STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:5320–5334
Published online 11 July 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2968

A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification

Wenyu Jiang[1,2,*,†] and Richard Simon[1]

[1] *Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, 6130 Executive Boulevard, Rockville, MD 20852, U.S.A.*
[2] *Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8*

SUMMARY

This paper first provides a critical review on some existing methods for estimating the prediction error in classifying microarray data where the number of genes greatly exceeds the number of specimens. Special attention is given to the bootstrap-related methods. When the sample size n is small, we find that all the reviewed methods suffer from either substantial bias or variability. We introduce a repeated leave-one-out bootstrap (RLOOB) method that predicts for each specimen in the sample using bootstrap learning sets of size ln. We then propose an adjusted bootstrap (ABS) method that fits a learning curve to the RLOOB estimates calculated with different bootstrap learning set sizes. The ABS method is robust across the situations we investigate and provides a slightly conservative estimate for the prediction error. Even with small samples, it does not suffer from large upward bias as the leave-one-out bootstrap and the 0.632+ bootstrap, and it does not suffer from large variability as the leave-one-out cross-validation in microarray applications. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: bootstrap; prediction error; class prediction; microarray data; learning curve; feature selection

1. INTRODUCTION

DNA microarray technology is now commonly used in cancer research and has an increasing impact on cancer treatment, diagnosis and prognosis. One major application of this technology is tumor classification. A typical statistical problem in the area of tumor classification is to classify tumor tissues (or patients) into predetermined classes of malignancies based on their gene expression profiles. Prediction rules are developed using the observed gene expression data and used to predict the tumor classes for future observations. An accurate prediction rule helps to improve the rates of

[*] Correspondence to: Wenyu Jiang, Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8.
[†] E-mail: wjiang@mathstat.concordia.ca, jiangwen@mail.nih.gov

Received 23 May 2006
Accepted 8 May 2007

correct diagnosis and proper treatment assignments for cancer patients. Performance of the class prediction procedures is usually assessed by prediction error rates.

In this paper, we focus on methods for estimating the prediction error in class prediction in microarray data analysis. A microarray experiment can monitor expression patterns of thousands of genes simultaneously. But due to their cost and complexity, such experiments are often restricted to a small number of specimens. Microarray analysis presents a unique challenge in statistics which is characterized by a small sample size n and a large number p of features (variables), often with $n \ll p$. In the traditional $n > p$ scenario, cross-validation methods [1] are widely used to estimate the prediction error. Various bootstrap methods such as the ordinary bootstrap, the leave-one-out bootstrap and the 0.632+ bootstrap are proposed and compared by Efron [2] and Efron and Tibshirani [3, 4]. Breiman [5] proposes an out-of-bag (OOB) estimation of the prediction error rate, which is a byproduct from a bagging predictor [6].

A microarray analysis typically starts with a feature (variable) selection procedure that determines the collection of genes to include in prediction modeling. When resampling methods are applied to microarray data, it is crucial to perform feature selection within each resampling step when estimating the prediction errors—a process known as *honest* performance assessment [7]. Molinaro *et al.* [8] compared several cross-validation methods, split-sample methods and the 0.632+ bootstrap for high-dimensional genomic studies. In this paper, we compare a number of existing bootstrap methods, the OOB estimation and a bootstrap cross-validation method [9], for estimating prediction errors when the number of features greatly exceeds the number of specimens. Such a study is needed to examine the performance of the bootstrap-related methods in microarray applications; the necessity of repeated feature selection is often overlooked in previous work.

It is commonly acknowledged that there is a bias–variance trade-off in estimating prediction errors. The methods reviewed in this paper suffer from either large upward or downward bias or very large variability in microarray situations. In the conventional $n > p$ situation, the 0.632+ bootstrap is very popular for having low variability and only moderate bias. However, the study in this paper and the work of Molinaro *et al.* [8] suggest that the 0.632+ bootstrap can run into problems in the $n < p$ situation. We propose an adjusted bootstrap (ABS) method, which performs robustly in various situations and achieves a good compromise in the bias–variance trade-off.

2. A REVIEW OF THE METHODS FOR PREDICTION ERROR ESTIMATION

In a microarray class prediction problem, we observe $x_i = (t_i, y_i)$, $i = 1, \ldots, n$, on n independent subjects, where t_i is a p-dimensional vector containing the gene expression measurements and y_i is the response for subject i. The observations $x_1, \ldots, x_n$ can be viewed as realizations of an underlying random variable $X = (T, Y)$. With dichotomous outcome, the response variable Y takes 0 or 1 values distinguishing the two classes. A prediction rule (model) $r(\cdot, x^{\text{learn}})$ is developed based on the information in the learning set x^{learn}. The true prediction error ($e_n = E[I\{Y \neq r(T, x)\}]$) is the probability that the prediction model built on the observed data $x = (x_1, \ldots, x_n)$ misclassifies a future item following the same random mechanism as X.

When the prediction rule is built for the observed data, the prediction accuracy should ideally be assessed on an independent large test set. But this is often impossible because of the relatively small sample sizes in microarray experiments. Methods for estimating prediction errors rely on partitioning or resampling the observed data to construct the learning and test sets. With a huge number of features, the prediction rules $r(\cdot, \cdot)$ contain two key steps: the feature selection and the class prediction (discrimination) step. Feature selection is administered prior to the class prediction step for every learning set. Failure to include feature selection in resampling steps results in serious downward bias in estimating the prediction error and overly optimistic assessment of the prediction rule [7, 10, 11]. Methods for class prediction include various versions of discriminant analysis, nearest neighbor classification, classification trees, etc. A comprehensive comparison of the class discrimination methods was conducted by Dudoit *et al.* [12]. In this section, we concentrate on the bootstrap-related methods for estimating prediction errors. For comparison purpose, we also include the resubstitution and the leave-one-out cross-validation (LOOCV) methods.

The *resubstitution* estimate is known to underestimate the prediction error for using the same data set to build and evaluate the prediction rule. Moreover, an overfitting problem arises in the $n<p$ situation; it is often possible to pick a number of features to build a model which fits the data perfectly, but such a model not very useful in predicting future observations.

The *leave-one-out cross-validation* estimate can be expressed as

$$\hat{e}_n^{\mathrm{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n} I\{y_i \neq r(t_i, x_{(-i)})\}$$

where $x_{(-i)}$ represents the learning set with x_i removed. It calculates the rate of misclassified responses when predicting for each specimen using a learning set containing all other observations in the sample. Correct application of the method to high-dimensional microarray data requires feature selection for every leave-one-out learning set $x_{(-i)}$ of size $n-1$. The LOOCV produces an almost unbiased estimate for the prediction error and has been a common choice for small sample problems. The investigation of Molinaro *et al.* [8] suggests that the LOOCV method performs no worse than other cross-validation methods and split-sample methods in genomic studies with small to moderate sample sizes. However, when the sample size is small, the LOOCV method is often criticized for having very large variation. The large variability is ascribed mainly to the similarity between the leave-one-out training sets $x_{(-i)}, i=1,\ldots,n$ and the sparseness of the data. The similarity between the sets $x_{(-i)}$ results in large covariance between the terms of $\hat{e}_n^{\mathrm{LOOCV}}$ and hence increases the overall variance of the estimate.

2.1. Methods through bootstrap resampling

These methods draw bootstrap samples of size n repeatedly from the original data x by simple random sampling with replacement.

Ordinary bootstrap. This method [4] has the problem that the learning and test sets overlap. In this, a prediction rule is built on a bootstrap sample and tested on the original sample. Averaging the misclassification rates across all bootstrap replications gives the ordinary bootstrap estimate. This method seriously underestimates the prediction error since a subset of the data is used both in building and in assessing the prediction model.

Bootstrap cross-validation. This method is proposed by Fu *et al.* [9] to handle small-sample problems. The procedure generates B bootstrap samples of size n from the observed sample and then calculates a LOOCV estimate on each bootstrap sample. Averaging the B cross-validation estimates gives the bootstrap cross-validation estimate for the prediction error. The paper of Fu *et al.* [9] did not carefully address the issue of feature selection. When the method is applied to high-dimensional gene expression data, we emphasize that feature selection must be conducted on every leave-one-out learning set derived from every bootstrap sample. Since an original observation can appear more than once in a bootstrap sample, a leave-one-out learning set may overlap with the

DOI: 10.1002/sim

left-out item when the cross-validation procedure is applied on a bootstrap sample. Consequently, the bootstrap cross-validation method tends to underestimate the true prediction error.

Leave-one-out bootstrap. This procedure [2] generates a total of B bootstrap samples of size n. Each observed specimen is predicted repeatedly using the bootstrap samples in which the particular observation does not appear. In this way, the method avoids testing a prediction model on the specimens used for constructing the model. The leave-one-out bootstrap estimate is given by

$$\hat{e}_n^{\text{LOOBS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|C_i|}\sum_{b \in C_i} I\{y_i \neq r(t_i, x^{*,b})\}$$

where C_i is the collection of bootstrap samples not containing observation i and $|C_i|$ is the number of such bootstrap samples. Feature selection and class prediction should be performed on each bootstrap sample $x^{*,b}$, $b = 1, \ldots, B$.

The leave-one-out bootstrap is basically a smoothed version of the LOOCV. To see this, the bootstrap samples in C_i can be viewed as random samples of size n generated from the leave-i-out data set $x_{(-i)}$. Bootstrap samples are more different between each other than the original leave-one-out sets. Moreover, for each specimen i, the leave-one-out bootstrap method averages on the errors from the multiple predictions made on the bootstrap samples in C_i. As a result, the leave-one-out bootstrap estimate has much smaller variability than the LOOCV estimate. On the other hand, a bootstrap sample of size n contains roughly $0.632n$ distinct observations from the original sample. It is often inadequate to represent the distribution of the original data when the sample size n is small. Hence the leave-one-out bootstrap estimate tends to overestimate the true prediction error.

Out-of-bag estimation. The OOB estimate [5] for the prediction error is a by-product of bagging predictors [6]. The OOB estimate is the misclassification rate when predicting for each observation by the class that wins the majority votes from the multiple predictions, made on the bootstrap samples in which the particular observation is out of bag (i.e. not included).

The OOB estimation makes an interesting comparison to the leave-one-out bootstrap. The OOB estimation employs a majority vote on the multiple predictions made for observation i based on the bootstrap samples in C_i, while the leave-one-out bootstrap takes an average on errors of these predictions. The OOB estimation can be viewed as a non-smooth variant and we envisage it to have larger variability than the leave-one-out bootstrap when the sample size is small.

0.632+ *Bootstrap*. The 0.632+ bootstrap is proposed by Efron and Tibshirani [3] in order to reduce the upward bias of the leave-one-out bootstrap. The estimate has the form $\hat{e}_n^{0.632+} = w\hat{e}_n^{\text{LOOBS}} + (1-w)\hat{e}_n^{\text{RSB}}$, where the weight w is between 0 and 1 and $\hat{e}_n^{\text{RSB}}$ is the resubstitution estimate. Taking $w = 0.632$ gives the *0.632 bootstrap* originally proposed by Efron [2]. When the resubstitution error is zero, the 0.632 bootstrap estimate becomes $0.632\hat{e}_n^{\text{LOOBS}}$. This results in systematic downward bias when there are no class differences [3, 13]. The 0.632+ bootstrap aims to circumvent this problem by increasing the weight w with respect to the growing level of overfitting. It often performs well in classification problems with $n > p$. For microarray data with $n < p$, the overfitting problem always exists and the resubstitution error estimate is often close to zero. The 0.632+ bootstrap tends to put too much weight on the leave-one-out bootstrap estimate in this situation.

3. AN ADJUSTED BOOTSTRAP APPROACH

As discussed in the previous section, all reviewed methods for estimating prediction error encounter problems (large downward or upward bias, or large variability) for small samples. In this section,

DOI: 10.1002/sim

we initially construct a repeated leave-one-out bootstrap (RLOOB) that generates bootstrap learning sets of size ln. The resulting estimates exhibit a decreasing pattern toward the true prediction error as l increases. We then propose an adjusted bootstrap (ABS) method which fits a learning curve on these estimates in order to improve on the accuracy of the estimation.

The RLOOB method is described as follows. For every original sample x, leave out one observation at a time and denote the resulting sets by $x_{(-1)}, \ldots, x_{(-n)}$. From each leave-one-out set $x_{(-i)}$, draw B_1 bootstrap learning sets of size ln. Build a prediction rule on every bootstrap learning set generated from $x_{(-i)}$ and apply the rule on the test observation x_i. The RLOOB estimate is the misclassification rate calculated across all the bootstrap runs and all n observations. It can be expressed as

$$\hat{e}_n^{\text{RLOOB}}(l) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B_1}\sum_{b_i=1}^{B_1}\{y_i \neq r(t_i, x_{(-i)}^{*,b_i})\}$$

where $x_{(-i)}^{*,b_i}$ is the b_ith bootstrap learning set of size ln drawn from the set $x_{(-i)}$. Feature selection should be carried out on every bootstrap learning set $x_{(-i)}^{*,b_i}$ for $b_i = 1, \ldots, B_1$ and $i = 1, \ldots, n$.

Let $c(l)$ be the chances that an observation appears in a bootstrap sample of size ln. A simple probabilistic argument indicates that $c(l) \approx 1 - \mathrm{e}^{-l}$. A bootstrap sample of size ln contains approximately $c(l) \cdot n$ distinct observations from the original sample. For example, for $l = 1, 2, 3$, the number of distinct observations is about $0.632n$, $0.865n$, $0.95n$, respectively. With $l = 1$, the RLOOB closely resembles the leave-one-out bootstrap procedure. As l increases, a bootstrap learning set for a left-out item contains more distinct observations. On one hand, the method acquires additional accuracy and brings a reduction on the upward bias. On the other hand, the bootstrap learning sets obtained from the same leave-one-out set become more similar in structure and this raises the variability of the estimation.

The learning behavior of the RLOOB can be modeled as a function of the number of distinct observations included in the bootstrap learning sets. The trend of a learning process as a function of sample size is often modeled in the machine learning literature by a flexible curve following an inverse power law. Let m be the expected number of distinct observations to appear in a bootstrap sample of size ln. Let $e(m)$ be the expected error rate given the observed sample using the repeated leave-one-out bootstrap method with bootstrap learning sets of size ln. Ideally, $e(m)$ should follow the inverse power law

$$e(m) = am^{-\alpha} + b$$

where a, α and b are the parameters.

The ABS method estimates the prediction error as follows. Pick J bootstrap learning set sizes $l_j n$, $j = 1, \ldots, J$. Compute the RLOOB estimate $\hat{e}_n^{\text{RLOOB}}(l_j)$ with bootstrap learning sets of size $l_j n$. Denote $\hat{e}_n^{\text{RLOOB}}(l_j)$ by e_{m_j}, where $m_j = c(l_j) \cdot n$ is the expected number of distinct original observations in a bootstrap learning set of size $l_j n$. Fit an empirical learning curve of the form $e_{m_j} = am_j^{-\alpha} + b$ with $j = 1, \ldots, J$. The estimates $\hat{a}$, $\hat{\alpha}$ and $\hat{b}$ for the parameters are obtained by minimizing the non-linear least-squares function $\sum_{j=1}^{J}\{e_{m_j} - am_j^{-\alpha} - b\}^2$.

The ABS estimate for the prediction error is given by

$$\hat{e}_n^{\text{ABS}} = \hat{a}n^{-\hat{\alpha}} + \hat{b}$$

It is the fitted value on the learning curve assuming that all original observations contributed to an individual bootstrap learning set.

In practice, the choice of l can range from somewhere close to 1 to a value greater than 5. RLOOB estimates typically have lower variability than leave-one-out cross-validation. They have an upward bias that decreases and their variability increases with the expected number of distinct original observations selected in bootstrap learning sets. Fitting an inverse power law curve to a series of RLOOB values enables us to define a conservative estimate (not subject to downward bias) that provides a compromise between estimates with large variability and large upward bias. The inverse power law curve is a flexible way to model a learning process, and is quite typical in describing machine learning, human and animal learning behavior [14]. Mukherjee *et al.* [15] studied sample size requirements in microarray classification using a similar learning curve.

4. COMPARISON OF METHODS

In this section, we compare the methods described in Sections 2 and 3 through simulation study and an application to a lymphoma data set.

4.1. Simulated data

Similar simulated data sets are considered in Molinaro *et al.* [8]. For each simulated data set, generate a sample of n patients, each with p genes (or features). Half of the patients are in class 0 and half in class 1, divided according to their disease status. Gene expression levels are generated from a normal distribution with covariance matrix $\Sigma=(\sigma_{ij}), i, j=1,\ldots,p$, where the only non-zero entries are $\sigma_{ii}=1$ and $\sigma_{ij}=0.2$ with $0<|i-j|\leqslant 5$. For class 0 patients, genes are generated with mean 0. For class 1 patients, 1 per cent of the genes are generated with mean μ_1, 1 per cent with mean μ_2 and the rest with mean 0.

In each simulation run, a prediction model is built on the sample of size n and tested on 1000 independent data generated with the same structure. The resulting error rate estimates the true prediction error for the sample and is denoted by $\tilde{e}_n$. For each method, we report the averaged estimate (Est.) and the standard deviation (STD) calculated across $R=1000$ simulation replications, as well as the averaged bias (Bias) and mean-squared error (MSE) with respect to the 'true' prediction error $\tilde{e}_n$. We have estimated the variability (STD) of a method of estimating prediction error using simulation, but such a variability estimate would not be available for analysis of a single data set.

For class discrimination, we consider in the simulation the diagonal linear discriminant analysis (DLDA), the one nearest neighbor with Euclidean distance (1NN) and the classification and regression tree (CART). These algorithms are available through built-in functions in the statistical package R [16]. Details of these R functions are described in Molinaro *et al.* [8]. For all methods reviewed in Section 2, we draw $B=100$ bootstrap samples. Running 50 to 100 bootstrap replications is often considered more than adequate [3, 4]. For the RLOOB and ABS, we run $B_1=50$ bootstrap replications on every leave-one-out set and use them to predict for the left-out observation. This should be quite sufficient in comparison to the leave-one-out bootstrap method, in which B bootstrap replications provide about $0.368B$ bootstrap samples not containing (and to predict for) a specific observation. The ABS method is fitted on $J=6$ RLOOB estimates with $l=0.75, 1, 1.5, 2, 3, 10$, so that the number of original observations contributing to the resampled learning set spreads out across a reasonable range for small-sample problems.

DOI: 10.1002/sim

Table I. Simulation study with samples of size 20 and 800 genes.

	Prediction error	Case 1: no differential genes				Case 2: 2 % differential genes			
Classifier	estimation method	Est.	STD	Bias	MSE	Est.	STD	Bias	MSE
DLDA	‘True’ error ($\tilde{e}_n$)	0.500	0.016			0.184	0.067		
	Resubstitution	0.009	0.020	−**0.491**	0.242	0.006	0.017	−**0.177**	0.036
	Ordinary bootstrap	0.196	0.022	−**0.304**	0.093	0.130	0.036	−**0.054**	0.006
	Bootstrap cross-validation	0.205	0.024	−**0.295**	0.088	0.139	0.037	−**0.045**	0.006
	0.632 Bootstrap	0.344	0.039	−**0.157**	0.026	0.229	0.064	0.045	0.007
	LOOCV	0.527	**0.206**	0.026	0.043	0.206	**0.152**	0.022	0.019
	Out-of-bag estimation	0.590	**0.156**	0.090	0.033	0.243	**0.153**	0.059	0.022
	Leave-one-out bootstrap	0.538	0.059	0.038	0.005	0.359	0.098	**0.175**	0.038
	0.632+ Bootstrap	0.516	0.054	0.015	0.003	0.318	0.111	**0.134**	0.027
	RLOOB:l = 1	0.539	0.058	0.039	0.005	0.358	0.098	0.175	0.038
	RLOOB:l = 2	0.537	0.098	0.036	0.011	0.278	0.121	0.095	0.020
	RLOOB:l = 10	0.532	0.160	0.032	0.027	0.217	0.136	0.034	0.015
	Adjusted bootstrap	0.534	0.128	0.033	0.018	0.237	0.133	0.053	0.016
1NN	‘True’ error ($\tilde{e}_n$)	0.501	0.016			0.211	0.071		
	Resubstitution	0	0	−**0.501**	0.251	0	0	−**0.211**	0.050
	Ordinary bootstrap	0.194	0.021	−**0.306**	0.094	0.127	0.038	−**0.085**	0.010
	Bootstrap cross-validation	0.205	0.022	−**0.296**	0.088	0.136	0.040	−**0.075**	0.009
	0.632 Bootstrap	0.342	0.036	−**0.159**	0.027	0.223	0.067	0.012	0.004
	LOOCV	0.529	**0.184**	0.029	0.035	0.241	**0.166**	0.030	0.020
	Out-of-bag estimation	0.600	**0.161**	0.100	0.036	0.243	**0.161**	0.031	0.018
	Leave-one-out bootstrap	0.541	0.058	0.040	0.005	0.354	0.106	**0.142**	0.027
	0.632+ Bootstrap	0.518	0.051	0.018	0.003	0.312	0.118	**0.100**	0.018
	RLOOB:l = 1	0.541	0.055	0.041	0.005	0.354	0.105	0.143	0.027
	RLOOB:l = 2	0.536	0.090	0.036	0.010	0.289	0.127	0.077	0.015
	RLOOB:l = 10	0.532	0.140	0.032	0.021	0.247	0.144	0.036	0.014
	Adjusted bootstrap	0.533	0.114	0.032	0.014	0.258	0.139	0.046	0.014
CART	‘True’ error ($\tilde{e}_n$)	0.500	0.016			0.290	0.089		
	Resubstitution	0	0	−**0.500**	0.250	0	0	−**0.290**	0.092
	Ordinary bootstrap	0.188	0.020	−**0.312**	0.098	0.129	0.041	−**0.162**	0.033
	Bootstrap cross-validation	0.197	0.022	−**0.303**	0.092	0.135	0.042	−**0.156**	0.031
	0.632 Bootstrap	0.331	0.034	−**0.169**	0.030	0.227	0.073	−**0.064**	0.013
	LOOCV	0.528	**0.225**	0.029	0.516	0.326	**0.194**	0.035	0.038
	Out-of-bag estimation	0.574	**0.180**	0.074	0.038	0.253	0.150	−**0.038**	0.022
	Leave-one-out bootstrap	0.524	0.054	0.024	0.004	0.359	0.115	**0.068**	0.018
	0.632+ Bootstrap	0.505	0.056	0.005	0.003	0.320	0.126	0.030	0.016
	RLOOB:l = 1	0.523	0.053	0.023	0.004	0.358	0.114	0.067	0.018
	RLOOB:l = 2	0.521	0.099	0.021	0.010	0.350	0.136	0.060	0.023
	RLOOB:l = 10	0.520	0.173	0.020	0.030	0.331	0.170	0.041	0.032
	Adjusted bootstrap	0.520	0.135	0.020	0.019	0.343	0.155	0.053	0.028

DLDA, diagonal linear discriminant analysis; 1NN, one nearest neighbor; CART, classification and regression tree; RLOOB, repeated leave-one-out bootstrap.

Two simulation setups are considered. Case 1: no genes are differentially expressed between the two classes. Case 2: gene expression levels of class 0 patients follow normal distribution with mean 0; for class 1 patients, the 2 per cent differentially expressed genes follow normal mixtures, half with mean $\mu_1 = 0.5$ and half with mean $\mu_2 = 1.5$. The ‘true’ prediction error $\tilde{e}_n$ is the misclassification rate when a prediction rule built on the sample is tested on 1000 independent data with the same structure.

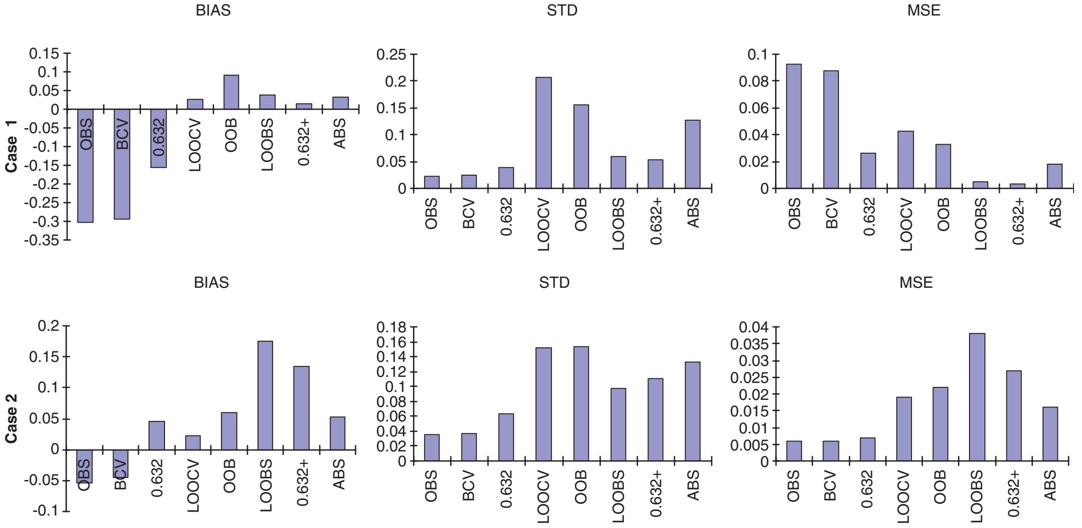

Figure 1. Comparison of prediction error estimation on simulated data sets with $n=20$, $p=800$. Case 1: no differentially expressed genes; Case 2: class 0 patients follow normal distribution with mean 0, for class 1 patients, the 2 per cent differentially expressed genes follow normal mixtures, half with mean 0.5 and half with mean 1.5. Diagonal linear discriminant analysis is used in class prediction. The 'true' prediction errors for the two cases are 0.500 and 0.184. Methods displayed are ordinary bootstrap (OBS), bootstrap cross-validation (BCV), 0.632 bootstrap, leave-one-out cross-validation (LOOCV), out-of-bag estimation (OOB), leave-one-out bootstrap (LOOBS), 0.632+ bootstrap, and adjusted bootstrap (ABS).

In Table I, we report the simulation results for two cases with $n=20$. In Case 1, we consider the no signal situation, where there are no differentially expressed genes between the two classes ($\mu_1=\mu_2=0$ in the simulation model). In Case 2, we consider a situation with moderate to strong signals ($\mu_1=0.5$, $\mu_2=1.5$ in the simulation model). We reported the methods in groups with substantial downward bias, large variability and large upward bias, and in all tables, cells with these unfavorable features are highlighted in boldface. Figure 1 clearly displays the comparative performance of the methods for the two contrasting cases with $n=20$ using DLDA.

We first look at the outcome when the DLDA and/or 1NN classifiers are used, in Table I and Figure 1. Results using these two classifiers are very similar. The resubstitution estimates are close to zero in the study. The ordinary bootstrap underestimates the prediction errors and the problem is more serious when there are weak or no signals distinguishing the classes (Case 1). The behavior of the bootstrap cross-validation method is very similar to that of the ordinary bootstrap across all situations. With moderate to strong signals, the 0.632 bootstrap performs well in terms of bias, STD and MSE. But the 0.632 bootstrap suffers from a systematic downward bias when there is no signal (Case 1). The LOOCV estimate is almost unbiased, but its STD becomes very large as the signal diminishes. The bootstrap-related methods reviewed in Section 2 generally have small variability. The only exception is the OOB estimate. It often gives STDs as large as the LOOCV and is much more unstable than the leave-one-out bootstrap. With strong signals and small sample sizes, the leave-one-out bootstrap has very large upward bias; the 0.632+ bootstrap reduces the bias of the leave-one-out bootstrap to some extent but still seriously overestimates the truth. In Case 2 with DLDA, for example, the bias-to-'true'-error ratios are more than 90 and 70 per cent

DOI: 10.1002/sim

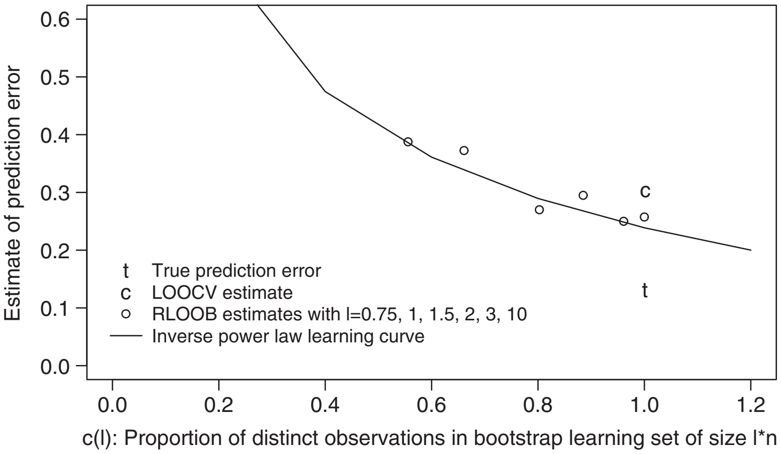

Figure 2. Illustration of the adjusted bootstrap method.

for leave-one-out bootstrap and the 0.632+ bootstrap with DLDA classifier. In situations with no signals, however, the leave-one-out bootstrap and the 0.632+ bootstrap both work very well. The ABS estimate evidently reduces the bias of the leave-one-out bootstrap estimate and variability of the LOOCV.

We also present in Table I the RLOOB estimates with $l=1$, 2 and 10. With small l, the RLOOB estimate gives small STD but can have a large upward bias. Increasing l in the method tends to reduce the upward bias but raise the STD. This along with Figure 2 illustrates the rationale for the ABS approach. In Figure 2, we plot the RLOOB estimates (using DLDA classifier) with $l=0.75, 1, 1.5, 2, 3, 10$ against the corresponding values of $c(l)$ for one of the simulated data set with $\mu_1=0.5$, $\mu_2=1.5$ and $n=20$. The quantity $c(l)$ shows roughly the proportion of the distinct observations from the original sample appearing in a bootstrap learning set of size ln. Also presented in Figure 2 is an inverse power law learning curve fitted through the ABS approach. The 'true' error rate and the LOOCV estimate are indicated by 't' and 'c' on the plot. When implementing the ABS method, we use the R function 'nlm', a non-linear minimization algorithm, to estimate for the parameters.

In Table I, we notice that using CART in the methods gives larger prediction error rates than using the other two classifiers. The CART classifier constantly overfits the data in the presence of large amount of noise. Not surprisingly, the resampling methods using CART become less sensitive to the varying sizes of the learning sets. Thus, the leave-one-out bootstrap has a smaller upward bias with CART than with the other two classifiers. The 0.632+ bootstrap performs well with CART when $n=20$. The RLOOB estimates using CART change only mildly as the size of the bootstrap learning set ln increases, and this reduces the benefit of the ABS method. But it remains slightly conservative and still reduces the large variability of the LOOCV and performs reasonably well under varying signal levels.

To see how the relative performance of the methods changes with sample sizes along with varying signal levels, in Tables II and III, we report the results for $n=40$ and 100. In each table, Case 1 is the no signal situation, and Case 2 is a situation with strong signals ($\mu=1.5$, i.e. $\mu_1=\mu_2$ in the simulation model). With these larger sample sizes (even with $n=100$), we find that when there are no real differences between the two classes, the OOB estimation and the LOOCV still

Table II. Simulation study with samples of size 40 and 800 genes.

Classifier	Prediction error estimation method	Case 1: no differential genes				Case 2: 2 % differential genes			
		Est.	STD	Bias	MSE	Est.	STD	Bias	MSE
DLDA	'True' error ($\tilde{e}_n$)	0.501	0.016			0.052	0.010		
	Resubstitution	0.066	0.035	**−0.435**	0.191	0.021	0.022	**−0.031**	0.0016
	Ordinary bootstrap	0.224	0.020	**−0.277**	0.077	0.040	0.020	**−0.012**	0.0007
	Bootstrap cross-validation	0.226	0.021	**−0.274**	0.076	0.042	0.021	**−0.010**	0.0007
	0.632 Bootstrap	0.351	0.034	**−0.150**	0.024	0.062	0.033	0.010	0.0013
	LOOCV	0.513	**0.160**	0.012	0.026	0.053	0.038	0.001	0.0015
	Out-of-bag estimation	0.539	**0.113**	0.038	0.014	0.049	0.036	**−0.003**	0.0014
	Leave-one-out bootstrap	0.517	0.043	0.016	0.002	0.087	0.043	**0.035**	0.0032
	0.632+ Bootstrap	0.502	0.043	0.001	0.002	0.065	0.035	0.013	0.0015
	Adjusted bootstrap	0.514	0.085	0.013	0.008	0.056	0.033	0.004	0.0012
1NN	'True' error ($\tilde{e}_n$)	0.501	0.016			0.079	0.020		
	Resubstitution	0	0	**−0.501**	0.251	0	0	**−0.079**	0.0066
	Ordinary bootstrap	0.188	0.013	**−0.312**	0.098	0.040	0.018	**−0.038**	0.0019
	Bootstrap cross-validation	0.193	0.014	**−0.308**	0.095	0.042	0.019	**−0.037**	0.0018
	0.632 Bootstrap	0.328	0.022	**−0.173**	0.031	0.070	0.032	**−0.008**	0.0009
	LOOCV	0.520	**0.132**	0.019	0.018	0.079	0.055	0.000	0.0023
	Out-of-bag estimation	0.550	**0.110**	0.049	0.015	0.058	0.042	**−0.021**	0.0019
	Leave-one-out bootstrap	0.519	0.036	0.018	0.002	0.111	0.050	**0.033**	0.0031
	0.632+ Bootstrap	0.505	0.035	0.005	0.001	0.078	0.038	−0.000	0.0012
	Adjusted bootstrap	0.517	0.064	0.016	0.005	0.085	0.047	0.006	0.0017
CART	'True' error ($\tilde{e}_n$)	0.500	0.016			0.221	0.042		
	Resubstitution	0.028	0.019	**−0.471**	0.223	0.020	0.017	**−0.201**	0.042
	Ordinary bootstrap	0.188	0.011	**−0.312**	0.097	0.085	0.019	**−0.136**	0.020
	Bootstrap cross-validation	0.191	0.013	**−0.309**	0.096	0.085	0.019	**−0.136**	0.020
	0.632 Bootstrap	0.335	0.019	**−0.165**	0.028	0.153	0.035	**−0.069**	0.007
	LOOCV	0.515	**0.175**	0.015	0.031	0.212	**0.112**	−0.010	0.012
	Out-of-bag estimation	0.543	**0.113**	0.043	0.015	0.123	0.049	**−0.099**	0.013
	Leave-one-out bootstrap	0.513	0.029	0.013	0.001	0.230	0.052	0.009	0.004
	0.632+ Bootstrap	0.503	0.029	0.003	0.001	0.180	0.047	**−0.041**	0.005
	Adjusted bootstrap	0.513	0.069	0.014	0.005	0.225	0.078	0.004	0.006

DLDA, diagonal linear discriminant analysis; 1NN, one nearest neighbor; CART, classification and regression tree.

Two simulation setups are considered. Case 1: no genes are differentially expressed between the two classes. Case 2: gene expression levels of class 0 patients follow normal distribution with mean 0; for class 1 patients, the 2 per cent differentially expressed genes follow normal distribution with mean $\mu = 1.5$. The 'true' prediction error $\tilde{e}_n$ is the misclassification rate when a prediction rule built on the sample is tested on 1000 independent data with the same structure.

give larger variability compared with other methods; the resubstitution, the ordinary bootstrap, the bootstrap cross-validation and the 0.632 bootstrap still have substantial downward bias (Case 1, Tables II and III). When there are strong differences between the classes, the LOOCV and OOB give smaller variability, and the resubstitution, the ordinary bootstrap, the bootstrap cross-validation and the 0.632 bootstrap give smaller downward bias (Case 2, Tables II and III) as sample size n increases in comparison to Case 2, Table I. The 0.632+ bootstrap and the OOB perform better as sample size increases, but they sometimes suffer from downward bias when $n = 40$ and 100

DOI: 10.1002/sim

Table III. Simulation study with samples of size 100 and 800 genes.

Classifier	Prediction error estimation method	Case 1: no differential genes				Case 2: 2 % differential genes			
		Est.	STD	Bias	MSE	Est.	STD	Bias	MSE
DLDA	'True' error ($\tilde{e}_n$)	0.500	0.016			0.049	0.008		
	Resubstitution	0.183	0.031	**−0.317**	0.102	0.031	0.018	**−0.018**	0.0007
	Ordinary bootstrap	0.291	0.015	**−0.209**	0.044	0.037	0.014	**−0.012**	0.0004
	Bootstrap cross-validation	0.292	0.016	**−0.208**	0.044	0.037	0.014	**−0.012**	0.0004
	0.632 Bootstrap	0.386	0.023	**−0.114**	0.014	0.043	0.016	**−0.006**	0.0003
	LOOCV	0.498	**0.110**	−0.002	0.012	0.047	0.023	−0.001	0.0006
	Out-of-bag estimation	0.510	**0.069**	0.010	0.005	0.040	0.019	**−0.009**	0.0005
	Leave-one-out bootstrap	0.504	0.026	0.004	0.001	0.050	0.017	0.001	0.0003
	0.632+ Bootstrap	0.496	0.028	−0.005	0.001	0.043	0.016	**−0.006**	0.0003
	Adjusted bootstrap	0.503	0.048	0.003	0.003	0.049	0.018	0.000	0.0004
1NN	'True' error ($\tilde{e}_n$)	0.500	0.015			0.076	0.015		
	Resubstitution	0	0	**−0.500**	0.251	0	0	**−0.076**	0.0059
	Ordinary bootstrap	0.185	0.007	**−0.315**	0.100	0.029	0.009	**−0.047**	0.0024
	Bootstrap cross-validation	0.187	0.008	**−0.314**	0.099	0.029	0.009	**−0.047**	0.0024
	0.632 Bootstrap	0.320	0.012	**−0.181**	0.033	0.049	0.015	**−0.026**	0.0009
	LOOCV	0.504	**0.081**	0.003	0.007	0.076	0.034	−0.000	0.0009
	Out-of-bag estimation	0.519	**0.068**	0.019	0.005	0.050	0.024	**−0.026**	0.0012
	Leave-one-out bootstrap	0.506	0.019	0.005	0.001	0.078	0.024	0.002	0.0005
	0.632+ Bootstrap	0.499	0.020	−0.001	0.001	0.053	0.018	**−0.023**	0.0008
	Adjusted bootstrap	0.505	0.031	0.005	0.001	0.077	0.027	0.001	0.0006
CART	'True' error ($\tilde{e}_n$)	0.500	0.016			0.188	0.025		
	Resubstitution	0.062	0.021	**−0.438**	0.193	0.041	0.017	**−0.146**	0.022
	Ordinary bootstrap	0.220	0.006	**−0.280**	0.079	0.100	0.013	**−0.088**	0.009
	Bootstrap cross-validation	0.222	0.008	**−0.278**	0.078	0.101	0.013	**−0.087**	0.008
	0.632 Bootstrap	0.342	0.012	**−0.157**	0.025	0.151	0.021	**−0.037**	0.002
	LOOCV	0.503	**0.124**	0.003	0.015	0.189	**0.076**	0.002	0.006
	Out-of-bag estimation	0.524	**0.066**	0.024	0.005	0.100	0.028	**−0.088**	0.009
	Leave-one-out bootstrap	0.506	0.015	0.006	0.001	0.214	0.029	0.026	0.002
	0.632+ Bootstrap	0.500	0.016	0.001	0.000	0.169	0.026	**−0.019**	0.002
	Adjusted bootstrap	0.506	0.033	0.006	0.001	0.200	0.039	0.012	0.002

DLDA, diagonal linear discriminant analysis; 1NN, one nearest neighbor; CART, classification and regression tree.

Two simulation setups are considered. Case 1: no genes are differentially expressed between the two classes. Case 2: gene expression levels of class 0 patients follow normal distribution with mean 0; for class 1 patients, the 2 per cent differentially expressed genes follow normal distribution with mean $\mu = 1.5$. The 'true' prediction error $\tilde{e}_n$ is the misclassification rate when a prediction rule built on the sample is tested on 1000 independent data with the same structure.

(Case 2, Tables II and III), and this is illustrated in Figure 3 using the results for $n = 40$ with CART classifier.

The ABS is robust in the sense that it remains conservative (has no downward bias) under all circumstances considered in the simulation with varying signal levels, classifiers and sample sizes. It does not suffer from extremely large upward bias or variability in comparison to other methods for small-to-moderate-sized samples (Tables I and II), and it performs no worse than the competitors such as the 0.632+ bootstrap, the OOB and the LOOCV for larger sample sizes (Table III).

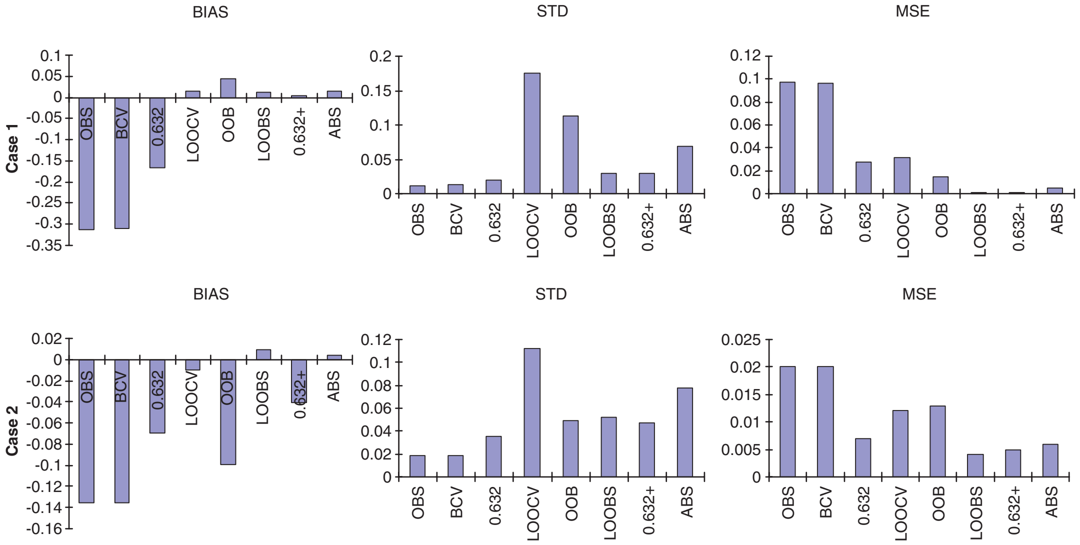

Figure 3. Comparison of prediction error estimation on simulated data sets with $n=40$, $p=800$. Case 1: no differentially expressed genes; Case 2: class 0 patients follow normal distribution with mean 0, for class 1 patients, the 2 per cent differentially expressed genes follow normal with mean 1.5. Classification and regression tree is used in class prediction. The 'true' prediction errors for the two cases are 0.500 and 0.221. Methods displayed are ordinary bootstrap (OBS), bootstrap cross-validation (BCV), 0.632 bootstrap, leave-one-out cross-validation (LOOCV), out-of-bag estimation (OOB), leave-one-out bootstrap (LOOBS), 0.632+ bootstrap, and adjusted bootstrap (ABS).

Additional simulations are conducted for varying signals and n/p ratios (Tables A1 and A2 in supplement). We found that the comparative conclusion does not depend on the n/p ratios (with $n \ll p$). When the DLDA or 1NN classifier is used, procedures for Tables I–III pick the 10 genes having the largest absolute value t-statistics in the feature selection steps. Choosing 30 genes instead in the feature selection steps does not affect the overall comparison of the methods as shown in the supplement (Tables A3 and A4). The CART algorithm selects features intrinsically and the number of variables selected is not fixed in advance. These should have covered a reasonable range of class prediction algorithms/feature selection methods/number of variables selected. Tables A1–A4 are reported in the supplementary material for this paper available at http://linus.nci.nih.gov/~brb/TechReport.htm.

4.2. Lymphoma data

Rosenwald *et al.* [17] conducted a microarray study among patients with large B-cell lymphoma and identified the germinal-center B-cell-like subgroup which had the highest five-year survival rate after chemotherapy. The study measured 7399 genes on 240 patients. In the following analysis, we define the classes of outcome by the lymphoma subgroups, the germinal-center B-cell-like as class 1, the activated B-cell-like and type 3 as class 0. These expression data provide only moderate signals to distinguish between the classes [18]. To assess the performance of the methods described in Sections 2 and 3, we repeatedly draw a sample of size n from the 240 patients, find the estimates

Table IV. Lymphoma data using a diagonal linear discriminant analysis (DLDA) classifier.

Prediction error estimation method	$n = 14$				$n = 20$			
	Est.	STD	Bias	MSE	Est.	STD	Bias	MSE
'True' error ($\tilde{e}_n$)	0.257	0.071			0.211	0.055		
Resubstitution	0.002	0.024	**−0.256**	0.070	0.010	0.023	**−0.201**	0.044
Ordinary bootstrap	0.184	0.057	**−0.073**	0.013	0.121	0.032	**−0.090**	0.012
Bootstrap cross-validation	0.209	0.058	**−0.049**	0.010	0.130	0.033	**−0.081**	0.010
0.632 Bootstrap	0.267	0.054	0.010	0.008	0.210	0.055	−0.001	0.005
LOOCV	0.306	**0.184**	0.048	0.039	0.234	**0.135**	0.023	0.019
Out-of-bag estimation	0.358	**0.156**	**0.100**	0.039	0.243	0.119	0.032	0.016
Leave-one-out bootstrap	0.422	0.084	**0.164**	0.039	0.327	0.084	**0.116**	0.022
0.632+ Bootstrap	0.390	0.100	**0.132**	0.032	0.279	0.092	**0.069**	0.015
Adjusted bootstrap	0.309	0.155	0.052	0.030	0.247	0.111	0.036	0.015

The 'true' prediction error $\tilde{e}_n$ is the misclassification rate when a prediction rule built on the sample is tested on those $240 - n$ patients not selected in the sample.

for prediction error and use the remaining patients as an independent test set to calculate the 'true' prediction error. This procedure is repeated $R = 1000$ times. The number of bootstrap replications is $B = 50$ for the methods reviewed in Section 2 and $B_1 = 20$ for the RLOOB.

We present the results for $n = 14$ and 20 in Table IV and apply only the DLDA classifier in this study. Feature selection is performed prior to each application of DLDA by choosing the 10 genes having the largest absolute-value t-statistics. In this moderate signal example, the methods behave in conformity with the trend observed in the previous simulation study. The ordinary bootstrap and the bootstrap cross-validation are less competitive because they underestimate the true prediction error; the OOB estimation is a less stable variant of the leave-one-out bootstrap, while the 0.632 bootstrap performs well for both sample sizes. The LOOCV has the largest variability. The leave-one-out bootstrap and the 0.632+ bootstrap overestimate the truth by about 64 and 51 per cent when $n = 14$ and by 55 and 33 per cent when $n = 20$. With the ABS, the ratios of overestimation drop to 20 and 14 per cent for $n = 14$ and 20.

5. DISCUSSION

As CART is not a satisfactory classifier for high-dimensional microarray data [12], we only included it in the simulation only to evaluate how the classifiers interact with the procedures for prediction error estimation. In the discussion, we mainly focus on methods using the two simpler and better-behaved classifiers, DLDA and 1NN [12].

There is a bias–variance trade-off in the behavior of prediction error estimates and no method is universally better than others with both bias and variability considerations. For example, the LOOCV and the leave-one-out bootstrap procedures include $n - 1$ and roughly $0.632n$ distinct observations, respectively, in each learning set. As a consequence, the LOOCV is almost unbiased but can have large variability; the leave-one-out bootstrap can seriously overestimate the true prediction error but has small variability. The LOOCV provides satisfactory estimates in strong signal situations and the leave-one-out bootstrap in the no-signal situations, but no method is the overall champion under all circumstances.

DOI: 10.1002/sim

Overlaps between the resampled learning and test sets cause serious underestimation of the prediction error. Such overlaps occur, for instance, in the ordinary bootstrap procedure and the bootstrap cross-validation. Both estimates suffer from downward bias, which becomes quite substantial as the signal to discriminate the classes weakens. The simulation study on the bootstrap cross-validation method by Fu *et al*. [9] was limited to situations with very strong signals, and it overlooked the necessity of feature selection in the resampling for high-dimensional data. We examine the bootstrap cross-validation method in more extensive situations with proper feature selection. Overall, the bootstrap cross-validation estimate performs not much better than the ordinary bootstrap estimate.

All the methods reviewed in Section 2 encounter difficulties when estimating prediction errors for high-dimensional data with small samples. The LOOCV and the OOB estimation suffer from large variability when the signal becomes weak. The leave-one-out bootstrap results in substantial upward bias as the signal level ranges from moderate to strong. Efron and Tibshirani [3] showed through simulation that the 0.632+ bootstrap works well for sample sizes as small as $n=14$ and 20 in the traditional $n>p$ situation. However, the 0.632+ bootstrap is not as satisfactory in the $n<p$ situation when repeated feature selection is needed. The design of the 0.632+ bootstrap encounters difficulty since the resubstitution estimate is close to zero in the presence of overfitting when $n<p$. With the DLDA and 1NN classifiers, the 0.632+ bootstrap works well in no-signal situations but has a large upward bias when the signal is moderate to strong and the sample size is small. The ordinary bootstrap and the bootstrap cross-validation yield significant downward bias in the moderate to no signal situations although the bias becomes smaller as the signal level increases. The 0.632 bootstrap behaves well in moderate to strong signal situations but leads to a systematic downward bias when there are no differences between the classes, and the problem persists for any sample sizes.

With the large amount of noisy information and limited sample sizes in microarray studies, *it is often preferable to provide conservative estimates for the prediction error in order to avoid false-positive reports on the prediction models*. The ordinary bootstrap, the bootstrap cross-validation and the 0.632 bootstrap are thus considered less competitive because they provide anti-conservative estimates under some circumstances, even though they can work well in terms of MSEs in strong signal situations. The OOB estimation can occasionally underestimate the truth because of its high variability; even the 0.632+ bootstrap can be downwardly biased with the CART classifier.

Although the size of microarray studies for classification purposes is increasing nowadays, there are still a considerable number of studies with small-to-moderate-sized samples. For example, in the recent papers of Ghadimi *et al*. [19] and Cleater *et al*. [20], the sample sizes used in cancer class predictions were $n=22$ and 40, respectively. The review of Dupuy and Simon [21] of 90 studies of expression profiling with cancer outcome data found that 65 per cent of the publications had fewer than 50 patients. In this paper, we first reviewed and compared existing methods and found that their performances were not satisfactory in the context of microarray applications. This study reveals how various methods behave in microarray situations in terms of bias and variability, and the results help investigators and reviewers understand the limitation of classifiers and estimates of prediction error developed with small sample sizes. Second, we proposed the ABS method and demonstrated that it is more robust than other methods across varying signal levels and classifiers in this context. For small-to-moderate-sized samples, we suggest using the ABS method since: (1) it remains conservative, hence avoids overly optimistic assessment of a prediction model; (2) it does not suffer from extremely large bias or variability in comparison to other methods.

DOI: 10.1002/sim

These features of the ABS method are particularly appealing for small samples when other prediction error estimation methods encounter difficulties.

REFERENCES

1. Stone M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* 1974; **36**:111–147.
2. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 1983; **78**:316–331.
3. Efron B, Tibshirani R. Improvement on cross-validation: the 0.632+ bootstrap method. *Journal of the American Statistical Association* 1997; **92**:548–560.
4. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall: London, 1998.
5. Breiman L. Out-of-bag estimation. *Technical Report*, Department of Statistics, University of California, Berkeley, CA, 1996.
6. Breiman L. Bagging predictors. *Machine Learning* 1996; **24**:123–140.
7. Dudoit S, Fridlyand J. Classification in microarray experiments. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall: London, 2003; 93–158.
8. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005; **21**:3301–3307.
9. Fu W, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* 2005; **21**:1979–1986.
10. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003; **95**:14–18.
11. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America* 2002; **99**:6562–6566.
12. Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 2002; **97**:77–87.
13. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Wadsworth: Belmont, CA, 1984.
14. Shrager J, Hogg T, Huberman BA. A graph-dynamic model of the power law of practice and the problem-solving fan effect. *Science* 1988; **242**:414–416.
15. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, Golub TR, Mesirov JP. Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* 2003; **10**:119–142.
16. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; **5**:299–314.
17. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Staudt LM. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *The New England Journal of Medicine* 2002; **346**:1937–1947.
18. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America* 2003; **100**:9991–9996.
19. Ghadimi BM, Grade M, Difilippantonio MJ, Varma S, Simon R, Montagna C, Fuzesi L, Langer C, Becker H, Liersch T, Ried T. Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. *Journal of Clinical Oncology* 2005; **23**:1826–1838.
20. Cleator S, Tsimelzon A, Ashworth A, Dowsett M, Dexter T, Powles T, Hilsenbeck S, Wong H, Osborne CK, O'Connell P, Chang JC. Gene expression patterns for doxorubicin (Adriamycin) and cyclophosphamide (Cytoxan) (AC) response and resistance. *Breast Cancer Research and Treatment* 2006; **95**:229–233.
21. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute* 2007; **99**:147–157.

DOI: 10.1002/sim