BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 6 2010, pages 822–830
doi:10.1093/bioinformatics/btq037

Data and text mining

Advance Access publication February 3, 2010

Small-sample precision of ROC-related estimates

Blaise Hanczar[1], Jianping Hua[2], Chao Sima[2], John Weinstein[3], Michael Bittner[2] and Edward R. Dougherty[2,3,4,*]

[1]LIPADE, University Paris Descartes, 45 rue des Saint-Peres, Paris, France, [2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, [3]Department of Bioinformatics and Computation Biology, MD Anderson Cancer Center, Houston and [4]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The receiver operator characteristic (ROC) curves are commonly used in biomedical applications to judge the performance of a discriminant across varying decision thresholds. The estimated ROC curve depends on the true positive rate (TPR) and false positive rate (FPR), with the key metric being the area under the curve (AUC). With small samples these rates need to be estimated from the training data, so a natural question arises: How well do the estimates of the AUC, TPR and FPR compare with the true metrics?

Results: Through a simulation study using data models and analysis of real microarray data, we show that (i) for small samples the root mean square differences of the estimated and true metrics are considerable; (ii) even for large samples, there is only weak correlation between the true and estimated metrics; and (iii) generally, there is weak regression of the true metric on the estimated metric. For classification rules, we consider linear discriminant analysis, linear support vector machine (SVM) and radial basis function SVM. For error estimation, we consider resubstitution, three kinds of cross-validation and bootstrap. Using resampling, we show the unreliability of some published ROC results.

Availability: Companion web site at http://compbio.tgen.org/paper_supp/ROC/roc.html

Contact: edward@mail.ece.tamu.edu

Received on December 14, 2009; revised on January 22, 2010; accepted on January 25, 2010

1 INTRODUCTION

High-throughput technologies, such as those based on microarrays or 'Next-Generation' sequencing, make it possible to generate data on large numbers of genes, transcripts or proteins simultaneously in biological samples. Typical variables assessed include mutations, DNA copy number, DNA methylation, mRNA expression, microRNA expression, protein expression and post-translational modifications. A central goal of current biomedical research is to use those molecular profiles to identify biomarkers or multi-gene biosignatures for 'personalization' of medicine—that is, to use them for the full range of medical management choices—in disease risk assessment, sub-classification of disease, early diagnosis, prognosis, choice of optimal therapy, evaluation of response to therapy and/or identification of relapse.

*To whom correspondence should be addressed.

The profile data are used to develop univariate or multivariate predictors of biologically or medically interesting outcomes. Often, the aim is to develop a binary classifier, for example, diseased versus normal, disease subtype 1 versus disease subtype 2, response versus non-response to a drug, or 5-year survival versus death. A large literature has developed on such classifiers, but the recurring question is, 'How accurate are their predictions and classifications?' This question is supposed to be answered by the error rate; however, recent Monte Carlo simulations have shown large uncertainty in the error estimates. In the presence of high-dimensional feature spaces and small samples, a ubiquitous situation with high-throughput technologies, resampling error estimation methods, for example, cross-validation (CV), suffer from high-deviation variance, that is, the variance of the difference between the true and estimated errors is large (Braga-Neto and Dougherty, 2004) [see Glick (1978) for an early criticism of CV]. Moreover, there tends to be a lack of correlation and regression between the true and estimated errors, to the extent that the regression line of the true error on the estimated error is nearly horizontal (Hanczar *et al.*, 2007). These Monte Carlo studies have been supported by analytical studies in the case of the discrete histogram rule (Braga-Neto and Dougherty, 2005b) and linear discriminant analysis (LDA; Zollanvari *et al.*, 2009).

For assessment of binary classifiers, in addition to the error rate, a favorite analytical tool is the receiver operator characteristic (ROC) representation (Pepe *et al.*, 2004; Spackman, 1989)—for instance, with regard to gene-expression profiling in cancer, see Table C1 on the companion web site (http://compbio.tgen.org/paper_supp/ROC/roc.html). An ROC curve is formulated by plotting the sensitivity and specificity of the classifier against each other as a function of some threshold criterion, for example, based on a biomarker or biosignature. The resulting ROC curve presents graphically the trade-off between false positives (FP) and false negatives (FN) in the classification process. The area under the ROC curve provides a scalar parameter that reflects the overall quality of the classifier. A natural question is whether parameters associated with ROC curves, such as the area under the curve (AUC), would suffer the same degree of uncertainty as discovered in the previous analyses of classifier error.

Accordingly, we have established the computational machinery to address this question for both simulated and real datasets, and have performed a variety of analyses based on different predictive algorithms and methods of validation. We have analyzed the effect of sample size and the effect of an unbalance in the number of cases per class. That type of imbalance is common in biological datasets.

Although ROC curves are insensitive to changes in class proportion, we show here that such imbalances have considerable impact on the estimation of both error rate and AUC. Through the simulations on both synthetic and real data, we identify how the training set size and class disproportion affect the performances of the different metrics. In particular, we show the unreliability of ROC performance metric estimations in small-sample settings.

2 SYSTEMS AND METHODS

2.1 ROC curves

Consider a two-class problem defined by the feature-label distribution F and a sample $S=\{(x_1,y_1),\ldots,(x_N,y_N)\}$ of N examples drawn from F. An example is a pair (x,y), where x is a d-dimensional vector and $y\in\{0,1\}$ is the class. A classification rule Ψ is used to design a discriminant $\Psi_S:\mathbb{R}^d\to\mathbb{R}$ from S. The output of Ψ_S is a probability or a score that reflects the degree of uncertainty with which an example is assigned to a class. A binary classifier $\Psi_{S,T}$ is derived from Ψ_S via a threshold T according to $\Psi_{S,T}(x)=0$ if $\Psi_S(x)>T$ and $\Psi_{S,T}(x)=1$ otherwise.

Given an example x, there are four possibilities when comparing the class predicted by $\Psi_{S,T}(x)$ to its true class y: true positive (TP) $y=0$ and $\Psi_{S,T}(x)=0$; FN $y=0$ and $\Psi_{S,T}(x)=1$; FP $y=1$ and $\Psi_{S,T}(x)=0$; true negative (TN) $y=1$ and $\Psi_{S,T}(x)=1$. From these four possibilities, we can define three performance metrics; classifier error, $\mathrm{ERR}=(\mathrm{FP}+\mathrm{FN})/N$; true positive rate, $\mathrm{TPR}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$; and false positive rate, $\mathrm{FPR}=\mathrm{FP}/(\mathrm{TN}+\mathrm{FP})$.

An ROC graph is a 2D graph in which the x-axis represents the FPR and y-axis represents the TPR. The point $(0,1)$ represents perfect classification: no negatives classified as positives and all positives classified as positive. On the diagonal, the points $(0,0)$ and $(1,1)$ correspond to all examples being assigned to the negative class and to the positive class, respectively. The performance of a classifier $\Psi_{S,T}$ for a fixed threshold T is represented by a single point in ROC space. If the decision threshold T is allowed to vary, then the performance of the discriminant Ψ_S is a variable depending on T and is represented by a curve in ROC space. A common metric to estimate the performance of a classifier independently of the decision threshold is the *Area Under the Curve* (AUC). $\mathrm{AUC}\in[0,1]$, AUC = 1 corresponds to the perfect classifier, for which the ROC curve goes directly from point $(0,0)$ to $(0,1)$ and then to $(1,1)$, AUC = 0 corresponds to the classifier assigning all examples to the wrong class, and the ROC curve that follows the diagonal line has $\mathrm{AUC}=0.5$.

A direct method to compute the AUC is to construct the ROC curve and then measure the AUC. If there are M test examples, then we obtain up to $M+1$ points in the ROC space with which to draw the curve. Accordingly, the AUC can be estimated by applying a rectangle or trapezoid area on each point. However, an alternative of AUC computation has been proposed in (Hand and Till, 2001), where it is shown that the AUC corresponds to the probability that an example from the positive class has a higher classifier output $\Psi_S(x)$ than an example from the negative class. In their procedure, the examples are sorted in increasing order according to the values $\Psi_S(x)$ and the AUC is computed by the following formula:

$$\mathrm{AUC}=\frac{S_0-n_0(n_0+1)/2}{n_0n_1} \tag{1}$$

where n_0 and n_1 are the numbers of examples of the positive and negative classes, respectively, in the test set, and S_0 is the sum of ranks of examples in the positive class.

2.2 Models for synthetic data

We have performed a set of experiments on synthetic data generated to reflect key properties of microarray data: small number of examples, high dimension and large number of irrelevant features. We construct two types of features: relevant and irrelevant. Relevant features follow different normal distributions for each class: $N(\mathbf{0},\sigma_0\Sigma)$ for the positive class and $N(\mathbf{A},\sigma_1\Sigma)$ for the negative class, where the elements of $\mathbf{A}$ are evenly drawn from $[0.5,1.5]$ and fixed throughout so as not to confound model variability with error estimation. An irrelevant feature follows the same normal distribution, $N(0,\sigma_0)$, for both classes. Inside a class, all relevant features have a common variance. Two covariance matrix structures Σ are considered: (i) Σ is the identity matrix $\mathbf{I}$ in which the features are uncorrelated and the class-conditional densities are spherical Gaussian. (ii) Σ is a block-structured matrix in which the features are equally divided into blocks of size 4: features from different blocks are uncorrelated and every two features within the same block have a common correlation coefficient $\rho=0.8$. In the linear models, the variances of covariance matrices of the two classes are equal, $\sigma_1=\sigma_0=1.8$; in the non-linear models, the variances of covariance matrices are different, with $\sigma_0=\sigma_1/1.5=1.4$. With these model characteristics, we generate training and test sets with N and 10 000 examples, respectively, containing 20 relevant features and 180 irrelevant features. The large test set is used to compute the true metrics. The number of training examples (N) and class prior probabilities (p_0 for class 0, and $p_1=1-p_0$ for Class 1) of the two classes are parameters of the dataset, with N varying from 50 to 1000, and p_0 from 0.2 to 0.8.

2.3 Classification rules

We consider three classification rules: LDA, linear support vector machine (SVM) and radial basis function SVM (RBF-SVM). The output of an LDA classifier is readily transformed into the posterior probability of the positive class, so the usual decision threshold is 0.5. The output of an SVM is a score that represents the distance of the example from the separating hyperplane and the sign of the score defines the predicted class. The usual decision threshold of an SVM is 0. The RBF-SVM is in general a non-linear classifier, although linear SVM can be viewed as a special form of it.

3 IMPLEMENTATION

Our simulation study uses the following protocol:

(i) A training set S_{tr} and a test set S_{ts} are generated. For the synthetic data, examples are sampled from the distribution determined by the model, N examples for S_{tr} and 10 000 examples for S_{ts}. For the microarray data, the examples are randomly separated into training and test sets with $N=50$ examples for the training set and the remaining for the test set.

(ii) To design a classifier based on a dataset S: first apply t-test to S and select 10 best features based on the t-test statistics; then build the classifier Ψ_S from the reduced set. For classification, the decision threshold T is as defined in Section 2.3.

(iii) For true performance:

- (a) Based on training data S_{tr}, build the classifier $\Psi_{S_{tr}}$.
- (b) Apply $\Psi_{S_{tr}}$ to test data S_{ts}. For each example x, compute class prediction $\Psi_{S_{tr},T}(x)$ and classifier output $\Psi_{S_{tr}}(x)$.
- (c) Based on class predictions and true labels, compute true error rate, TPR and FPR.
- (d) Sort $\Psi_{S_{tr}}(x)$, then compute true AUC according to Equation (1).

(iv) For estimated performance, consider the following estimators:

- (a) resubstitution:
  - (1) Based on training data S_{tr}, build the classifier $\Psi_{S_{tr}}$. Apply $\Psi_{S_{tr}}$ back to training data S_{tr}. For each example compute class prediction and classifier output.
  - (2) Compute estimated error rate, TPR and FPR with class predictions, and estimated AUC with sorted classifier outputs.
- (b) k-fold CV:

(1) Randomly partition the training data into k folds $S_{tr}^{(i)}$, $i=1,\ldots,k$. For each fold $S_{tr}^{(i)}$, based on the remaining data in the training set $S_{tr}\setminus S_{tr}^{(i)}$, build the classifier $\Psi_{S_{tr}\setminus S_{tr}^{(i)}}$. Apply $\Psi_{S_{tr}\setminus S_{tr}^{(i)}}$ to fold $S_{tr}^{(i)}$ to generate each example's class prediction and classifier output.

(2) Collect the class predictions and classifier outputs from all folds. Compute estimated error rate, TPR, FPR and AUC.

(3) For leave-one-out (LOO), set k to the training sample size; for 10-fold CV (10CV), set $k=10$; for 10CV with 10 repetitions (10CV10), repeat Step (1) for 10 times before entering Step (2).

(c) .632 bootstrap (BOOT):

(1) Form a bootstrap sample S^* of size N by drawing with replacement from S_{tr}. Using S^*, build the classifier Ψ_{S^*}. Apply Ψ_{S^*} to the examples that are in S_{tr} but not in S^*.

(2) Repeat the above step 100 times, collect class predictions and classifier outputs from all repetitions, compute estimated error rate, TPR, FPR and AUC, and denote them as ε_0, TPR_0, FPR_0 and AUC_0, respectively.

(3) Obtain the resubstitution estimate of error rate, TPR, FPR and AUC, and denote them as $\varepsilon_{\text{resub}}$, $\text{TPR}_{\text{resub}}$, $\text{FPR}_{\text{resub}}$ and $\text{AUC}_{\text{resub}}$, respectively.

(4) Compute the .632 bootstrap estimated error rate by

$$\varepsilon_{.632}=(1-0.632)\varepsilon_{\text{resub}}+0.632\varepsilon_0.$$

Replace ε in above equation with TPR, FPR and AUC to obtain the .632 bootstrap estimation of TPR, FPR and AUC, respectively.

(v) Repeat the above procedure for 5000 times and collect all results.

4 RESULTS AND DISCUSSION

In this article, we discuss representative results with the full set of results being given on the companion web site. For the synthetic data, we restrict ourselves here to the linear SVM and CV error estimation for the linear model with uncorrelated data, unless specifically indicated. For real data, again we demonstrate only the results of linear SVM with CV error estimation, unless specifically indicated.

4.1 Results for synthetic data

Figure 1 presents the deviation distributions (true minus estimated metric) for classifier error, AUC, TPR and FPR, with $N=100$ and $p_0=0.5,0.7$ and for five error estimators. As expected, leave-one-out is practically unbiased but has the largest deviation variance. Generally, for this case, bootstrap and the CV estimators are nearly unbiased. Bootstrap has the smallest deviation variance except for resubstitution, which suffers from severe bias. However, one must be careful about generalizing from the bootstrap bias results, since .632 bootstrap is known to have substantial bias for certain models and classifiers.

Figure 2 shows scatter plots comparing the true and estimated values of the four metrics for $N=50,100,200$ (500 and 1000 are on the companion web site) and $p_0=0.5,0.7$. The estimated and true values are on the x- and y-axis, respectively, and the small triangles indicate the mean true and mean estimated values. The black line shows the linear regression for the true error on the estimated error. The lack of error regression and wide dispersion for small N is consistent with what we have previously reported

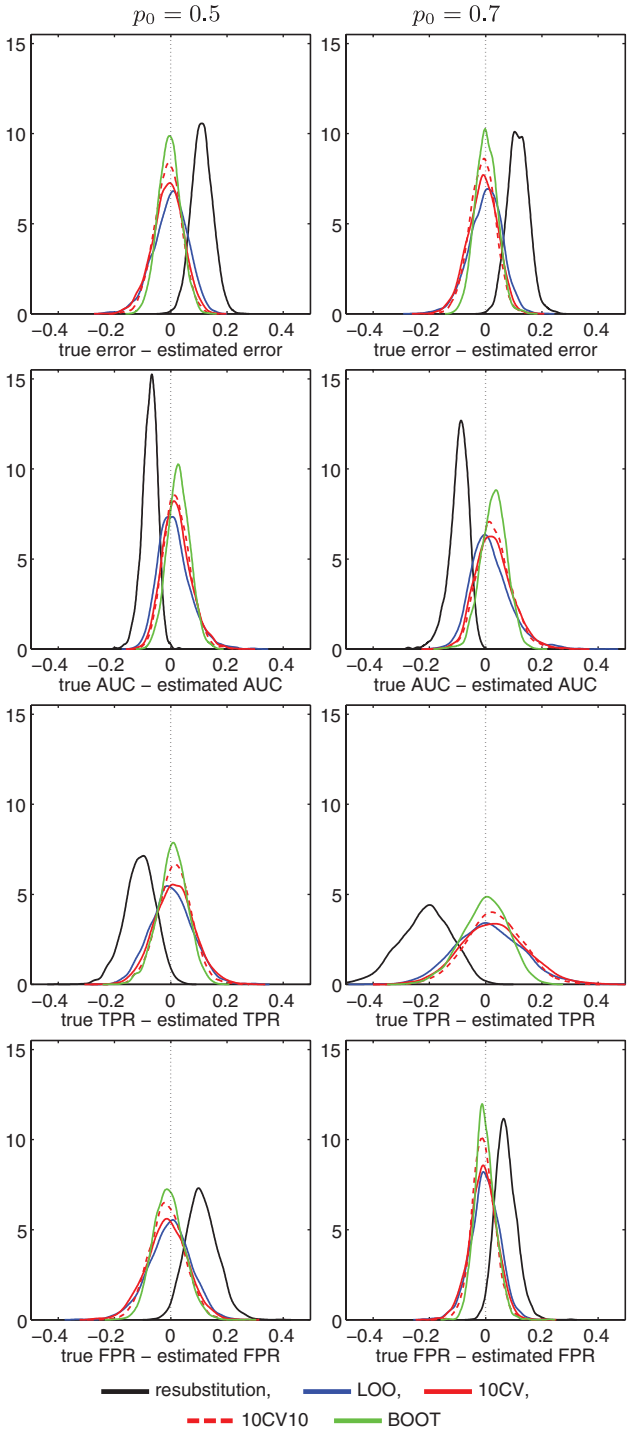

Fig. 1. Deviation distribution of various error estimation schemes for four performance measures: synthetic data, linear uncorrelated model, linear SVM and sample size $N=100$.

for error estimation (Hanczar *et al.*, 2007). Of interest here, the dispersion is worse for the AUC than for the classifier error and that it is worse for the unbalanced prior ($p_0=0.7$) than the balanced prior ($p_0=0.5$). Note that the TPR variance is particularly bad for the unbalanced prior, a finding common throughout this study. There

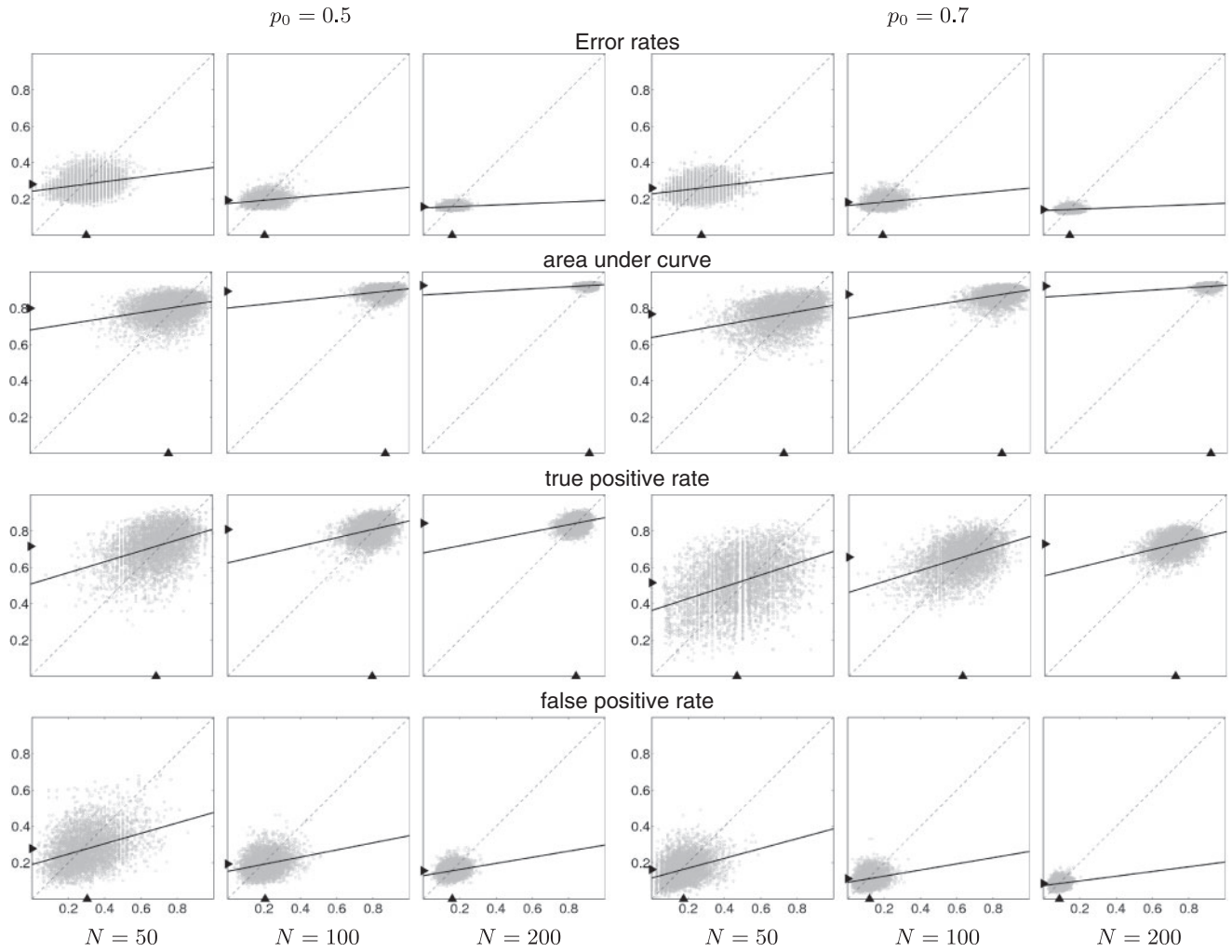

Fig. 2. The scatter plots of various performance measures at different sample size *N*s and priors: synthetic data, linear uncorrelated model, linear SVM and 10CV. The *x*-axis is the estimated performance, whereas the *y*-axis is the true performance.

is also little regression for the true AUC on the estimated AUC. On the companion web site, it can be seen that AUC regression generally does not improve for large sample sizes; however, the variance decreases greatly for $N = 500, 1000$. Hence, estimation is good, even with a lack of regression.

Figure 3 shows the root mean square (RMS) error and the correlation between the true and estimated metrics as functions of the class prior probabilities (on the *x*-axis) for $N = 50, 100, 200, 500, 1000$. RMS of the first row is defined by

$$\mathrm{RMS} = \sqrt{E[|\varepsilon_{\mathrm{est}} - \varepsilon_{\mathrm{tru}}|^2]}$$

where $\varepsilon_{\mathrm{tru}}$ and $\varepsilon_{\mathrm{est}}$ are the true and estimated metrics. The second row in the figure shows the correlation between true and estimated metrics. The RMS is strongly negatively correlated with the training set size, RMS decreases as N increases. The prior probability also impacts the RMS. The classifier error decreases slightly and the AUC increases slightly with imbalance between the classes. Relative to the prior probability, RMS for the TPR increases substantially and RMS for the FPR decreases substantially for increasing prior probability. In all cases, sensitivity to the prior is substantial for small samples and decreases with increasing N.

RMS can be decomposed into the bias and deviation variance as

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

Figure 2 shows that the bias is small for the CV estimator being considered, so that both classifier error and AUC imprecision result from the deviation variance. The deviation variance can be further decomposed into the variances of the true and estimated metrics, along with the correlation, ρ, between true and estimated metrics:

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

According to Figure 3, the correlation is typically not large and cannot offset the $\sigma^2_{\mathrm{est}} + \sigma^2_{\mathrm{tru}}$ term in the deviation variance for the classifier error or the AUC. In sum, the AUC is poorly estimated for small samples, particularly so for $N \leq 100$.

The effect of imprecise estimation is observable in the ROC curves themselves. Figure 4 shows ROC-related curves for LDA classification and the non-linear uncorrelated model.

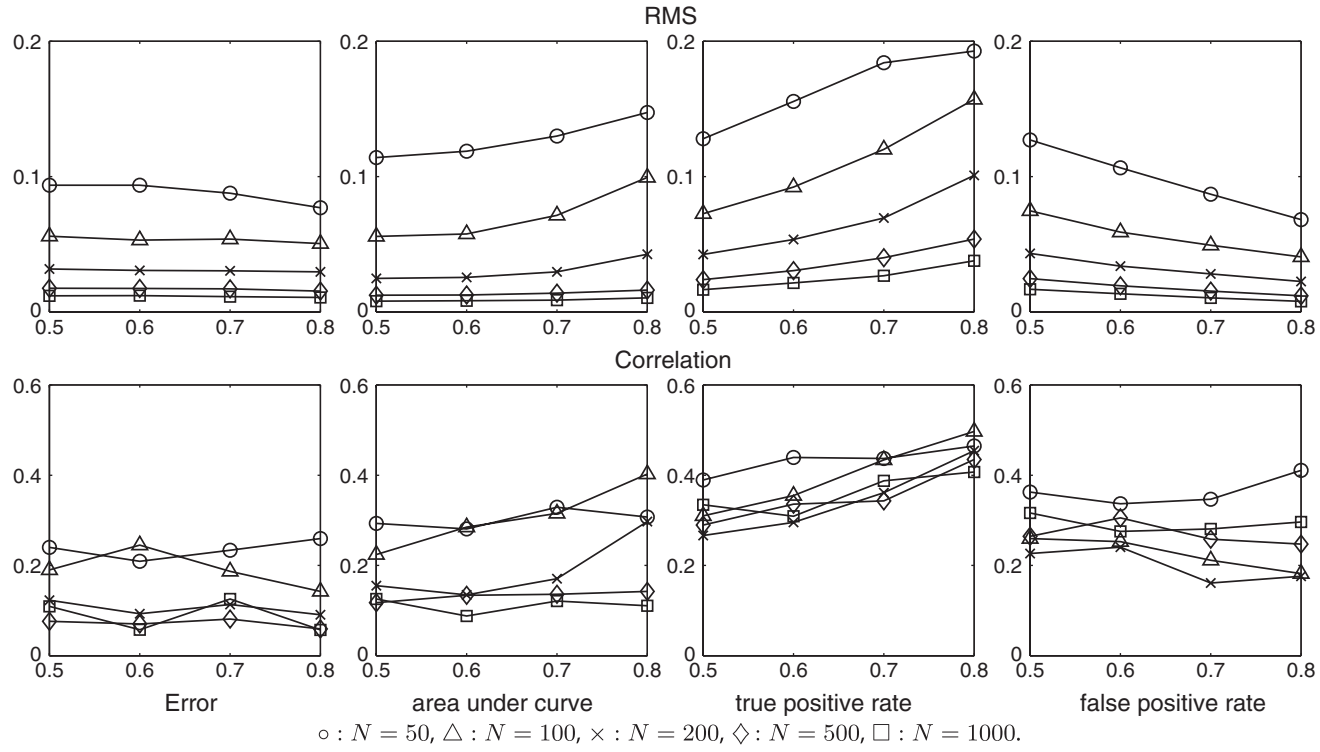

Fig. 3. The precision of four performance measures, under different sample size Ns and priors: synthetic data, linear uncorrelated model, linear SVM and 10CV. The x-axis is the p_0 of the distribution and the y-axis is the precision based on either RMS (top row) or correlation (bottom row).

The left- and right-hand columns show results for $p_0=0.5$ and $p_0=0.7$, respectively, and $N=50,100$ and 200. The cross- and circle-marked curves correspond to using LOO error estimation and the true error, respectively. The solid lines are the mean ROC curves and the upper and lower dashed lines are the 95% confidence interval. The circle-marked dashed lines represent the variation associated with computing ROC curves from samples, that is, constructing the curves from sample-based TPRs and FPRs. The extra variance represented by the cross-marked dashed lines results from using the estimated TPR and estimated FPR instead of the true TPR and true FPR. For $N\leq 100$, this extra variance is substantial.

4.2 Analytic representation of estimated AUC variance

We have observed that the variance of the estimated AUC often exceeds the variance of the estimated error in the simulations. Although we cannot prove a general theorem to that effect, we can give an analytic proof for a special case. For simplicity, we consider the one-split training–testing scheme, but the conclusion can easily be extended to many other schemes.

For the feature vector X and binary label variable Y, let the two class-conditional distributions be identical, the corresponding cumulative distribution functions (CDFs) be continuous over $\mathcal{R}$, and the prior class probabilities be given by $P\{Y=0\}=p$ and $P\{Y=1\}=1-p$. The Bayes error is $\min(p,1-p)$. For AUC estimation, the classification rule adopted will yield a discriminant $\Psi_S:\mathcal{R}\rightarrow[0,1]$ from the training data. Let the CDFs of $\Psi_S(X|Y=0)$ and $\Psi_S(X|Y=1)$ be continuous over $[0,1]$. Assume there are altogether n testing sample points $x_1,\ldots,x_n$, with $n_0=pn$ points from class 0 and $n_1=(1-p)n$ points from Class 1. The AUC is computed according to Equation (1). Owing to the continuity assumption, $P[\Psi_S(x_i)=\Psi_S(x_j)]=0$. Hence, the ranking order is unique with probability 1.

Since the class-conditional distributions are identical, $\Psi_S(X|Y=0)$ and $\Psi_S(X|Y=1)$. Thus, the rank-sum S_0 follows the same distribution of the null-hypothesis in the Mann–Whitney test (Mann and Whitney, 1947), whose variance has been shown to be $n_0n_1(n_0+n_1+1)/12$. Hence, the variance of the estimated AUC is

$$\sigma^2_{\text{est-AUC}}=\frac{n_0+n_1+1}{12n_0n_1}=\frac{n+1}{12p(1-p)n^2}\geq\frac{n+1}{3n^2}>\frac{1}{3n},$$

where we use $p(1-p)\leq 1/4$. It is well known that the variance of the estimated error in this testing scenario is bounded according to $\sigma^2_{\text{est-ERR}}\leq\frac{1}{4n}$ (Devroye *et al.*, 1996), which is smaller than $\frac{1}{3n}$.

The same argument applies to k-fold CV and many other resampling-based error estimation schemes, as long as the data partitioning is stratified so that the testing sample points are represented in the same proportion as the training data for every fold/resampling. For non-trivial distributions, where $\Psi_S(X|Y=0)\neq\Psi_S(X|Y=1)$, the distribution of the rank-sum is generally unknown. Moreover, the variance of the true AUC and true error rate, which are functions of sample size and classification rule, are rarely known. Hence, we depend on simulation.

4.3 Results for microarray data

We present two sets of experiments based on real microarray datasets. In the first experiment, we apply a hold-out-based

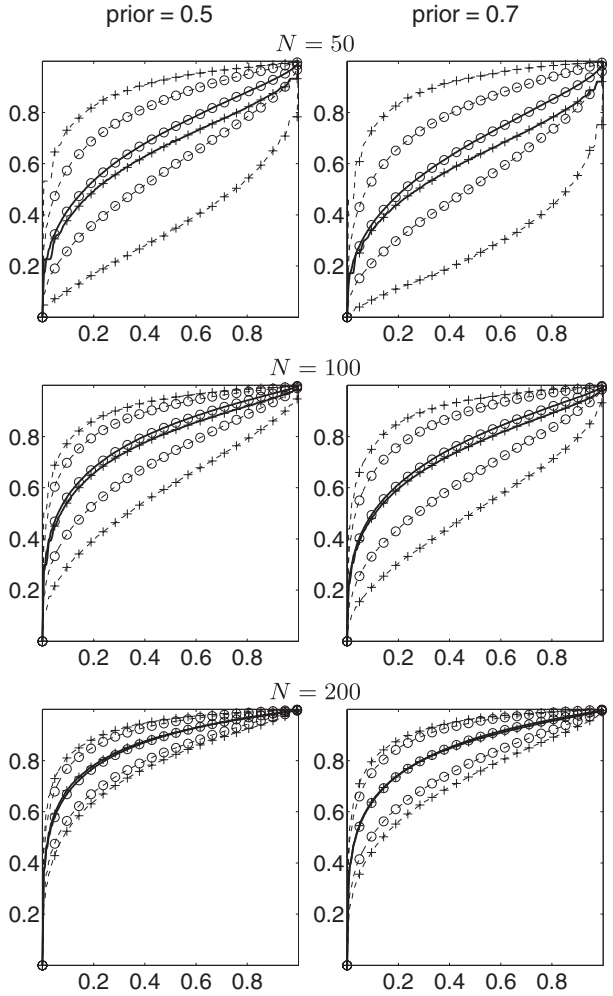

Fig. 4. Comparison of ROC plots: non-linear uncorrelated model, LDA classifier, LOO. The x-axis is the FPR and the y-axis is the TPR. The curves with cross marks and curves with circle mark correspond to using LOO error estimation and the true error, respectively. The solid lines are the mean ROC curves and the upper and lower dashed lines are the 95% confidence interval.

scheme on two existing cancer datasets to compare the true and estimated metrics and confirm the conclusions drawn from artificial data simulations. In the second experiment, we reproduce the experiments in two publications to verify the imprecise estimation observable in ROC plots.

In the first set of experiments, we use microarray data from two published sources: breast cancer (van de Vijver *et al.*, 2002) and lung cancer (Bhattacharjee *et al.*, 2001) studies. The breast cancer dataset includes 295 patients, 115 belonging to the good-prognosis class and 180 belonging to the poor-prognosis class, with prior probabilities 0.39 and 0.61, respectively. The lung cancer dataset contains 203 tumor samples, 139 being adenocarcinoma and 64 being of some other type of tumor, the prior probabilities being 0.68 and 0.32, respectively. We have reduced the two datasets to a selection of the 2000 genes with highest variance. For each iteration of our procedure, the datasets are divided into a training set and a test set. The training set is formed by 50 examples drawn without replacement from the dataset. 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 dataset, 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 dataset and 12 examples for the lung cancer dataset. The dependence among the samples is, therefore, not expected to have a large impact on the results (see Braga-Neto and Dougherty, 2004, for a discussion of this issue). We apply on these data the same protocol used for artificial data and detailed in Section 3. The results allow us to compare the true values of the metrics to the estimated metrics. Figure 5 shows the scatter plots for the breast and lung cancer microarray dataset with the linear SVM. As for the synthetic data, there is very little regression, wide dispersion, and the AUC dispersion is comparable or even greater than that for the classifier error. The RMS of the lung cancer dataset is smaller and there is less variance because the classification is easier (e.g. the error is low and the AUC is high). Table 1 gives the RMS values and the correlation coefficients, where the latter are very small.

In the second set of experiments, we demonstrate the large variance observable in ROC plots in real data cases. We use data in two published studies and repeat the same classification scheme multiple times with the only variance being the randomness in data partitioning.

The first dataset is from a study on the prediction of Parkinson's disease based on gene expression from blood samples (Scherzer *et al.*, 2007). The authors have identified a set of biomarkers, constructed a classifier and validated their results with an ROC analysis of the classifier. The dataset contains 105 subjects, 50 at early stages of Parkinson's disease, 22 healthy and 33 having another brain disease. The classification task is to detect only the Parkinson's disease, so the prior of the problem is 0.52. The authors use an algorithm that selects the best genes based on the Pearson's correlation between their expression level and the class label. Then a template of each class is formed from the mean expression of the selected genes. The classification outcome is determined by the risk score, which is defined as its correlation with the Parkinson's disease template minus its correlation with the non-Parkinson disease template. To evaluate the performance of this classifier, the original dataset is randomly divided into a training set (66 examples) and a test set (39 examples). The training set is used for gene selection and classifier construction, and the test set for evaluation. The authors support the validity of the identified biomakers using the classifier ROC curves. These ROC curves are computed using both the test set and an LOO procedure.

In our experiment, we apply the same scheme of gene selection, classification and evaluation as in the original study; however, we run this procedure 100 times to estimate the variance of the results. Figure 6 shows the average curve for the 100 ROC curves computed by LOO (solid line with cross marks) and the test set (solid line with circle marks). The dashed lines represent the 95% confidence intervals. These represent a kind of internal confidence bounds relative to the sample. We see that the results of our experiments are more pessimistic than the ones in the original paper. The AUC for LOO is 0.561 with confidence interval [0.354; 0.766]. The AUC for the test set is 0.536 with confidence interval [0.345; 0.727]. Our ROC curves are much more closer to the axis $y=x$. The confidence intervals are wide and include the axis $y=x$. In these conditions it is not possible to validate the genes, identified by the methodology, as

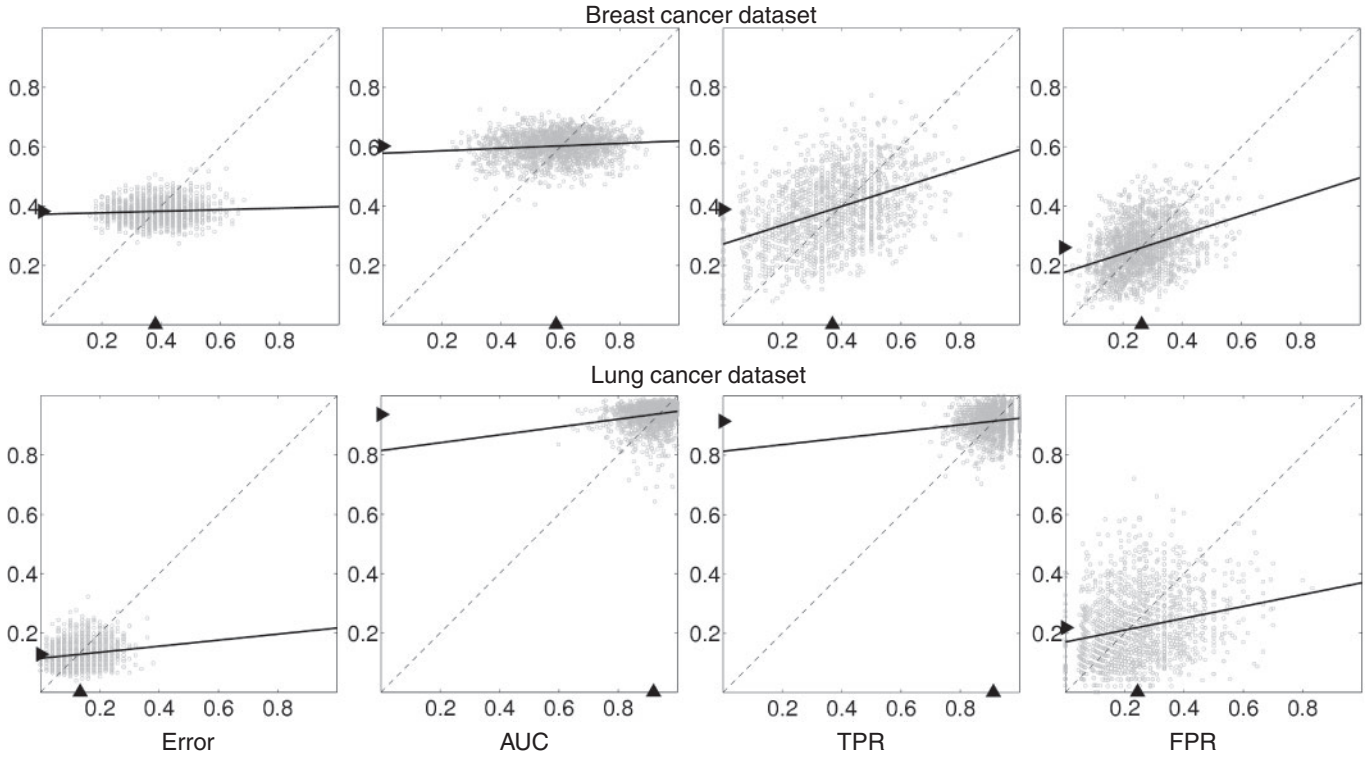

Fig. 5. The scatter plots of performance measures for breast and lung cancer dataset: linear SVM, 10CV. The *x*-axis is the estimated performance, whereas the *y*-axis is the true performance.

Table 1. The RMS and correlation of performance measures of breast and lung cancer datasets: linear SVM and 10CV

	Error	AUC	TPR	FPR
Breast cancer				
RMS	0.089	0.124	0.149	0.102
Correlation	0.070	0.114	0.421	0.369
Lung cancer				
RMS	0.065	0.064	0.063	0.153
Correlation	0.161	0.197	0.136	0.226

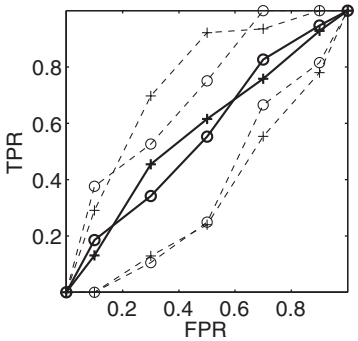

Fig. 6. ROC curves for the Parkinson's disease dataset. Lines with cross marks are the ROC curves by LOO and lines with circle marks are by the test set. The solid lines represent the average ROC, whereas the dashed lines the 95% confidence intervals.

predictors of Parkinson's disease. The situation is even worse than what is depicted here because the confidence intervals are internal to the sample. This only accounts for resampling variation, not the variance across samples, which would have to be included to obtain the full variance (Braga-Neto and Dougherty, 2004, 2005a).

The second dataset is from a study of the loss of phosphatase and tensin homolog (PTEN) associated with the presence of solid tumor. The authors develop and validate a microarray gene expression signature for immunohistochemistry (IHC) detectable PTEN loss in breast cancer (Saal *et al.*, 2007). The data contain 105 examples, 70 being IHC negative and 35 being IHC positive. The genes set is reduced to genes containing <20% of missing values, thereby resulting in 16 039 genes. The best discriminant genes are selected by a Mann–Whitney test and the classifier is constructed from a linear SVM. The performance of the classifier is estimated by 3-fold CV with 10 repetitions. The AUC of the final classifier is estimated to be 0.758.

We use the same procedure as described in the original study, except the feature size in classifier design, which was not clearly described in the original publication. In our experiment, the final feature size is fixed as 100. We compute the ROC curve and AUC of the obtained classifier. We also estimate the internal variance of the ROC curve and AUC via repetitions of the procedure. Figure 7 gives the average ROC curve (solid line) with its 95% confidence interval (dashed lines). This figure illustrates the large variance of the ROC curve, with the 95% confidence interval including the

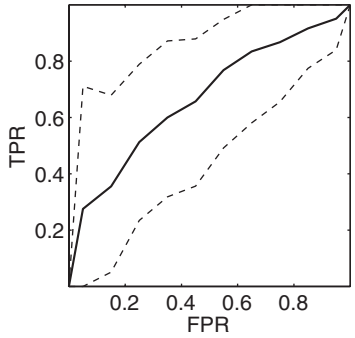

Fig. 7. ROC curve for the PTEN dataset. The solid line is the ROC curve by 3-fold CV with 10 repetitions and the dashed lines represent the 95% confidence intervals.

axis $y=x$. This means that, even if the ROC curve is above the axis $y=x$, we cannot reject the hypothesis that the classifier is meaningless. The AUC of the classifier is estimated to 0.642 and its 95% confidence interval is [0.483; 0.810]. Note that the AUC of the original published classifier (AUC = 0.758) is included in our confidence interval, as is a random guess (AUC = 0.5). As with the Parkinson's data, the situation is worse because the confidence internal only reflects internal variance from resampling within the sample, not the full variance. These experiments demonstrate that there are insufficient examples in the microarray datasets to draw ROC-based conclusions with acceptable precision.

4.4 Concluding remarks

This article and several preceding it have shown that even for synthetic data of a simple type, two Gaussian distributions, it is difficult to find good feature sets (Sima and Dougherty, 2006) and difficult to identify the error rate of a classifier composed of the features that one does find with a small sample. The essential reason is that one cannot make sufficiently good estimates of predictive error at any of the steps required to select features or to characterize the error of the classifier finally developed (Braga-Neto and Dougherty, 2004; Hanczar *et al.*, 2007). Here, we have demonstrated that small sample size leads to large inaccuracies in the estimated validation parameters associated with ROC analysis, even from well-behaved distributions.

A previous study (Saal *et al.*, 2007) applied permutation *P* tests to the AUC and obtained good *P*-values. There is no contradiction here because permutation *P* tests, when applied to classification, are essentially unrelated to classifier performance. Specifically, the *P*-values have virtually no regression with the error estimates (Hsing *et al.*, 2003).

A procedure for error estimation cannot provide more information than that exists in the distribution of samples with which it is presented: if that distribution is a poor estimate of the actual distribution, then the error estimate will be poor as well. In the case of actual biological samples, it can be seen that the differences between true and estimated error are even larger—considerably larger—than for a similarly small sample of well-distributed synthetic data.

ROC curves must be used with extreme caution unless one has a very large sample. In other cases, it would be nice to have some simple rule of thumb to determine if a sample is sufficiently large for the problem at hand; however, since in practice there is only a single sample available, no simple solution is possible. Nonetheless, an experimenter can take some precautions. First, one could use the kind of model-based analysis done in the present article. This would not reflect the actual population but it would allow the kind of confidence analysis demonstrated in Figure 4. This could be performed using a model developed for the specific technology being used or, lacking the availability of such a model, a Gaussian model like the one employed herein. It can be expected that the true biological population is less well behaved than the model so that the resulting confidence bounds could be taken as a performance floor for determining a sufficient sample size. A second approach would be to use the internal variance of the AUC if a resampling procedure has been employed, as in the PTEN example (see Appendix A for a description of the internal variance). Again, this would provide a floor because it only provides an estimate of the internal variance, not the full variance of the AUC. Neither method is perfect, but certainly if the 95% confidence interval contains the line $y=x$ in either case, then the sample size is insufficient. If forced to choose between the two approaches, we would choose the model-based approach because often the internal variance is much less than the full variance, so the resampling approach may be more optimistic. If these approaches are used, prudence would dictate utilizing both and making no conclusion unless the lower 95% confidence bound for the AUC exceeds 0.5 for both.

Funding: National Science Foundation (CCF-0634794, partially).

Conflict of Interest: none declared.

REFERENCES

Bhattacharjee,A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

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

Braga-Neto,U.M. and Dougherty,E.R. (2005a). Classification. In Dougherty,E.R. *et al.* (eds) *Genomic Signal Processing and Statistics*, *EURASIP Book Series on Signal Processing and Communication*. Hindawi Publishing Corporation, New York, NY.

Braga-Neto,U.M. and Dougherty,E.R. (2005b) Exact performance of error estimators for discrete classifiers. *Pattern Recogn.*, **38**, 1799–1814.

Devroye,L. *et al.* (1996) *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Glick,N. (1978) Additive estimators for probabilities of correct classification. *Pattern Recogn.*, **10**, 211–222.

Hanczar,B. *et al.* (2007) Decorrelation of the true and estimated classifier errors in high-dimensional settings. *EURASIP J. Bioinform. Syst. Biol.*, Article ID 38473.

Hand,D.J. and Till,R.J. (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.*, **45**, 171–186.

Hsing,T. *et al.* (2003) Relation between permutation-test p values and classifier error estimates. *Mach. Learn.*, **52**, 11–30.

Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.

Pepe,M.S. *et al.* (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J. Epidemiol.*, **159**, 882–890.

Saal,L.H. *et al.* (2007) Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl Acad. Sci. USA*, **104**, 7564–7569.

Scherzer,C.R. *et al.* (2007) Molecular markers of early parkinson's disease based on gene expression in blood. *Proc. Natl Acad. Sci. USA*, **104**, 955–960.

Sima,C. and Dougherty,E.R. (2006) What should be expected from feature selection in small-sample settings. *Bioinformatics*, **22**, 2430–2436.

Spackman,K.A. (1989) Signal detection theory: Valuable tools for evaluating inductive learning. In San Mateo,C.M.K. (ed.) *Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishers, San Francisco.

van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

Zollanvari,A. *et al.* (2009)On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. *Pattern Recogn.*, **42**, 2705–2723.

APPENDIX A

The *internal variance* of a randomized error estimator, such as k-fold CV, is the variance of the estimator given the sample, namely, the variance due only to its internal random factors (Braga-Neto and Dougherty, 2004, 2005a). It is expressed as $\mathrm{Var}_{\mathrm{int}} = \mathrm{Var}(\hat{\epsilon}|S)$, where $\hat{\epsilon}$ is a randomized error estimator and S is the sample. This variance is zero for non-randomized error estimators. The full variance, $\mathrm{Var}(\hat{\epsilon})$, of the error estimator is the one we are really concerned about, since it takes into account the uncertainty introduced by randomly sampling the data from the population. Using the well-known conditional-variance formula,

$$\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}(E[X|Y]),$$

we can break down $\mathrm{Var}(\hat{\epsilon})$ in the following way:

$$\mathrm{Var}(\hat{\epsilon}) = E[\mathrm{Var}_{\mathrm{int}}] + \mathrm{Var}(E[\hat{\epsilon}|S]).$$

The second term on the right-hand side is the one that includes the variability due to random sampling.