Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 38473, 12 pages
doi:10.1155/2007/38473

Research Article

Decorrelation of the True and Estimated Classifier Errors in High-Dimensional Settings

Blaise Hanczar,[1,2] Jianping Hua,[3] and Edward R. Dougherty[1,3]

[1] *Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA*
[2] *Laboratoire d'Informatique Medicale et Bio-informatique (Lim&Bio), Universite Paris 13, 93017 Bobigny cedex, France*
[3] *Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA*

Received 14 May 2007; Revised 11 August 2007; Accepted 27 August 2007

Recommended by John Goutsias

The aim of many microarray experiments is to build discriminatory diagnosis and prognosis models. Given the huge number of features and the small number of examples, model validity which refers to the precision of error estimation is a critical issue. Previous studies have addressed this issue via the deviation distribution (estimated error minus true error), in particular, the deterioration of cross-validation precision in high-dimensional settings where feature selection is used to mitigate the peaking phenomenon (overfitting). Because classifier design is based upon random samples, both the true and estimated errors are sample-dependent random variables, and one would expect a loss of precision if the estimated and true errors are not well correlated, so that natural questions arise as to the degree of correlation and the manner in which lack of correlation impacts error estimation. We demonstrate the effect of correlation on error precision via a decomposition of the variance of the deviation distribution, observe that the correlation is often severely decreased in high-dimensional settings, and show that the effect of high dimensionality on error estimation tends to result more from its decorrelating effects than from its impact on the variance of the estimated error. We consider the correlation between the true and estimated errors under different experimental conditions using both synthetic and real data, several feature-selection methods, different classification rules, and three error estimators commonly used (leave-one-out cross-validation, k-fold cross-validation, and .632 bootstrap). Moreover, three scenarios are considered: (1) feature selection, (2) known-feature set, and (3) all features. Only the first is of practical interest; however, the other two are needed for comparison purposes. We will observe that the true and estimated errors tend to be much more correlated in the case of a known feature set than with either feature selection or using all features, with the better correlation between the latter two showing no general trend, but differing for different models.

1. INTRODUCTION

The validity of a classifier model, the designed classifier, and prediction error depends upon the relationship between the estimated and true errors of the classifier. Model validity is different from classifier goodness. A good classifier is one with small error, but this error is unknown when a classifier is designed and its error is estimated from sample data. In this case, its performance must be judged from the estimated error. Since the error estimate characterizes our understanding of the predicted classifier performance on future observations and since we do not know the true error, model validity relates to the design process as a whole. What is the relationship between the estimated and true errors resulting from applying the classification and error-estimation rules to the feature-label distribution when using samples of a given size? Since classifier design is based upon random samples, the classifier is a random function and both the true and estimated errors are random variables, depending on the sample. Hence, we are concerned with the estimation of one random variable, the true error, by another random variable, the estimated error. Naturally, we would like the true and estimated errors to be strongly correlated. In this paper, using a number of feature-label models, classification rules, feature selection procedures, and error-estimation methods, we demonstrate that when there is high dimensionality, meaning a large number of potential features and a small sample, one should not expect significant correlation between the true and estimated errors. This conclusion has serious ramifications in the domain of high-throughput genomic classification, such as gene expression or SNP classification. For instance, with gene-expression microarrays, the number of

potential features (gene expressions) is usually in the tens of thousands and the number of sample points (microarrays) is often under one hundred. The relationship between the two errors depends on the feature-label distribution, the classification rule, the error-estimation procedure, and the sample size. According to the usual design protocol, a sample S of a given size is drawn from a feature-label distribution, a classification rule is applied to the sample to design a classifier, and the classifier error is estimated from the sample data by an error-estimation procedure. Within this general protocol, there are two standard issues to address. First, should the sample be split into training and test data? Since our interest is in small samples, we only consider the case where the same data is used for training and testing. The second issue is whether the feature set for the classifier is known ahead of time or it has to be chosen by a feature-selection algorithm. Since we are interested in high dimensionality, our focus is on the case where there is feature selection; nonetheless, in order to accent the effect of the feature-selection paradigm on the correlation between the estimated and true errors, for comparison purposes, we will also consider the situation where the feature set is known beforehand.

Keeping in mind that the feature-selection algorithm is part of the classification rule, we have the model $\mathrm{M}(F, \Omega, \Lambda, \Xi, D, d, n)$, where F is the feature-label distribution, Ω is the feature selection part of the classification rule, Λ is the classifier construction part of the classification rule, Ξ is the error-estimation procedure, D is the total number of available features, d is the number of features to be used as variables for the designed classifier, and n is the sample size. As an example, F is composed of two class-conditional Gaussian distributions over some number D of variables, Λ is linear-discriminant analysis, Ω is t-test feature selection, Ξ is leave-one-out cross-validation, $d = 5$ features, and $n = 50$ data points. In this model, feature selection is accomplished without reference to the classifier construction. If instead we let Ω be sequential forward selection, then it is accomplished in conjunction with classifier construction, and is referred to as a wrapper method. We will denote the designed classifier by ψ_n, where we recognize that ψ_n is a random function depending on the random sample.

The correlation between the true and estimated errors relates to the joint distribution of the random vector $(\varepsilon_{\mathrm{tru}}, \varepsilon_{\mathrm{est}})$, whose component random variables are the true error, $\varepsilon_{\mathrm{tru}}$, and the estimated error, $\varepsilon_{\mathrm{est}}$, of the designed classifier. This distribution is a function of the model $\mathrm{M}(F, \Omega, \Lambda, \Xi, D, d, n)$. A realization of the random vector $(\varepsilon_{\mathrm{tru}}, \varepsilon_{\mathrm{est}})$ occurs each time a sample is drawn from the feature-label distribution and a classifier is designed from the sample. In effect, we are considering the linear regression model

$$\mu_{\varepsilon_{\mathrm{tru}}|\varepsilon_{\mathrm{est}}} = a\varepsilon_{\mathrm{est}} + b, \tag{1}$$

where $\mu_{\varepsilon_{\mathrm{tru}}|\varepsilon_{\mathrm{est}}}$ is the conditional mean of $\varepsilon_{\mathrm{tru}}$, given $\varepsilon_{\mathrm{est}}$. The least-squares estimate of the regression coefficient a is given by

$$\hat{a} = \frac{\hat{\sigma}_{\mathrm{tru}}}{\hat{\sigma}_{\mathrm{est}}}\hat{\rho}, \tag{2}$$

where $\hat{\sigma}_{\mathrm{tru}}$, $\hat{\sigma}_{\mathrm{est}}$, and $\hat{\rho}$ are the sample-based estimates of the standard deviation σ_{tru} of $\varepsilon_{\mathrm{tru}}$, the standard deviation σ_{est} of $\varepsilon_{\mathrm{est}}$, and the correlation coefficient ρ for $\varepsilon_{\mathrm{tru}}$ and $\varepsilon_{\mathrm{est}}$, respectively, where we assume that $\hat{\sigma}_{\mathrm{est}} \neq 0$. In our experiments, we will see that $\hat{a} < 1$. The closer $\hat{a}$ is to 1, the stronger the regression, the closer $\hat{\rho}$ is to 1, the better the regression. As will be seen in our experiments (see figure C1 on the companion website at gsp.tamu.edu/Publications/error_fs/), it needs not be the case that $\hat{\sigma}_{\mathrm{tru}}/\hat{\sigma}_{\mathrm{est}} \leq 1$. Here, one might think of a pathological case: the resubstitution estimate for nearest-neighbor classification is always 0.

We will observe that, with feature selection, $\hat{\rho}$ will typically be very small, so that $\hat{a} \approx 0$ and the regression line is close to being horizontal: there is negligible correlation and regression between the true and estimated errors. When the feature set is known, there will be greater correlation between the true and estimated errors, and $\hat{a}$, while not large, will be significantly greater than zero. In the case of feature selection, this is a strong limiting result and brings into question the efficacy of the classification methodology, in particular, as it pertains to microarray-based classification, which usually involves extremely large sets of potential features.

While our simulations will show that there tends to be much less correlation between the true and estimated errors when using feature selection than when there is a known feature set, we must be careful about attributing responsibility for lack of correlation. In the absence of being given a feature set, feature selection is employed to mitigate overfitting the data and avoid falling prey to the peaking phenomenon, which refers to increasing classifier error when using too many features [1–3]. Feature selection is necessary and the result is decorrelation of the true and estimated errors; however, does the feature-selection process cause the decreased correlation or does it result from having a large number of features to begin with? To address this issue, in the absence of being given a feature set, we will consider both feature selection and using the full set of given features for classification. While the latter approach is not realistic, the comparison will help reveal the effect of the feature-selection procedure itself. In all, we will consider three scenarios: (1) feature selection, (2) known feature set, and (3) all features, the first one being the one of practical interest. We will observe that the true and estimated errors tend to be much more correlated in the case of a known feature set than with either feature selection or using all features, with the better correlation between the latter two showing no general trend, but differing for different models.

This is not the first time that concerns have been raised regarding the microarray classification paradigm. These concerns go back to practically the outset of the expression-based classification using microarray data [4]. Of particular relevance to the present paper are problems relating to small-sample error estimation. A basic concern is the deleterious effect of cross-validation variance on error-estimation accuracy [5], and specific concern has been raised as to the even worse performance of cross-validation when there is feature selection [6, 7]. Whereas the preceding studies focus on the increased variance of the deviation distribution between the estimated and true errors, here we utilize regression and

a decomposition of that variance to show that it is the decorrelation of the estimated and true errors in the case of feature selection that is the root of the problem.

Whereas here we focus on correlation and regression between the true and estimated errors, we note that various problems with error estimation and feature selection have been addressed in the context of high dimensionality and small samples. These include the effect of error estimation on gene ranking [8, 9], the effect of error estimation on feature selection [10], the effect of error estimation on cross-validation error estimation [6, 7], the impact of ties resulting from counting-based error estimators on feature selection algorithms [11], and the overall ability of feature selection to find good features sets [12]. With papers addressing single issues relating to error estimation and feature selection in small-sample settings, there have been a number of papers critiquing general statistical and methodological problems [13–19].

2. ERROR ESTIMATION

A classification task consists of predicting the value of a label Y from a feature vector $\mathbf{X} = (X_1, \ldots, X_D)$. Consider a two-class problem with a D-dimensional input space defined by the feature-label distribution F. A classifier is a function $\psi : R^D \to \{0, 1\}$ and its true-error rate is given by the expectation $\varepsilon[\psi] = E[|Y - \psi(\mathbf{X})|]$, taken relative to F. In practice, F is unknown and a classifier ψ_n is built, via a classification rule from a training sample S_n containing n examples drawn from F. The training sample is set of n independent pairs (feature vector, label), $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$. Assuming there is no feature selection, relative to the model $\mathrm{M}(F, \Lambda, \Xi, D, n)$, the true error of ψ_n is given by

$$\varepsilon_{\mathrm{tru}} = \varepsilon[\psi_n] = \varepsilon[\Lambda(S_n)] = E[|Y - \Lambda(S_n)(\mathbf{X})|]. \tag{3}$$

With feature selection, the model is of the form $\mathrm{M}(F, \Lambda, \Omega, \Xi, D, d, n)$ and (with feature selection being part of the classification rule), the true error takes the form

$$\varepsilon_{\mathrm{tru}} = \varepsilon[\psi_n] = \varepsilon[(\Lambda, \Omega)(S_n)] = E(|Y - (\Lambda, \Omega)(S_n)(\mathbf{X})|). \tag{4}$$

Computing the true error requires the feature-label distribution F. Since F is not available in practice, we compute only an estimate of the error. For small samples, this estimate must be done on the training data. Among the popular estimation rules are leave-one-out cross-validation, k-fold cross-validation, and bootstrap.

Cross-validation estimation is based on an iterative algorithm that partitions the training sample into k example subsets, S^i. At each iteration i, the ith subset is left out of classifier construction and used as a testing subset. The final k-fold cross-validation estimate is the mean of the errors obtained on all of the testing subsets:

$$\varepsilon_{\mathrm{cv}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n/k} \left| Y_j^i - (\Lambda, \Omega)(S_n - S^i)(\mathbf{X}_j^i) \right|, \tag{5}$$

where $(\mathbf{X}_j^i, Y_j^i)$ is an example in the ith subset. Cross-validation, although typically not too biased, suffers from high variance when sample sizes are small. To try to reduce the variance, one can repeat the procedure several times and average the results. The leave-one-out estimator, $\varepsilon_{\mathrm{loo}}$, is a special case of cross-validation where the number of subsets equals the number of examples, $k = n$. This estimator is approximately unbiased but has a high variance.

The 0.632 bootstrap estimator is based on resampling. A bootstrap sample S_n^* consists of n equally likely draws with replacement from S_n. At each iteration, a bootstrap sample is generated and used as a training sample. The examples not selected are used as a test sample. The bootstrap zero estimator is the average of the test-sample errors:

$$\varepsilon_{b0} = \frac{\sum_{b=1}^{B} \left(\sum_{i=1}^{n_b} |Y_i^{-b} - (\Lambda, \Omega)(S_n^{*b})(\mathbf{X}_i^{-b})| \right)}{\sum_{b=1}^{B} n_b}, \tag{6}$$

where the examples $\{(\mathbf{X}_i^{-b}, Y_i^{-b}),\ i = 1, \ldots, n_b\}$ do not belong to the bth bootstrap sample. The 0.632 bootstrap estimator is a weighted sum of the resubstitution error and the bootstrap zero error,

$$\varepsilon_{b632} = (1 - 0.632)\varepsilon_{\mathrm{resub}} + 0.632\varepsilon_{b0} \tag{7}$$

the resubstitution error, $\varepsilon_{\mathrm{resub}}$, being the error of the classifier on the training data. The 0.632 bootstrap estimator is known to have a lower variance than cross-validation but can possess different amounts of bias, depending on the classification rule and feature-label distribution. For instance, it can be strongly optimistically biased when using the CART classification rule.

3. PRECISION OF THE ERROR ESTIMATION

The precision of an error estimator relates to the difference between the true and estimated errors, and we require a probabilistic measure of this difference. Here we use the root-mean-square error (square root of the expectation of the squared difference),

$$\mathrm{RMS} = \mathrm{RMS}(F, \Omega, \Lambda, \Xi, D, d, n) = \sqrt{E\left[|\varepsilon_{\mathrm{est}} - \varepsilon_{\mathrm{tru}}|^2 \right]}. \tag{8}$$

It is helpful to understand the RMS in terms of the deviation distribution, $\varepsilon_{\mathrm{est}} - \varepsilon_{\mathrm{tru}}$. The RMS can be decomposed into the bias, $\mathrm{Bias}[\varepsilon_{\mathrm{est}}] = E[\varepsilon_{\mathrm{est}} - \varepsilon_{\mathrm{tru}}]$, of the error estimator relative to the true error, and the deviation variance, $\mathrm{Var}_{\mathrm{dev}}[\varepsilon_{\mathrm{est}}] = \mathrm{Var}[\varepsilon_{\mathrm{est}} - \varepsilon_{\mathrm{tru}}]$, namely,

$$\mathrm{RMS} = \sqrt{\mathrm{Var}_{\mathrm{dev}}[\varepsilon_{\mathrm{est}}] + \mathrm{Bias}[\varepsilon_{\mathrm{est}}]^2}. \tag{9}$$

Moreover, the deviation variance can be further decomposed into

$$\mathrm{Var}_{\mathrm{dev}}[\varepsilon_{\mathrm{est}}] = \sigma_{\mathrm{est}}^2 + \sigma_{\mathrm{tru}}^2 - 2\rho\sigma_{\mathrm{est}}\sigma_{\mathrm{tru}}. \tag{10}$$

This relation is demonstrated in the following manner:

$$\begin{aligned}\mathrm{Var}_{\mathrm{dev}}[\epsilon_{\mathrm{est}}] &= \mathrm{Var}[\epsilon_{\mathrm{est}} - \epsilon_{\mathrm{tru}}] \\ &= E\Big[(\epsilon_{\mathrm{est}} - \epsilon_{\mathrm{tru}} - E[\epsilon_{\mathrm{est}} - \epsilon_{\mathrm{tru}}])^2\Big] \\ &= E\Big[(\epsilon_{\mathrm{est}} - E[\epsilon_{\mathrm{est}}])^2\Big] + E\Big[(\epsilon_{\mathrm{tru}} - E[\epsilon_{\mathrm{tru}}])^2\Big] \\ &\quad - 2E[(\epsilon_{\mathrm{est}} - E[\epsilon_{\mathrm{est}}])(\epsilon_{\mathrm{tru}} - E[\epsilon_{\mathrm{tru}}])] \\ &= \mathrm{Var}[\epsilon_{\mathrm{est}}] + \mathrm{Var}[\epsilon_{\mathrm{tru}}] - 2\mathrm{cov}(\epsilon_{\mathrm{est}}, \epsilon_{\mathrm{tru}}) \\ &= \sigma^2_{\mathrm{est}} + \sigma^2_{\mathrm{tru}} - 2\rho\sigma_{\mathrm{est}}\sigma_{\mathrm{tru}}.\end{aligned} \tag{11}$$

Large samples tend to provide good approximations of the feature-label distribution, and therefore their differences tend not to have a large impact on the corresponding designed classifiers. The stability of these classifiers across different samples means that the variance of the true error is low, so that $\sigma^2_{\mathrm{tru}} \approx 0$. If the classification rule is consistent, then the expected difference between the error of the designed classifier and the Bayes error tends to 0. Moreover, popular error estimates tend to be precise for large samples. The variance caused by random sampling decreases with increasing sample size. Therefore, for a large sample, we have $\sigma^2_{\mathrm{est}} \approx 0$, so that $\mathrm{Var}_{\mathrm{dev}}[\epsilon_{\mathrm{est}}] \approx 0$ for any value of ρ, and the correlation between the true and estimated errors is inconsequential. The situation is starkly different for small samples. Different samples typically yield very different classifiers possessing widely varying errors. For these, σ^2_{tru} is not small, and σ^2_{est} can be substantially larger, depending on the error estimator. If σ^2_{tru} and σ^2_{est} are large, then the correlation plays an important role. For instance, if $\rho = 1$, then

$$\mathrm{Var}_{\mathrm{dev}}[\epsilon_{\mathrm{est}}] \approx (\sigma_{\mathrm{est}} - \sigma_{\mathrm{tru}})^2. \tag{12}$$

But if $\rho \approx 0$, then

$$\mathrm{Var}_{\mathrm{dev}}[\epsilon_{\mathrm{est}}] \approx \sigma^2_{\mathrm{est}} + \sigma^2_{\mathrm{tru}}. \tag{13}$$

This is a substantial difference when σ^2_{tru} and σ^2_{est} are not small. As we will see, small-sample problems with feature selection produce high variance and low correlation between the true and estimated errors.

4. SIMULATION STUDY

The objective of our simulations is to compare the true and estimated errors in several conditions: low dimensional, high-dimensional without feature selection, and high-dimensional with feature selection. These correspond to the three scenarios discussed in the introduction. We have performed three kinds of experiments:

- No feature selection (ns): the data contain a large number of features and no feature selection is performed.
- Feature preselection (ps): a small feature set is selected before the learning process. The selection is not data-driven and the classification design is performed on a low-dimensional data set.
- A feature selection (fs): a feature selection is performed using the data. The selection is included in the learning process.

Our simulation study is based two kinds of data: synthetic data generated from Gaussian models and patient data from two microarray studies, breast cancer, and lung cancer.

4.1. Experimental design

Our simulation study uses the following protocol when using feature selection:

(1) a training set S_{tr} and test set S_{ts} are generated. For the synthetic data, n examples are created for the training set and 10000 examples for the test set. For the microarray data, the examples are separated into training and test sets with 50 examples for the training set and the remaining for the test set;
(2) a feature-selection method is applied on the training set to find a feature subset $\Omega_d(S_{\mathrm{tr}})$, where d is the number of selected features chosen from the original D features;
(3) a classification rule is used on the training set to build a classifier $(\Lambda, \Omega_d)(S_{\mathrm{tr}})$;
(4) the true classification error rate is computed using the test set, $\varepsilon_{\mathrm{tru}} = (1/10000)\sum_{i \in S_{\mathrm{ts}}} |Y_i^{\mathrm{ts}} - (\Lambda, \Omega_d)(S_{\mathrm{tr}})(\mathbf{X}_i^{\mathrm{ts}})|$;
(5) three estimates of the error rate are computed from S_{tr} using the three estimators: leave-one-out, cross-validation, and 0.632 bootstrap.

This procedure is repeated 10 000 times. We consider three feature-selection methods: t-test, relief, and mutual information. And we consider five classification rules: 3-nearest-neighbor (3NN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), linear support vector machine (SVM), and decision trees (CART). For cross-validation, we use 5 runs of 5-fold cross-validation and for 0.632 bootstrap, we do 100 replications.

In the case of feature preselection, a subset of d features is randomly selected before this process, step (2) is omitted, and $d \ll D$. In the case of no feature selection, step (2) is omitted and $d = D$. Also in the case of no feature selection, we do not consider the uncorrelated model. This is because the independence of the features in the uncorrelated Gaussian model suppresses the peaking phenomenon and yields errors very close to 0 with the given variances. This problem could be avoided by increasing the variances, but then the feature-selection procedure would have to yield very high errors (near 0.5) to obtain significant errors with uncorrelated features. The key point is that we cannot compare the feature selection and no feature selection procedures using the same uncorrelated model, and comparison would not be meaningful if we compared them with different uncorrelated models. Since the no feature selection scenario is not used in practice and included only for comparison purposes, we omit it for the uncorrelated models.

4.2. Simulations based on synthetic data

The synthetic data are generated from two-class Gaussian models. The classes are equally likely and the class-conditional densities are defined by $N(\mu_0, \sigma_0\Sigma)$ and $N(\mu_1, \sigma_1\Sigma)$. The mean of the first class is at the origin $\mu_0 = \vec{0}$ and the

Table 1: Parameters of the experiments.

Model	Features	σ	n	D	d	Feat. select.	Classif. rule	Error estimation
Linear	Uncorrelated	0.2	50	200	5	No selection	LDA	Resubstitution
Nonlinear	$\rho = 0$	to	100	400	10	t-test	QDA	Leave-one-out
	Correlated	5			20	Relief	3NN	5×5-fold cross-valid
	$\rho = 0.5$					Mutual information	SVM	0.632 bootstrap
Breast cancer data set			50	2000	10	t-test	LDA	Resubstitution
Lung cancer data set					20	Relief	3NN	Leave-one-out
					30	Mutual information	SVM	5×5-fold cross-valid
					40		CART	0.632 bootstrap

mean of the second is located at $\mu_1 = \vec{A} = [a_0, \ldots, a_D]$, where the a_i are drawn from a beta distribution, $\beta(2, 2)$. Inside a class, all features possess common variance. We consider two structures Σ for the covariance matrix. The first is the identity $\Sigma = \mathbf{I}$, in which the features are uncorrelated and the class-conditional densities are spherical Gaussian. The second is a block matrix in which the features are equally divided into 10 blocks. Features from different groups are uncorrelated and every two features within the same group possess a common correlation coefficient ρ. In the linear models, the variance and covariance matrices of the two classes are equal, $\sigma_0 = \sigma_1$, and the Bayes classifier is a hyperplane perpendicular. In the nonlinear models, the variance and covariance matrices are different, with $\sigma_0 = \sigma_1/\sqrt{2}$. The different values of the parameters can be found in Table 1. Our basic set of synthetic data-based simulations consists of 60 experiments across 15 models. These are listed in Table C1 on the companion website, as experiments 1 through 60. The results about no feature selection experiments can be found on Table C7 on the companion website.

When there is a feature preselection, $\mu_1 = \vec{A} = [a_0, \ldots, a_d]$, the d features are randomly chosen from the original D features. As opposed to the feature-selection case, the selection is done before the learning process and is not data-driven. There is no absolute way to compare the true-error and estimated-error variances between experiments with feature selection and preselection. However, this is not important because our interest is in comparing the regressions and correlations.

4.3. Simulations based on microarray data

The microarray data come from two published studies, one on breast cancer [20] and the other on lung cancer [21]. The breast-cancer data set contains 295 patients, 115 belonging to the good-prognosis class, and 180 belonging to the poor-prognosis class. The lung-cancer data set contains 203 tumor samples, 139 being adenocarcinoma, and 64 being of some other type of tumor. We have reduced the two data sets to a selection of the 2000 genes with highest variance. The simulations follow the same protocol as the synthetic data simulation. The training set is formed by 50 examples drawn without replacement from the data set. The examples not drawn are used as the test set. Note that the training sets are not fully independent. Since they are all drawn from the same data set, there is an overlap between the training sets; however, for a training set size of 50 out of a pool of 295 or 203, the amount of overlap between the training sets is small. The average size of the overlap is about 8 examples for the breast-cancer data sets and 12 examples for the lung-cancer data set. The dependence between the samples is therefore weak and does not have a big impact on the results. The different values of the parameters can be found in Table 1. Our microarray data-based simulations consist of a set of 24 experiments, 12 for breast cancer, and 12 for lung cancer. These are listed in Tables C3 and C5 on the companion website, as experiments 61 through 72 and 73 through 84, respectively.

Note that on microarray data, we cannot perform experiments with feature preselection. The reason is that we do not know the actual relevant features for microarray data. If we do a random selection, then it is likely that the selected features will be irrelevant, so that the estimated and true errors will be close to 0.5, which is a meaningless scenario.

5. RESULTS

5.1. Synthetic data results

Our discussion of synthetic data results focus on experiment 18; similar results can be seen for the other experiments on the companion website. Experiment 18 is based on a linear model with correlated features ($\rho = 0.5$), $n = 100$, $D = 400$, $d = 10$, feature selection by the t-test, and classification by 3NN. The class-conditional densities are Gaussian and possess common variance $\sigma_1 = \sigma_2 = 1$.

Figure 1 shows the estimated- and true-error pairs. The horizontal and vertical axes represent ε_{est} and ε_{tru}, respectively. The dotted 45-degree line corresponds to $\varepsilon_{\text{est}} = \varepsilon_{\text{tru}}$. The black line is the regression line. The means of the estimated and true errors are marked by dots on the horizontal and vertical axes, respectively. The three plots in Figure 1(a) represent the comparison of the true error with the leave-one-out error, the 5×5-fold cross-validation error, and the 0.632-bootstrap error. The difference between the means of the true and estimated errors give the biases of the estimators: $E[\varepsilon_{\text{tru}}] = 0.26$, whereas $E[\varepsilon_{\text{loo}}] = 0.26$, $E[\varepsilon_{\text{cv}}] = 0.27$, and $E[\varepsilon_{b632}] = 0.23$. The leave-one-out and cross-validation estimators are virtually unbiased and the bootstrap is slightly biased. Estimator variance is represented by the width of the scatter plot.

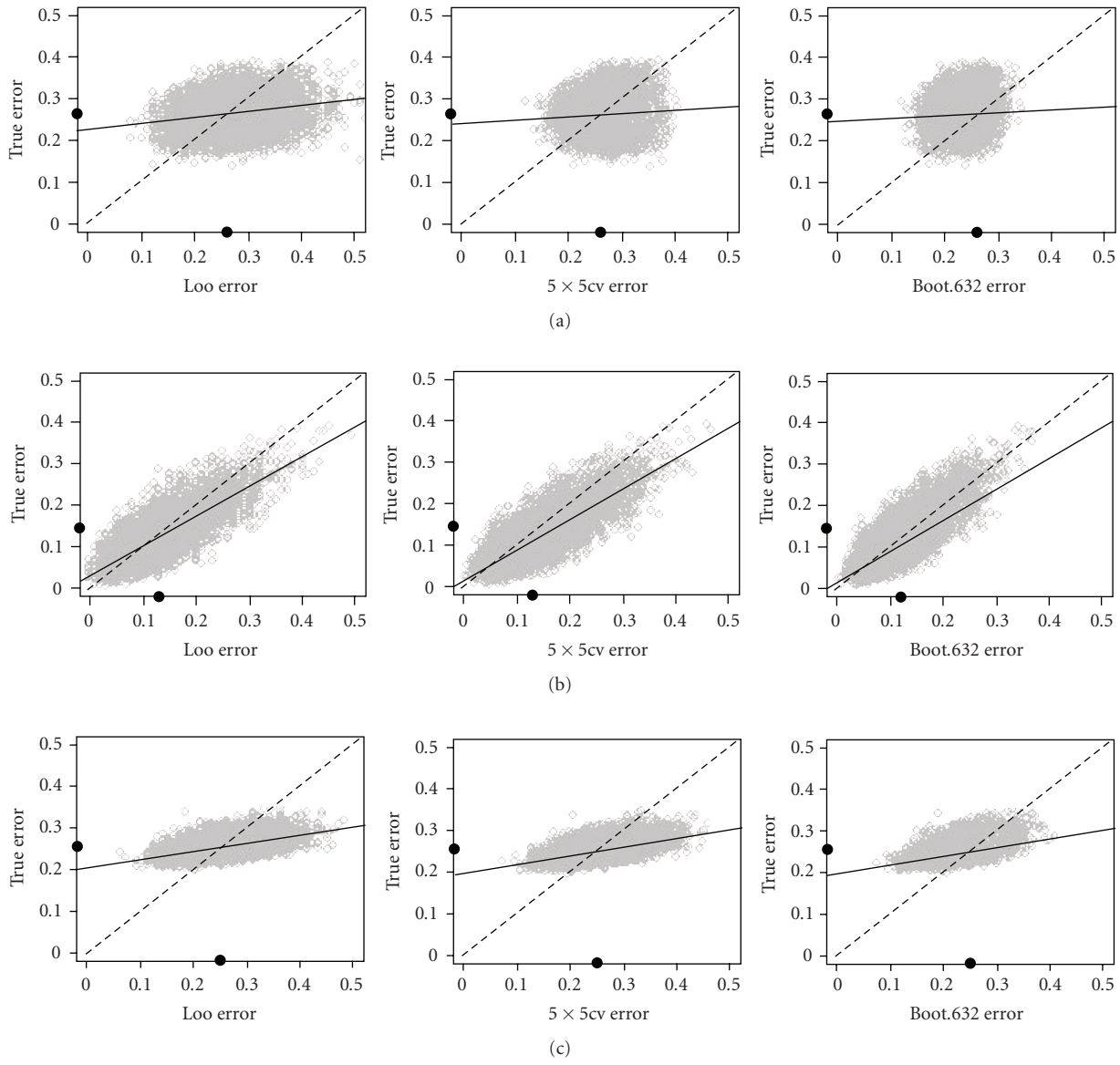

Figure 1: Comparison of the true and estimated errors on artificial data: (a) experiment 18 with linear model, $n = 100$, $D = 400$, $d = 10$, t-test selection and 3NN; (b) experiment 17 with linear model, $n = 100$, $D = 10$, feature preselection and 3NN; (c) experiment 115 with linear model, $n = 100$, $D = 400$, no feature selection and 3NN.

Our focus is on the correlation and regression for the estimated and true errors. When we wish to distinguish feature selection from feature preselection from no feature selection, we will denote these by $\widehat{\rho}_{\mathrm{fs}}$, $\widehat{\rho}_{\mathrm{ps}}$, and $\widehat{\rho}_{\mathrm{ns}}$, respectively. When we wish to emphasize the error estimator, for instance, leave-one-out, we will write $\widehat{\rho}^{\mathrm{loo}}_{\mathrm{fs}}$, $\widehat{\rho}^{\mathrm{loo}}_{\mathrm{ps}}$, or $\widehat{\rho}^{\mathrm{loo}}_{\mathrm{ns}}$. In Figure 1(a), the regression lines are almost parallel to the x-axis. Referring to (2), we see the role of the correlation in this lack of regression, that is, the correlation is small for each estimation rule: $\widehat{\rho}^{\mathrm{loo}}_{\mathrm{fs}} = 0.23$, $\widehat{\rho}^{\mathrm{cv}}_{\mathrm{fs}} = 0.07$, and $\widehat{\rho}^{b632}_{\mathrm{fs}} = 0.18$. Ignoring the bias, which is small in all cases, the virtual loss of the correlation term in (10) means that $\mathrm{RMS}^2 \approx \mathrm{Var}_{\mathrm{dev}}[\varepsilon_{\mathrm{est}}] \approx \sigma^2_{\mathrm{est}} + \sigma^2_{\mathrm{tru}}$, which is not small because σ^2_{est} and σ^2_{tru} are not small.

Let us compare the preceding feature-selection setting with experiment 17 (linear model, 10 correlated features, $n = 100$, feature preselection, 3NN), whose parameters are the same except that there is a feature preselection, the classifier being generated from $d = 10$ features. Figure 1(b) shows the data plots and regression lines for experiment 17. In this case, there is significant regression in all three cases with $\widehat{\rho}^{\mathrm{loo}}_{\mathrm{ps}} = 0.80$, $\widehat{\rho}^{\mathrm{cv}}_{\mathrm{ps}} = 0.80$, and $\widehat{\rho}^{b632}_{\mathrm{ps}} = 0.81$. There is a drastic difference in correlation and regression between the two experiments. We compare now these results with experiment 115 (linear model, 400 correlated features, $n = 100$, no feature selection, 3NN) whose parameters are the same except that there is no feature selection. Figure 1(c) shows the data plots and regression lines for experiment 115. In this case,

Table 2: Correlation of the true and estimated error on the artificial data. "ps" columns contains the correlation where a feature pre-selection is performed, "ns" for no feature selection, "tt" for the t-test selection, "rf" for relief, and "mi" for mutual information. The blanks in the table correspond to the experiments where the covariance matrix is not full rank and not invertible, and therefore the classifiers LDA and QDA cannot be computed, and to the no feature selection case for uncorrelated models.

	Model 1					Model 2					Model 3				
	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi
loo	0.62	—	0.17	0.29	0.22	0.43	—	0.07	0.17	0.11	0.56	—	0.14	0.28	0.21
cv	0.64	—	0.19	0.33	0.25	0.48	—	0.08	0.17	0.12	0.58	—	0.18	0.32	0.26
Boot632	0.64	—	0.19	0.34	0.24	0.49	—	0.06	0.16	0.13	0.59	—	0.18	0.32	0.24
	Model 4					Model 5					Model 6				
	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi
loo	0.32	—	0.19	0.22	0.18	0.8	0.52	0.23	0.32	0.22	0.75	0.1	−0.07	0.07	0.18
cv	0.38	—	0.18	0.21	0.18	0.8	0.56	0.07	0.15	0.06	0.79	0.11	−0.18	−0.07	0.05
Boot632	0.4	—	0.14	0.17	0.16	0.81	0.53	0.18	0.23	0.15	0.78	0.11	0.06	0.19	0.18
	Model 7					Model 8					Model 9				
	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi
loo	0.54	0.52	0.19	0.26	0.18	0.43	0.1	0.17	0.29	0.37	0.32	—	0.29	0.28	0.3
cv	0.54	0.56	0.02	0.11	0.04	0.53	0.11	0.15	0.19	0.32	0.32	—	0.4	0.36	0.39
Boot632	0.57	0.53	0.1	0.16	0.08	0.53	0.11	0.21	0.29	0.29	0.25	—	0.27	0.22	0.28
	Model 10					Model 11					Model 12				
	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi
loo	0.55	—	0.12	0.17	0.13	0.37	—	0.25	0.35	0.28	0.82	—	0.24	0.29	0.29
cv	0.61	—	0.11	0.2	0.14	0.47	—	0.25	0.34	0.25	0.84	—	0.15	0.21	0.2
Boot632	0.62	—	0.09	0.19	0.1	0.48	—	0.2	0.26	0.17	0.84	—	0.21	0.24	0.22
	Model 13					Model 14					Model 15				
	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi	ps	ns	tt	rf	mi
loo	0.92	0.14	0.38	0.45	0.39	0.67	—	0.36	0.38	0.41	0.83	0.14	0.31	0.29	0.28
cv	0.93	0.15	0.26	0.32	0.24	0.72	—	0.4	0.43	0.45	0.86	0.15	0.22	0.21	0.16
Boot632	0.93	0.16	0.42	0.45	0.4	0.6	—	0.29	0.31	0.32	0.86	0.16	0.28	0.27	0.24

there is some regression in all three cases with $\hat{\rho}_{ns}^{loo} = 0.52$, $\hat{\rho}_{ns}^{cv} = 0.56$, and $\hat{\rho}_{ns}^{b632} = 0.53$. The correlation of the no feature selection experiment is lower than the feature preselection experiment but higher than the feature-selection experiment.

Table 2 shows the correlation between the estimated and true errors for all experiments. For each of the 15 models, the 5 columns show the correlations obtained with feature pre-selection (ps), no feature selection (ns), t-test (tt), relief (rf), and mutual information (mi) selection. Recall that we cannot compare no feature selection experiments with the other experiments in uncorrelated models, that is why there are blanks in the columns "ns" of models 1, 2, 3, 4, 9, 10, 11. The other blanks in Table 2 correspond to the experiments where the covariance matrix is not full-rank and not invertible, therefore the classifiers LDA and QDA cannot be computed. In all cases, except with model 9, $\hat{\rho}_{fs} < \hat{\rho}_{ps}$, and often $\hat{\rho}_{fs}$ is very small. In model 9, $\hat{\rho}_{fs} \approx \hat{\rho}_{ps}$, and in several cases, $\hat{\rho}_{fs} > \hat{\rho}_{ps}$. What we observe is that $\hat{\rho}_{ps}$ is unusually small in this model, which has sample size 50 and QDA classification. If we change the sample size to 100 or use LDA instead of QDA, then we have the typical results for all estimation rules: $\hat{\rho}_{ps}$ gets larger and $\hat{\rho}_{fs}$ is substantially smaller than $\hat{\rho}_{ps}$. The correlation in no feature selection experiments depends on the classification rule.

As might be expected, the correlation increases with increasing sample size. This is illustrated in Figure 2, which shows the correlation for increasing sample sizes using model 2 (linear model, 200 uncorrelated features, $n = 50$, $d = 5$, t-test, SVM). As illustrated, the increase tends to be slower with feature selection than with feature preselection. Figure 3 shows the corresponding increase in regression with increasing sample size (see experiments 85 through 97 in Table C1 on the companion website). This increase has little practical impact because, as seen in (10), small error variances imply a small deviation variance, irrespective of the correlation.

Figure 4 compares regression coefficients between no feature selection, feature preselection, and feature-selection experiment: (a) $\hat{a}_{ns}$ and $\hat{a}_{fs}$, (b) $\hat{a}_{ps}$ and $\hat{a}_{ns}$, (c) $\hat{a}_{ps}$ and $\hat{a}_{fs}$. The regression coefficients are compared on models using 3NN and SVM: models 2, 4, 5, 6, 7, 8, 10, 11, 13, and 15. For each model, the comparison is done with the 3 estimation rules (loo, cv, boot). Figures 4(b) and 4(c) show that $\hat{a}_{ps}$ is clearly higher than both $\hat{a}_{ns}$ and $\hat{a}_{fs}$. Figure 4(a) shows that when compared to each other, neither $\hat{a}_{ns}$ nor $\hat{a}_{fs}$ is dominant. In general, no feature selection and feature-selection experiments produce poor regression between

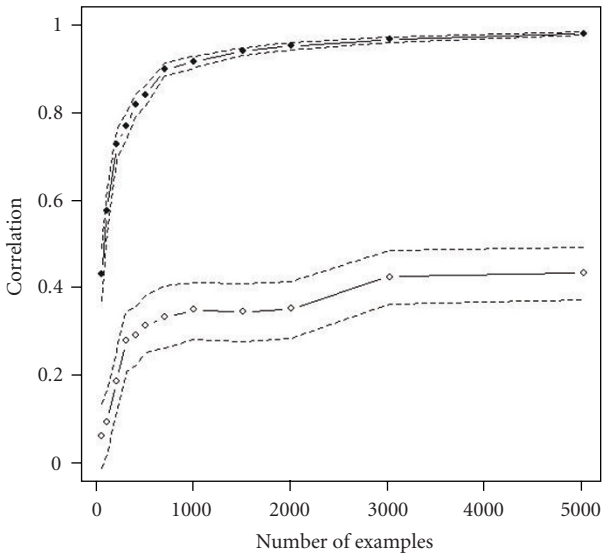

FIGURE 2: Correlation between estimated and true errors as a function of the number of examples. The black dot curve corresponds to the experiments with feature preselection and the white-dot curve to the experiments with feature selection. The dashed lines represent the 95% confidence intervals.

the true and estimated errors, with both $\hat{a}_{\text{ns}}$ and $\hat{a}_{\text{fs}}$ below 0.4.

5.2. Microarray data results

For the microarray data results, we focus on two experiments: 68 (breast-cancer data set, $d = 30$, relief, SVM) and 84 (lung-cancer data set, $d = 40$, mutual information, CART). The results are presented in Figures 5(a) and 5(b), respectively. In each case, there is very little correlation between the estimated and true errors: in the breast-cancer data set, 0.13 for leave-on-out, 0.19 for cross-validation, and 0.16 for bootstrap; in the lung-cancer data set, 0.02 for leave-on-out, 0.06 for cross-validation, and 0.07 for bootstrap. Tables 3 and 4 give the correlation values of all microarray experiments. The results are similar to those obtained with the synthetic data.

5.3. Discussion

It has long been appreciated that the variance of an error estimator is important for its performance [22], but here we have seen the effect of the correlation on the RMS of the error estimator when samples are small. Looking at the decomposition of (10), a natural question arises: which is more critical, the increase in estimator variance or the decrease in correlation between the estimated and true errors? To answer this question, we begin by recognizing that the ideal estimator would have $\hat{a} = 1$ in (2), since this would mean that the estimated and true errors are always equal. The loss of regression, that is, the degree to which $\hat{a}$ falls below 1, depends on the two factors in (2).

Letting

$$\hat{\nu} = \frac{\hat{\sigma}_{\text{tru}}}{\hat{\sigma}_{\text{est}}}. \tag{14}$$

Equation (2) becomes $\hat{a} = \hat{\nu}\hat{\rho}$. What causes more loss of regression, the increase in estimator variance or the loss of correlation, can be analyzed by quantifying the effect of feature selection on the factors $\hat{\nu}$ and $\hat{\rho}$. The question is this: which is smaller, $\hat{\nu}_{\text{fs}}/\hat{\nu}_{\text{ps}}$ or $\hat{\rho}_{\text{fs}}/\hat{\rho}_{\text{ps}}$? If $\hat{\nu}_{\text{fs}}/\hat{\nu}_{\text{ps}} < \hat{\rho}_{\text{fs}}/\hat{\rho}_{\text{ps}}$, then the effect of feature selection on regression is due more to estimator variance than to the correlation; however, if $\hat{\rho}_{\text{fs}}/\hat{\rho}_{\text{ps}} < \hat{\nu}_{\text{fs}}/\hat{\nu}_{\text{ps}}$, then the effect owes more to the correlation.

Figure 6 plots the ratio pairs $(\hat{\rho}_{\text{fs}}/\hat{\rho}_{\text{ps}}, \hat{\nu}_{\text{fs}}/\hat{\nu}_{\text{ps}})$ for the 15 models considered, with t-test and leave-one-out (squares), cross-validation (circles), and bootstrap (triangles). The closed and open dots refer to the correlated and uncorrelated models, respectively. In all cases, $\hat{\rho}_{\text{fs}}/\hat{\rho}_{\text{ps}} < \hat{\nu}_{\text{fs}}/\hat{\nu}_{\text{ps}}$, so that decorrelation is the main reason for loss of regression. For all three error estimators, $\hat{\rho}_{\text{fs}}/\hat{\rho}_{\text{ps}}$ tends to be less than $\hat{\nu}_{\text{fs}}/\hat{\nu}_{\text{ps}}$ to a greater extent in the correlated models, with this effect being less pronounced for bootstrap.

In the same way, Figure 7 shows the comparison of the ratios $\hat{\rho}_{\text{ns}}/\hat{\rho}_{\text{ps}}$ and $\hat{\nu}_{\text{ns}}/\hat{\nu}_{\text{ps}}$. In the majority of the cases, $\hat{\rho}_{\text{ns}}/\hat{\rho}_{\text{ps}} < \hat{\nu}_{\text{ns}}/\hat{\nu}_{\text{ps}}$ demonstrates that again the main reason for loss of regression is the decorrelation between the true and estimated errors.

5.4. Conclusion

Owing to the peaking phenomenon, feature selection is a necessary part of classifier design in the kind of high-dimensional, small-sample settings commonplace in bioinformatics, in particular, with genomic phenotype classification. Throughout our experiments for both synthetic and microarray data, regardless of the classification rule, feature-selection procedure, and estimation method, we have observed that in such settings there is very little correlation between the true and estimated errors. In some sense, it is odd that one would use the random variable ε_{est} to estimate the random variable ε_{tru}, with which it is essentially uncorrelated; however, for large samples, the random variables are more correlated and, in any event, their variances are then so small that the lack of correlation is not problematic. It is the advent of high-feature dimensionality with small samples in bioinformatics, that has brought into play the decorrelation phenomenon, which goes a long way towards explaining the negative impact of feature selection on cross-validation error estimation previously reported [6, 7]. A key observation is that the decrease in correlation between the estimated and true errors in high-dimensional settings has more effect on the loss of regression for estimating ε_{tru} via ε_{est} than does the change in the estimated-error variance relative to true-error variance—with an actual decrease in variance often being the case.

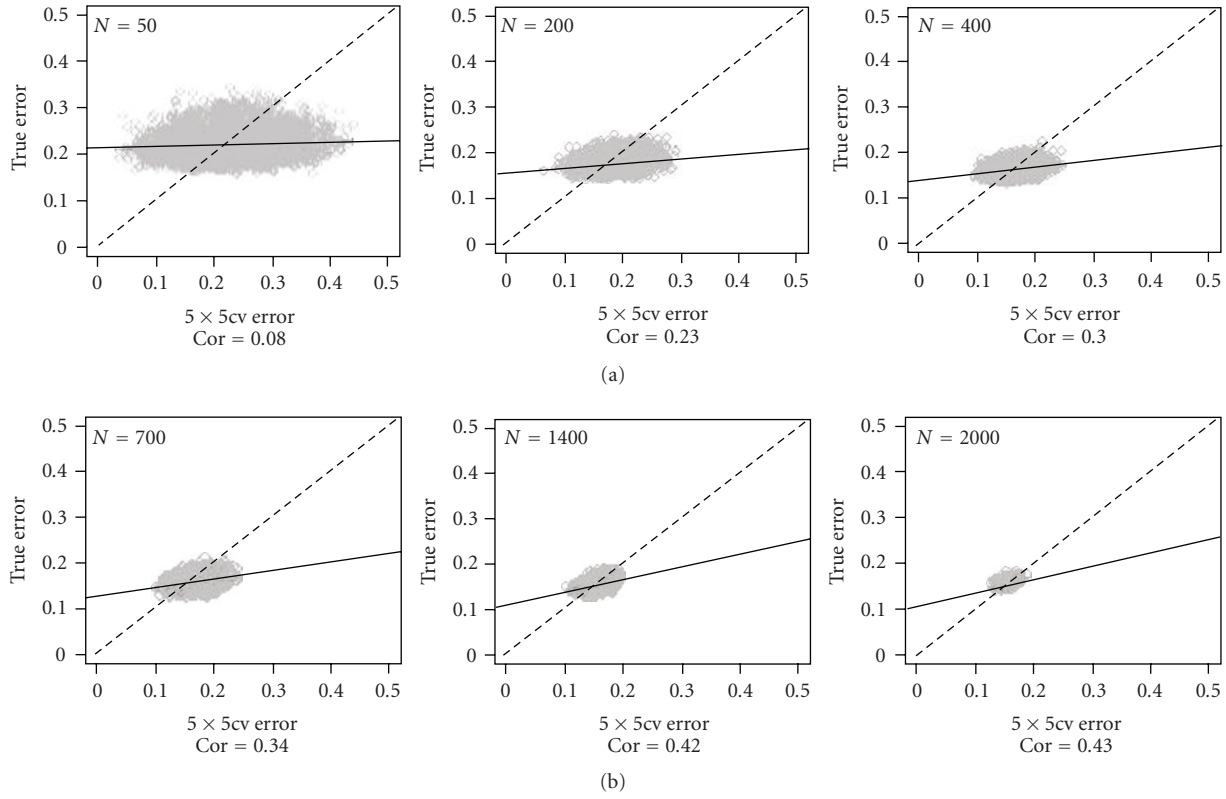

Figure 3: Comparison of the true and estimated errors in the experiment 6 (linear model, 200 uncorrelated features, $d = 5$, t-test, SVM) with different number of examples.

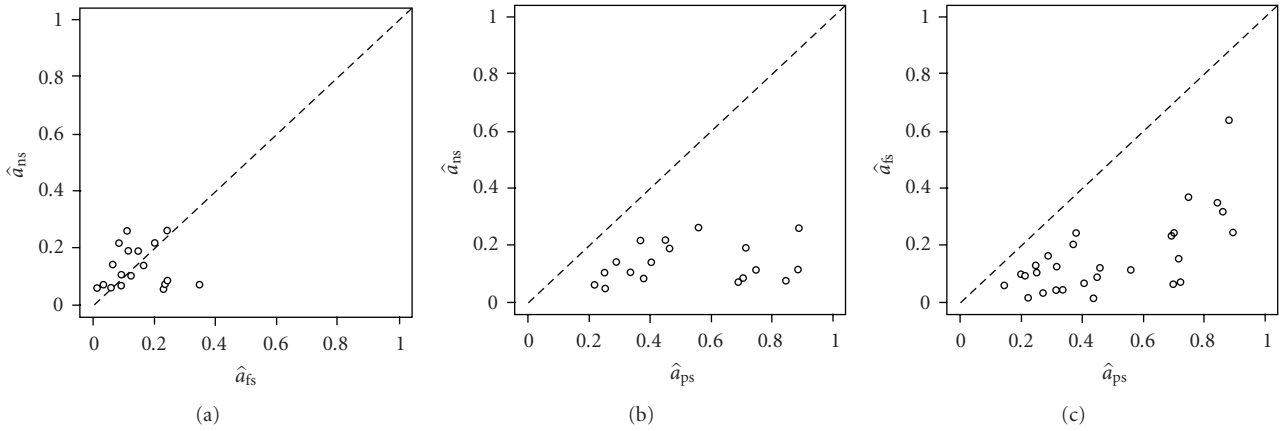

Figure 4: Comparison of the regression coefficient $\hat{a}$ on the artificial data. The left figure shows the comparison between feature selection and no feature selection experiments. The center figure shows the comparison between feature preselection and no feature selection experiments. The right figure shows the comparison between feature preselection and feature-selection experiments.

Table 3: Correlation of the true and estimated error on the breast-cancer data set. "ns" columns contains the correlation where no feature selection is performed, "tt" for the t-test selection, "rf" for relief, and "mi" for mutual information.

	LDA, $d = 10$			3NN, $d = 20$				SVM, $d = 30$				CART, $d = 40$		
	tt	rf	mi	ns	tt	rf	mi	ns	tt	rf	mi	tt	rf	mi
loo	0.06	0.21	0.11	0.28	0.11	0.13	0.15	0.06	0.03	0.13	0.06	0.07	0.13	0.09
cv	0.06	0.24	0.13	0.30	0.12	0.15	0.18	0.08	0.05	0.19	0.08	0.15	0.18	0.15
Boot632	0.03	0.22	0.11	0.21	0.13	0.13	0.17	0.08	0.06	0.16	0.08	0.13	0.17	0.16

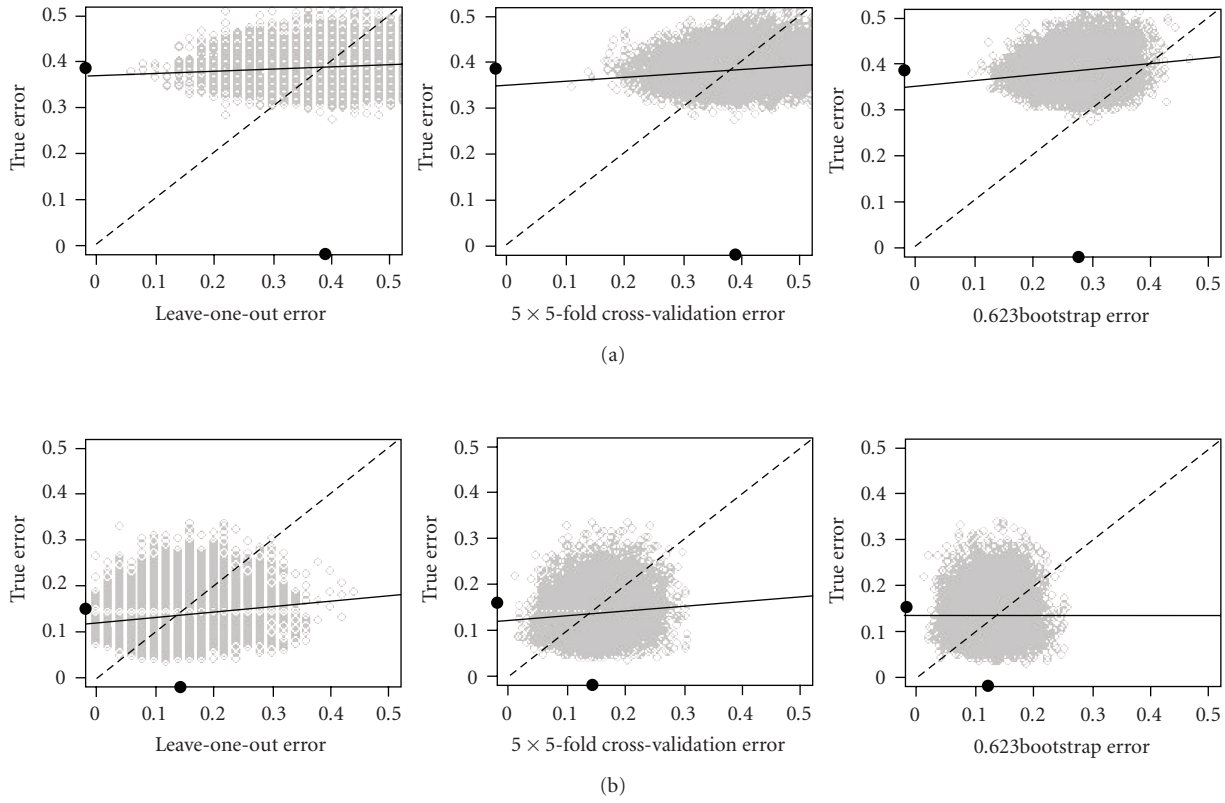

Figure 5: Comparison of the true and estimated errors on microarray data. (a) Experiment 68 with the breast-cancer data set, $d = 30$, relief, and SVM. (b) Experiment 84 with the lung-cancer data set, $d = 40$, mutual information, and CART.

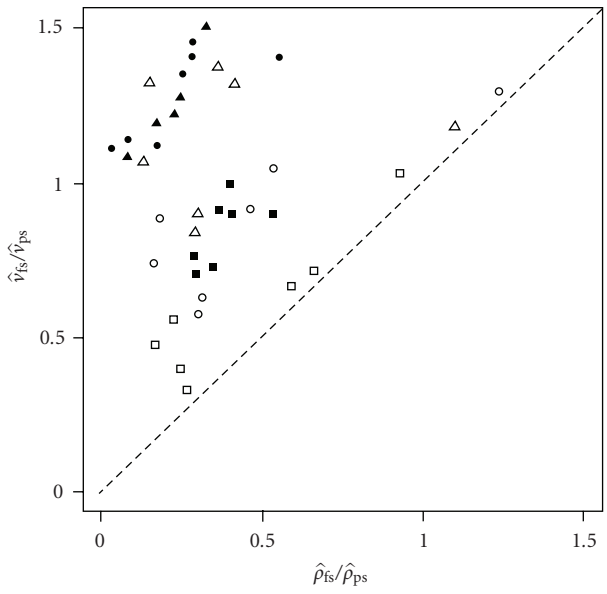

Figure 6: Comparison of the variance and correlation ratios between feature selection and feature preselection experiments. Squares corresponds to experiments with leave-one-out estimators, circles with cross-validation, and triangles with bootstrap. The closed and open dots refer to the correlated and uncorrelated models.

APPENDIX

t-test score

The t-test score measures how much a feature distinguishes two classes: $t = |\mu_0 - \mu_1|/\sqrt{\sigma_0^2/n_0 + \sigma_1^2/n_1}$, where, μ, σ^2, and n are the mean, variance, and number of examples of the classes, respectively.

Mutual information

Mutual information measures the dependence between two variables. It is used to estimate the information that a feature contains to predict the class. A high value of mutual information means that the feature contains a lot of information for the class prediction. The mutual information, $I(X, C)$, is based on the Shannon entropy and is defined in the following manner: $H(X) = -\sum_{i=1}^{m} p(X = x_i) \log p(X = x_i)$ and $I(X, C) = H(X) - H(X \mid C)$.

Relief

Relief is a popular feature selection method in machine learning community [6, 7]. A key idea of the relief algorithm is to estimate the quality of features according to how well their values distinguish between examples that are near to

Table 4: Correlation of the true and estimated error on the lung-cancer data. "ns" columns contains the correlation where no feature selection is performed, "tt" for the t-test selection, "rf" for relief, and "mi" for mutual information.

	LDA, $d = 10$			3NN, $d = 20$				SVM, $d = 30$				CART, $d = 40$		
	tt	rf	mi	ns	tt	rf	mi	ns	tt	rf	mi	tt	rf	mi
loo	0.19	0.43	0.17	0.21	0.07	0.11	−0.02	−0.06	0.19	0.32	0.21	0.03	0.04	0.02
cv	0.16	0.39	0.11	0.27	0.04	0.02	−0.09	−0.12	0.17	0.27	0.16	0.07	0.03	0.06
Boot632	0.11	0.37	0.00	0.21	0.03	0.00	−0.09	−0.18	0.12	0.10	0.08	0.07	0.05	0.07

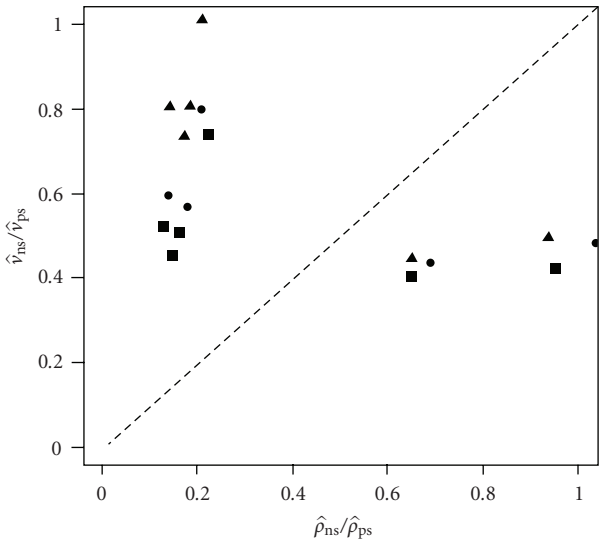

Figure 7: Comparison of the variance and correlation ratios between feature preselection and no feature selection experiments. Squares correspond to experiments with leave-one-out estimators, circles with cross-validation, and triangles with bootstrap.

Require: A data set containing n examples and d features
Require: parameter k
Ensure: Weight of each feature W
  Initialize all feature weights $W[i] \leftarrow 0$
  for all features i **do**
    for all examples j **do**
      $C_j \leftarrow$ class of example j
      $Z_j^s \leftarrow$ the k nearest neighbors belonging to C_j
      $Z_j^o \leftarrow$ the k nearest neighbors belonging to an other class than C_j
      for l such that $l \in Z_j^s$ **do**
        $W[i] \leftarrow W[i] - \sum_l \text{distance}(j, l)$
      end for
      for l such that $l \in Z_j^o$ **do**
        $W[i] \leftarrow W[i] + \sum_l \text{distance}(j, l)$
      end for
    end for
  end for

Algorithm 1: Relief algorithm.

each other. For that purpose, given a randomly selected example X, relief searches for its $2k$ nearest neighbors: k from the same class Z_i^s and k from the other class Z_i^o. It updates the quality estimation $W[F]$ for all features F depending on the values of X, Z_i^s, and Z_i^o. If X and Z_i^s have different values for the feature F, then this feature separates two examples of the same class. It is not desirable, and therefore its quality estimation $W[F]$ is decreased. On the other hand, if example X and Z_i^o have different values of the feature F, then the feature F separates two examples of different classes. It is desirable, and therefore its quality estimation $W[F]$ is increased. This process is repeated for each example.

ACKNOWLEDGMENTS

The authors would like acknowledge the Translational Genomics Research Institute, the National Science Foundation (CCF-0634794), and the French Ministry of Foreign Affairs for providing support for this research.

REFERENCES

[1] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[3] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005.

[4] E. R. Dougherty, "Small sample issues for microarray-based classification," *Comparative and Functional Genomics*, vol. 2, no. 1, pp. 28–34, 2001.

[5] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[6] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005.

[7] Y. Xiao, J. Hua, and E. R. Dougherty, "Quantification of the impact of feature selection on the variance of cross-validation error estimation," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, Article ID 16354, 11 pages, 2007.

[8] U. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, "Is cross-validation better than resubstitution for ranking genes?" *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.

[9] C. Sima, U. Braga-Neto, and E. R. Dougherty, "Superior feature-set ranking for small samples using bolstered error estimation," *Bioinformatics*, vol. 21, no. 7, pp. 1046–1054, 2005.

[10] C. Sima, S. Attoor, U. Braga-Neto, J. Lowey, E. Suh, and E. R. Dougherty, "Impact of error estimation on feature-selection algorithms," *Pattern Recognition*, vol. 38, no. 12, pp. 2472–2482, 2005.

[11] X. Zhou and K. Z. Mao, "The ties problem resulting from counting-based error estimators and its impact on gene selection algorithms," *Bioinformatics*, vol. 22, no. 20, pp. 2507–2515, 2006.

[12] C. Sima and E. R. Dougherty, "What should be expected from feature selection in small-sample settings," *Bioinformatics*, vol. 22, no. 19, pp. 2430–2436, 2006.

[13] T. Mehta, M. Tanik, and D. B. Allison, "Towards sound epistemological foundations of statistical methods for high-dimensional biology," *Nature Genetics*, vol. 36, no. 9, pp. 943–947, 2004.

[14] E. R. Dougherty, A. Datta, and C. Sima, "Research issues in genomic signal processing," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 46–68, 2005.

[15] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[16] E. R. Dougherty and U. Braga-Neto, "Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity," *Journal of Biological Systems*, vol. 14, no. 1, pp. 65–90, 2006.

[17] U. Braga-Neto, "Fads and fallacies in the name of small-sample microarray classification—a highlight of misunderstanding and erroneous usage in the applications of genomic signal processing," *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 91–99, 2007.

[18] A. Dupuy and R. M. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *Journal of the National Cancer Institute*, vol. 99, no. 2, pp. 147–157, 2007.

[19] E. R. Dougherty, J. Hua, and M. L. Bittner, "Validation of computational methods in genomics," *Current Genomics*, vol. 8, no. 1, pp. 1–19, 2007.

[20] M. J. van de Vijver, Y. D. He, L. J. van't Veer, et al., "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[21] A. Bhattacharjee, W. G. Richards, J. Staunton, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[22] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, NY, USA, 1996.