the training folds in 33/60 $\approx$55% of cases. A classifier that assigns every point to the larger class in the training set can thus be expected to show an error rate of 45% under this type of cross-validation. However, we know that it would achieve an expected error rate of 50% if applied to new data from the same population or underlying distribution.

Efron and Tibshirani (1997) (p. 552) define the no-information error rate γ to be the error rate if the true response or classification is independent of all the predictors. They estimate γ by

$$\hat{\gamma} = \sum_{g=1}^{G} \hat{\pi}_g (1 - \hat{q}_g), \tag{4}$$

where, $\hat{\pi}_g$ is the observed proportion of responses in class g and $\hat{q}_g$ is the observed proportion of predictions in class g. In the above example, we might have $\hat{\pi}_1 = 33/60$, $\hat{q}_1 = 1$, $\hat{\pi}_2 = 27/60, \hat{q}_2 = 0$, so $\hat{\gamma} = 27/60 \approx 0.45$.

Consider the following three trivial classifiers. These are simple to implement and use no predictor information. They are presented here in order of expected increasing error rate. The first trivial classifier can be seen as providing a baseline for classifier error rates.

Trivial classifier 1 (TC1): classify all observations as belonging to the largest class in the sample. Without loss of generality let this be class 1, so $\hat{q}_1 = 1$, $\hat{\gamma}_{\text{TC1}} = 1 - \hat{\pi}_1$ and $\text{Err}_{\text{TC1}} = 1 - \pi_1$.

Trivial classifier 2 (TC2): classify observations randomly with class probabilities equal to the sample proportions, so $q_g = \hat{q}_g = \hat{\pi}_g, g = 1, \ldots, G$. Then $\hat{\gamma}_{\text{TC2}} = 1 - \sum_{g=1}^{G} \hat{\pi}_g^2$ and $\text{Err}_{\text{TC2}} = 1 - \sum_{g=1}^{G} \pi_g \hat{\pi}_g$.

Trivial classifier 3 (TC3): classify observations randomly with equal probability for each class, so $q_g = 1/G, g = 1, \ldots, G$. If we use q_g in equation (4) instead of $\hat{q}_g$, we obtain: $\hat{\gamma}_{\text{TC3}} = (G-1)/G = \text{Err}_{\text{TC3}}$.

Using $\hat{\pi}_1 \geq \hat{\pi}_g, g \neq 1$ and the Cauchy–Schwartz inequality, it can be shown that $\hat{\gamma}_{\text{TC1}} \leq \hat{\gamma}_{\text{TC2}} \leq \hat{\gamma}_{\text{TC3}}$. If $\pi_1 \geq \pi_g, g \neq 1$, then $\text{Err}_{\text{TC1}} \leq \text{Err}_{\text{TC2}}$ and if $\pi_1 \geq 1/G$, $\text{Err}_{\text{TC1}} \leq \text{Err}_{\text{TC3}}$. The true averages of the class error rates are the same for all three trivial classifiers and will equal the estimated average for trivial classifier 3, i.e. $\hat{\text{Ea}}_{\text{TC3}} = \text{Ea}_{\text{TCg}} = (G-1)/G, g = 1, 2, 3$.

Efron and Tibshirani (1997) (p. 552) studied classification based on uninformative data with responses of 0 or 1 with probability 0.5. They erroneously claim that on this type of data, the leave-one-out cross-validation estimate $\hat{\text{Err}}$ for the nearest neighbour classifier would have the correct expectation of 0.5. In fact, as described above, it will generally be slightly lower for this and most other classifiers since sampling variation will usually produce one class larger than the other. Classifiers will tend to exploit this imbalance and cross-validation estimates of the error rate derived from the same dataset will be unable to correct for its effect. By default, the NSC takes the prior class probabilities to be the sample class proportions. If no genes are selected, a prior term causes the NSC to classify all observations into the largest sample class, so becoming TC1.

2.6 Permutation assessment

The permutation or randomization test is an exact test which can be used to determine a significance level for the acceptance or rejection of a null hypothesis (Good, 1994). The statistic of interest here is the estimated error rate of the classifier. The null hypothesis is that the value of this statistic does not depend upon the given set of labels, i.e. there is no meaningful relationship between the predictors and the given labels. This implies that the classifier would be expected to yield a similar estimated error rate even under a random permutation of the labels.

We can obtain a reasonable approximation by taking a subset of the possible permutations chosen via a uniform distribution over the $n!$ possible relabellings. The p-value for this test is then given by the fraction of the statistics obtained under permutation which are more extreme than the value obtained using the original labelling.

The mean of a statistic under a large number of permutations is also worth consideration. If the permutations successfully remove the relationship between predictors and response, and the trivial classifiers dominate as expected, then we can expect the permutation mean of $\hat{E}a$ to be close to $(G-1)/G$. If it is not, then the method used to estimate error is likely to be biased.

3 METHODS

We performed computer experiments to test for bias in the estimation of error rates using 1-external and 2-external cv. In addition, the experiments were designed to test for differences between $\hat{E}rr$ and $\hat{E}a$ (see Equations (2) and (3) and Section 2.1). In each case, the NSC classifier was used since this allowed a clear comparison with the results of two previous studies of interest. The NSC is implemented in a well-known R package (PAMR) which offers only 1-external cv. We wrote R code to wrap another level of cross-validation around it to allow the use of 2-external cv. We used balanced cross-validation throughout our experiments via our own code and through the routines in PAMR (http://www-stat.stanford.edu/\%7Etibs/PAM/). PAMR controls variable selection by automatically trying 30 threshold values in a linear series, with a value of 0 corresponding to no genes selected.

The 1-external and 2-external cv versions of the NSC classifier were applied to a simulated non-informative dataset and the Khan (Khan *et al*., 2001) and Sharma (Sharma *et al*., 2005) datasets. For each dataset, we estimated the simple, average and class-conditional error rates for the NSC classifier using 1-external and 2-external cv. We also recorded the number of genes selected using the optimal threshold value under 1-external cv and the average number of genes selected across the K folds under 2-external cv. Balanced cross-validation randomly allocates data values to folds, so we repeated each procedure 1000 times to reduce variability and estimated the mean and standard deviation of each of the above estimates across these repetitions. Standard errors were calculated across folds, then averaged over the repetitions.

We also carried out a series of permutation tests for each dataset. We permuted the data labels (responses) 1000 times and refit the NSC classifier under 1-external and 2-external cv. Each time we recorded the simple, average and class error rates and the number of genes selected. We also calculated the mean of each estimate over the permutations. For the non-informative dataset, permutation would be expected to make little difference to any of the estimates.

Since the true distribution is available here, we were also able to estimate the optimization bias with the NSC on sample sizes of 100 by simulating 1000 additional samples of this size and performing 1- and 2-external cv on each.

3.1 Simulated data

The non-informative simulated dataset comprised 100 data points $\{x_i = (t_i, y_i), i = 1, \ldots, 100\}$, each intended to represent an individual drawn from a population of interest. Each individual was given 2000 real-valued predictor measurements $t_i = \{t_{ij}, j = 1, \ldots, 2000\}$, with each $T_{ij} \sim N(0,1)$ and a binary response y_i with $Y_i \sim \text{Bin}(1, 0.5)$, i.e. $\pi_1 = \pi_2 = 0.5$. Hence each predictor and response value is drawn independently of all others and any relationships between predictors and response are purely due to chance. The dataset was generated randomly once and then used throughout the experiments. The number of observations in classes 1 and 2 were 53 and 47, respectively.

3.2 Khan data

Khan *et al*. (2001) described a gene expression dataset of 83 observations, each from a child who was determined by clinicians to have a type of small round blue cell tumour (SRBCT). These included the following four classes: neuroblastoma (N), rhabdomyosarcoma (R), Burkitt lymphoma (B; a subset of the non-Hodgkin lymphomas) and the Ewing's sarcoma family of tumours (E). The numbers in each class are: 18 N, 25 R, 11 B, 29 E.

For each tissue sample the levels of gene expression were estimated using a cDNA microarray. A total of 2308 genes and ESTs passed the intensity requirements imposed and the values were normalized (Khan *et al*., 2001) The full dataset is publicly available at http://home.ccr.cancer.gov/oncology/oncogenomics/. We ignored five additional observations which were not determined to be SBRCTs.

3.3 Sharma data

Sharma *et al*. (2005) described and made public a dataset containing the expression levels (mRNA) of 1368 genes from 60 blood samples taken from 56 women. Some of the blood samples were analyzed more than once in separate batches giving a total of 102 labelled blood samples. Each blood sample was labelled by clinicians, with 24 labelled as having breast cancer (BC) and 36 labelled as not having it (NC).

The supplementary section of Sharma *et al*. (2005) supplies both the raw data from macroarray measurements and batch-adjusted data, obtained using ANOVA. The authors found a clear batch effect and removed it for their analysis, so we also used only the batch-adjusted data. To avoid consideration of the method of aggregation, we chose to use just one measurement per blood sample and ignore the others. Hence the Sharma data set used here is a randomly selected subset of 60 observations, rather than the whole 102. The subset used here is publicly available on the website.

Table 1 lists our calculations of the estimated no-information error rates $\hat{\gamma}$, and the true error rates Err for the three types of trivial classifiers described in Section 2.5 on the three datasets. The missing

Table 1. Calculated estimated no-information error rate $\hat{\gamma}$ and true error rate Err for three types of trivial classifiers on the three datasets studied here

Dataset	$\hat{\gamma}_{TC1}$	Err_{TC1}	$\hat{\gamma}_{TC2}$	Err_{TC2}	$\hat{\gamma}_{TC3}$	Err_{TC3}
Simulated	0.47	0.5	0.498	0.5	0.5	0.5
Khan	0.651	–	0.723	–	0.75	0.75
Sharma	0.4	–	0.48	–	0.5	0.5

Table 2. Error rates of the NSC classifier on the simulated non-informative (2 class) dataset, as estimated using 1-external and 2-external 10-fold cross-validation

Test	$\hat{\text{Err}}$	$\hat{\text{Ea}}$	$\hat{\text{Err}}_1$	$\hat{\text{Err}}_2$	Number of genes
1-external cv	0.435	0.461	0.023	0.899	9.55
1-external cv sd	0.0176	0.0192	0.0308	0.056	6.05
1-external cv se folds	0.0229	0.0215	0.0149	0.040	0.70
1-ext cv perm mean	0.420	0.435	0.180	0.689	409
1-ext cv perm mean sd	0.0441	0.051	0.129	0.201	608
2-external cv	0.472	0.498	0.0548	0.942	33.9
2-external cv sd	0.024	0.023	0.0449	0.0328	59.1
2-external cv se folds	0.0278	0.0223	0.0384	0.0349	27.8
2-ext cv perm mean	0.487	0.503	0.229	0.778	376
2-ext cv perm mean sd	0.0491	0.0516	0.0974	0.139	343

The non-permutation results give the means, standard deviations (sd) and fold-based standard errors (se) from 1000 repetitions of balanced cross-validation. The permutation (perm) results are based on a single instance of cross-validation for each of 1000 permutations of the labels. The numbers of selected genes are based on the 10 outer cross-validation folds.

entries for true error rates could be filled in if one knew the prior probabilities of class membership π_g for the populations sampled by Khan *et al.* (2001) and Sharma *et al.* (2005). These may be available, but are beyond the scope of this article.

4 RESULTS AND DISCUSSION

4.1 Results on simulated data

The results on the simulated dataset are detailed in Table 2. They show that 1-external cv yielded mean (standard error) estimates of $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ of 0.435 (0.0229) and 0.461 (0.0215), respectively. Given the non-informative model which generated this dataset, we can be confident that the true Err and Ea and their class-conditional counterparts would all be 0.5. Using 2-external cv, the mean (standard error) estimates of $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ were 0.472 (0.0278) and 0.498 (0.0223), respectively. The difference between the 1-external and 2-external cv estimates can be largely attributed to optimization bias.

As discussed in Section 2.5, the estimate $\hat{\text{Err}}$ is also influenced by differences between the sample and true class proportions. Although the true proportions for the two classes were equal, the sample proportions were 0.53 and 0.47. Since class 1 has the larger sample proportion, the baseline no-information error rate $\hat{\gamma}_{\text{TC1}} = 0.47$, which is similar to the $\hat{\text{Err}}$ estimate found using 2-external cv. On the basis of these results, the optimization bias seems to have reduced both $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ by around 0.04 on this dataset.

This example also illustrates the value of estimating $\hat{\text{Ea}}$ in addition to $\hat{\text{Err}}$. $\hat{\text{Ea}}$ was unaffected by the difference between the sampling and true proportions and so offers a valuable diagnostic tool for determining whether or not a given method of estimating error is biased when the true proportions are unknown.

The NSC returned large p-values in the range 0.34–0.66 for $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ with both 1-external and 2-external cv under the permutation test on this dataset. This was expected since the given labelling was assigned randomly and uninformatively. The average values of $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ with the given labelling were slightly different to the permutation mean values, but fell well inside a standard deviation.

Under 2-external cv, the mean estimate of $\hat{\text{Err}}$ with permuted labels was 0.487, which is slightly above the 0.47 $\hat{\gamma}_{\text{TC1}}$ baseline. The mean estimate of $\hat{\text{Err}}$ using 1-external cv under permuted labellings was 0.420. The 2-external cv estimate $\hat{\text{Ea}}$ was 0.503, which is close to the expected 0.5, while 1-external cv produced an $\hat{\text{Ea}}$ of 0.435. Hence the use of 1-external cv, seems to incur an optimization bias of around -0.07 in both $\hat{\text{Err}}$ and $\hat{\text{Ea}}$.

Based on the 1000 additional datasets of size 100, the mean (standard error over the 1000) results for $\hat{\text{Err}}$ were 0.410 (0.0014) and 0.476 (0.0018) for 1- and 2-external cv, respectively. For $\hat{\text{Ea}}$, the respective values were 0.439 (0.0016) and 0.503 (0.0017). Hence, for this true distribution and sample sizes of 100, we estimate the optimization bias in $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ under 1-external cv to be -0.06.

4.2 Results on Khan and Sharma data

The results on the Khan and Sharma datasets are detailed in Table 3 and 4. On the Khan dataset, 1-external cv produced mean (standard error) results of 0.00026 (0.00027) for $\hat{\text{Err}}$ and 0.00023 (0.00023) for $\hat{\text{Ea}}$. The mean (standard error) estimates for $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ from 2-external cv were 0.00717 (0.0069) and 0.00563 (0.0052), respectively. Tibshirani *et al.* (2002) reported an estimated error rate of zero for the NSC using a separate test set, but the 2-external cv estimate given here is expected to be more accurate. On this dataset, optimization bias reduced both $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ by an order of magnitude under 1-external cv.

On the Sharma dataset the mean (standard error) estimates of $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ using 1-external cv, were 0.186 (0.0494) and 0.201 (0.0537), respectively. Using 2-external cv the estimates for $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ were 0.212 (0.0524) and 0.232 (0.0576), respectively. These differences of around 3% can be attributed to optimization bias. Due to differences with the original dataset, the results here cannot be directly compared with those of Sharma *et al.* (2005), but their reported error rates based on 1-external cv are likely to include a similar level of bias.

For both the Khan and Sharma datasets, the permutation tests rejected the null hypothesis with p-values < 0.001 for $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ estimated using 1-external and 2-external cv. This is unsurprising, and supports an association between the predictors and the given labels.

As with the simulated data set, it is more interesting to consider the permutation mean of $\hat{\text{Err}}$ and $\hat{\text{Ea}}$. The effects of optimization bias are illustrated for the Sharma dataset in Figure 1 through the different distributions of $\hat{\text{Err}}$ and $\hat{\text{Ea}}$ as estimated by 1-external and 2-external cv under 1000 permutations of the labels. The baseline error rates $\hat{\gamma}_{\text{TC1}}$ for trivial classifier 1 are 0.651 and 0.4 for the Khan and Sharma datasets, respectively. These values are approximately midway between the permutation means of $\hat{\text{Err}}$ using 1-external and 2-external cv on these datasets.

Table 3. Error rate results for the given labels and 1000 permutations using the NSC classifier on the Khan (4 class) dataset, estimated using 1-external and 2-external 10-fold cross-validation

Test	$\hat{E}rr$	$\hat{E}a$	$\hat{E}rr_B$	$\hat{E}rr_E$	$\hat{E}rr_N$	$\hat{E}rr_R$	Number of genes
Reported	0	0	0	0	0	0	43
1-external cv	0.00026	0.00023	0.00027	0.00066	0	0	242.7
1-external cv sd	0.00177	0.00017	0.00497	0.00471	0	0	192.5
1-external cv se folds	0.00027	0.00023	0.0003	0.00063	0	0	2.8
1-ext cv perm mean	0.631	0.717	0.951	0.144	0.933	0.838	108.1
1-ext cv perm mean sd	0.0294	0.0461	0.124	0.192	0.138	0.201	369.5
2-external cv	0.00717	0.00563	0.0030	0.0189	0.00017	0.00048	198
2-external cv sd	0.00728	0.00667	0.0172	0.0194	0.00304	0.00436	49.5
2-external cv se folds	0.0069	0.0052	0.0025	0.0184	0.0002	0.00043	56.0
2-ext cv perm mean	0.667	0.751	0.976	0.163	0.965	0.900	98.6
2-ext cv perm mean sd	0.031	0.030	0.072	0.141	0.079	0.130	195

The reported values are from Tibshirani *et al*. (2002). The class-specific error rates are abbreviated by subscripts as follows: Burkitt lymphoma (B), Ewing's sarcoma (E), neuroblastoma (N), rhabdomyosarcoma (R).

Table 4. Error rate results for the given labels and 1000 permutations using the NSC classifier on the Sharma (2 class) dataset, estimated using 1-external and 2-external 10-fold cross-validation

Test	$\hat{E}rr$	$\hat{E}a$	$\hat{E}rr_{BC}$	$\hat{E}rr_{NC}$	Number of genes
Reported A	0.217	0.230	0.292	0.167	37
Reported B	0.167	0.188	0.292	0.0833	25
1-external cv	0.186	0.201	0.277	0.125	53.4
1-external cv sd	0.0156	0.0174	0.0354	0.0208	44.6
1-external cv se folds	0.0494	0.0537	0.092	0.057	2.03
1-ext cv perm mean	0.384	0.465	0.866	0.063	72.3
1-ext cv perm mean sd	0.0278	0.0558	0.212	0.108	215
2-external cv	0.212	0.232	0.329	0.135	56.1
2-external cv sd	0.0259	0.0288	0.051	0.0258	19.5
2-external cv se folds	0.0524	0.0576	0.0999	0.0574	13.4
2-ext cv perm mean	0.421	0.503	0.912	0.0937	65.2
2-ext cv perm mean sd	0.0368	0.0415	0.129	0.0926	105

The reported values are from Sharma *et al*. (2005) and are based on the full 102 observations. Their first analysis (A) contained three non-decisions, which were treated as errors. Their second analysis (B) made a decision (classification) for every observation. The class-specific error rates are abbreviated by subscripts as follows: breast cancer (BC), no breast cancer (NC).

The permutation means of $\hat{E}a$ are far more clearly in favour of 2-external cv. Under label permutations, the average error rates Ea should be centred around 0.5 on the Sharma dataset and 0.75 on the Khan dataset. The estimates $\hat{E}a$ derived from 2-external cv were centered in this way, but those from 1-external cv were noticeably lower. 1-external cv gave mean results of 0.717 and 0.465 on the Khan and Sharma datasets, respectively, while 2-external cv gave results of 0.751 and 0.503, respectively. The 2-external cv estimates are highly accurate and the 1-external cv estimates are noticeably biased downward, giving an assessment of classifier accuracy which is too optimistic by around 3%.

Mean class error rates under permutation were very high for the smaller observed classes (B, N and R on the Khan dataset and BC on the Sharma dataset), which indicates that the NSC may have frequently become trivial classifier 1. By checking the raw results, we found that under 1-external cv the NSC became in effect TC1 in 38% of cases on the Khan dataset and in 60% of cases on the Sharma dataset. Under 2-external cv there is another layer of diversity, but class error rates matching TC1 were seen in 16% of cases on the Khan dataset and in 27% of cases on the Sharma dataset. This shows that trivial classifiers are relevant in deriving a baseline error rate.

5 CONCLUSIONS

We have quantified the bias and precision of error rates in classification based upon gene expression data from simulations and using real datasets, and have shown how common methods of estimation can lead to bias. We have proposed

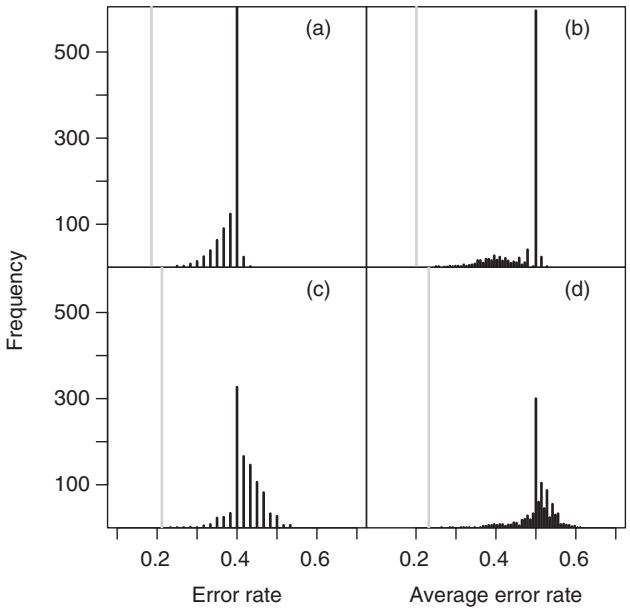

Fig. 1. Histogram of error rates estimated using 1- and 2-external cv under 1000 permutations of the labels on the Sharma dataset. Superimposed are the estimated error rates using the original labels. 1-external cv was used to estimate the error rate in (**a**) and the average error rate in (**b**). 2-external cv was used to estimate the error rate in (**c**) and the average error rate in (**d**).

a novel permutation approach to detect bias and shown the effectiveness of two-level external cross-validation in reducing it.

We urge all investigators performing classification tasks to calculate and examine the permutation mean of the average of the estimated class error rates $\hat{E}a$. If this is noticeably below the expected $(G-1)/G$, the procedure may be incurring selection or optimization bias. These can be avoided by using two-level external cross-validation.

6 ACKNOWLEDGEMENTS

The authors appreciate discussions with Geoff McLachlan, David Duffy, Ross McVinish, Clair Alston and Georgia Chenevix-Trench and the helpful comments of two anonymous reviewers. This research was primarily supported by the ARC Center for Complex Dynamic Systems and Control CEO348165 and NHMRC Medical Bioinformatics, Genomics and Proteomics Program Grant 389892.

Conflict of Interest: none declared.

REFERENCES

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, **99**, 6562–6566.

Braga-Neto,U.M. and Dougherty,E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.

Breiman,L. *et al.* (1984) *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Dabney,A.R. (2005) Classification of microarrays to nearest centroids. *Bioinformatics*, **21**, 4148–4154.

Dudoit,S. and Fridlyand,J. (2003) Classification in microarray experiments. In Speed,T.P. (ed.), *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall, Boca Raton, pp. 93–158.

Efron,B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.*, **78**, 316–331.

Efron,B. and Tibshirani,R. (1997) Improvements on cross-validation: The .632 + bootstrap method. *J. Am. Stat. Assoc.*, **92**, 548–560.

Good,P. (1994) *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New York..

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.

Johnson,S. *et al.* (1994) *Continuous Univariate Distributions*, 2nd edn, Vol. 1, Wiley, New York.

Khan,J. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.*, **7**, 673–679.

McLachlan,G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

McLachlan,G.J. *et al.* (2004) *Analyzing Microarray Gene Expression Data*.Wiley, Hoboken, NJ, USA.

Molinaro,A.M. *et al.* (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301–3307.

Sharma,P. *et al.* (2005) Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. *Breast Cancer Res.*, **7**, R634–R644.

Stone,M. (1974) Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B*, **36**, 111–147.

Statnikov,A. *et al.* (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, **99**, 6567–6572.

Tibshirani,R. *et al.* (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.*, **18**, 104–117.

Varma,S. and Simon,R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**, 91.

Wessels,L.F.A. *et al.* (2005) A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**, 3755–3762.

Zhu,J.X. *et al.* (2007) On selection biases with prediction rules formed from gene expression data. *J. Stat. Plan. Inference*, in press.