BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 12 2011, pages 1675–1683
doi:10.1093/bioinformatics/btr262

Data and text mining

Advance Access publication May 5, 2011

Multiple-rule bias in the comparison of classification rules

Mohammadmahdi R. Yousefi[1], Jianping Hua[2] and Edward R. Dougherty[1,2,3,*]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, [2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004 and [3]Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: There is growing discussion in the bioinformatics community concerning overoptimism of reported results. Two approaches contributing to overoptimism in classification are (i) the reporting of results on datasets for which a proposed classification rule performs well and (ii) the comparison of multiple classification rules on a single dataset that purports to show the advantage of a certain rule.

Results: This article provides a careful probabilistic analysis of the second issue and the ‘multiple-rule bias’, resulting from choosing a classification rule having minimum estimated error on the dataset. It quantifies this bias corresponding to estimating the expected true error of the classification rule possessing minimum estimated error and it characterizes the bias from estimating the true comparative advantage of the chosen classification rule relative to the others by the estimated comparative advantage on the dataset. The analysis is applied to both synthetic and real data using a number of classification rules and error estimators.

Availability: We have implemented in C code the synthetic data distribution model, classification rules, feature selection routines and error estimation methods. The code for multiple-rule analysis is implemented in MATLAB. The source code is available at http://gsp.tamu.edu/Publications/supplementary/yousefi11a/. Supplementary simulation results are also included.

Contact: edward@ece.tamu.edu

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

Received on February 19, 2011; revised on April 9, 2011; accepted on April 15, 2011

1 INTRODUCTION

Three recent articles in *Bioiformatics* have lamented the difficulty in establishing performance advantages for proposed classification rules (Boulesteix, 2010; Jelizarow *et al*., 2010; Rocke *et al*., 2009). Two statistically grounded sources of overoptimism have been highlighted. One considers applying a classification rule to numerous datasets and then reporting only the results on the dataset for which the designed classifier possesses the lowest estimated error. The optimistic bias from this kind of dataset picking is quantitatively analyzed in Yousefi *et al*. (2010), where it is termed ‘reporting bias’ and where this bias is characterized as a function of the number of considered datasets. A second kind of overoptimism concerns the comparison of a collection of classification rules by applying the classification rules to a dataset and comparing them according to the estimated errors of the designed classifiers. This kind of bias, which we will call ‘multiple-rule bias’, has been considered in Boulesteix and Strobl (2009) by applying a battery of classification rules to colon cancer and prostate cancer datasets and then examining the effects of choosing classification rules having minimum cross-validation error estimates.

Whereas the thrust of Boulesteix and Strobl (2009) is to compare the sources of multiple-rule bias in classification rules, namely, gene selection, parameter selection and classifier function construction, our interest is in studying multiple-rule bias as a function of the number of rules being considered. In particular, we are interested in the joint distribution, as a function of the number of compared rules, between the minimum estimated error among a collection of classification rules and the true error for the classification rule having minimum estimated error, as well as certain moments associated with this joint distribution. Although different with regard to distributional specifics, this approach is analogous to the approach taken in Yousefi *et al*. (2010), where the joint distribution involved the minimum estimated error of the designed classifier over a collection of datasets and the true error of the designed classifier on the population corresponding to the dataset resulting in minimum estimated error. This is a natural way to proceed because any bias ultimately results from inaccuracy in error estimation, so that the behavior of the joint distribution of interest and its moments are consequent to the joint distribution of the error estimator and the true error. Owing to the methodology in Boulesteix and Strobl (2009), it would have been impossible for them to study this joint distribution because they never concern themselves with true errors, only cross-validation estimates. Hence, when they compare a minimal error to a baseline error to arrive at a measure of optimistic bias, they are comparing cross-validation estimates.

In characterizing multiple-rule bias, we begin with a more general framework than the one just described; rather than simply considering multiple classification rules, we consider multiple classifier rule models, so that there are not only multiple classification rules, but also multiple error estimation rules being employed. We define a classifier rule model as a pair (Ψ, Ξ), where Ψ is a classification rule, including feature selection if feature selection is employed, and Ξ is an error estimation rule (Dougherty and Braga-Neto, 2006). The scenario in the preceding paragraph results where there is only a single error estimation rule.

[*]To whom correspondence should be addressed.

2 SYSTEMS AND METHODS

We consider r classification rules, $\Psi_1, \Psi_2, \ldots, \Psi_r$, and s error estimation rules, $\Xi_1, \Xi_2, \ldots, \Xi_s$, on a feature-label distribution F. These are combined to form $m = rs$ classifier rule models: $(\Psi_1, \Xi_1), (\Psi_1, \Xi_2), \ldots, (\Psi_1, \Xi_s), (\Psi_2, \Xi_1), (\Psi_2, \Xi_2), \ldots, (\Psi_r, \Xi_s)$. Given a random sample $\mathcal{S}_n$ of size n drawn from F, the classification rules yield r designed classifiers: $\psi_i = \Psi_i(\mathcal{S}_n), i = 1, 2, \ldots, r$. The true error of ψ_i is given by $\varepsilon^i_{\text{true}} = P(\psi_i(\mathbf{X}) \neq Y)$, where $(\mathbf{X}, Y)$ is a feature-label pair. For $j = 1, 2, \ldots, s$, Ξ_j provides an error estimate, $\varepsilon^{i,j}_{\text{est}}$, for ψ_i. Since the classification rules are not identical, neither are the distributions of $\varepsilon^1_{\text{true}}, \varepsilon^2_{\text{true}}, \ldots, \varepsilon^r_{\text{true}}$ nor are the distributions of $\varepsilon^{i,1}_{\text{est}}, \varepsilon^{i,2}_{\text{est}}, \ldots, \varepsilon^{i,s}_{\text{est}}$. All true-error and estimated-error distributions are functions of the random sample $\mathcal{S}_n$. Since all classification rules operate on the same sample, the true errors can be highly correlated, as will be the estimated errors. Without loss of generality, we assume the classifier models are enumerated so that $E_{\mathcal{S}_n}[\varepsilon^1_{\text{true}}] \leq E_{\mathcal{S}_n}[\varepsilon^2_{\text{true}}] \leq \ldots \leq E_{\mathcal{S}_n}[\varepsilon^r_{\text{true}}]$.

The minimum estimated error is

$$\varepsilon^{\min}_{\text{est}} = \min\{\varepsilon^{1,1}_{\text{est}}, \varepsilon^{1,2}_{\text{est}}, \ldots, \varepsilon^{1,s}_{\text{est}}, \varepsilon^{2,1}_{\text{est}}, \ldots, \varepsilon^{r,s}_{\text{est}}\}. \tag{1}$$

Letting $i_{\min}$ and $j_{\min}$ denote the classifier number and error estimator number, respectively, for which the error estimate is minimum, we have $\varepsilon^{\min}_{\text{est}} = \varepsilon^{i_{\min}, j_{\min}}_{\text{est}}$.

Suppose a researcher wishes to select the best performing of r classification rules on a feature-label distribution F and proceeds by taking a random sample from F, designing a classifier for each classification rule, and estimating the errors of the designed classifiers. If F is known, then the true error of each designed classifier can be evaluated, the classifier with minimum true error can be chosen, and the classification rule leading to that classifier be declared ‘ best’. The truly best classification rule has the minimum expected error across all samples from F, so that what is happening is that a single observation of $\varepsilon^i_{\text{true}}$ is being used as an estimate for $E_{\mathcal{S}_n}[\varepsilon^i_{\text{true}}]$. On the other hand, if F is unknown as is used in practice, then the errors of the designed classifiers are estimated from sample data and the classification rule leading to the classifier, $\psi_{i_{\min}}$, with minimum estimated error is chosen as ‘best’. We are assuming that the researcher tries s error estimators and selects the one with lowest error estimate. In this case, a single observation of $\varepsilon^{\min}_{\text{est}}$ is being used to estimate $E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}]$, the basic performance measure for $\Psi_{i_{\min}}$. At issue is the goodness of this estimation. This involves the distribution of the deviation $\Delta = \varepsilon^{\min}_{\text{est}} - E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}]$, which is marginal to the joint distribution of $(\varepsilon^{\min}_{\text{est}}, \varepsilon^{i_{\min}}_{\text{true}})$.

A key performance measure derived from the deviation distribution is the bias of $\varepsilon^{\min}_{\text{est}}$ as an estimator of $E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}]$, namely,

$$\text{Bias}(m,n) = E_{\mathcal{S}_n}[\Delta] = E_{\mathcal{S}_n}\left[\varepsilon^{\min}_{\text{est}}\right] - E_{\mathcal{S}_n}\left[\varepsilon^{i_{\min}}_{\text{true}}\right]. \tag{2}$$

Estimation is optimistic if $\text{Bias}(m,n) < 0$. This can happen even if none of the error estimation rules are optimistically biased, that is, even if $E_{\mathcal{S}_n}[\varepsilon^{i,j}_{\text{est}}] \geq E_{\mathcal{S}_n}\left[\varepsilon^i_{\text{true}}\right]$ for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, s$. Indeed, even if this is so, owing to estimation-rule variance, on any given sample it may be that $\varepsilon^{\min}_{\text{est}} < \varepsilon^1_{\text{true}}$. For instance, if among $\Xi_1, \Xi_2, \ldots, \Xi_s$ there is an error estimation rule, such as leave-one-out cross-validation, that is slightly (pessimistically) biased and possesses large variance, then we should expect that $E_{\mathcal{S}_n}\left[\varepsilon^{\min}_{\text{est}}\right] < E_{\mathcal{S}_n}\left[\varepsilon^1_{\text{true}}\right]$. In this case, $E_{\mathcal{S}_n}\left[\varepsilon^{\min}_{\text{est}}\right] < E_{\mathcal{S}_n}\left[\varepsilon^1_{\text{true}}\right] \leq E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}]$ and $\text{Bias}(m,n) < 0$. In the way we have set up the general problem, not only can optimistic bias result from considering multiple estimated errors among classifiers but also from applying multiple error estimates for each classifier.

From the generic arguments made thus far, we can state two properties concerning the bias. First, as the number m of classifier models grows, the minimum of Equation (1) is taken over more estimated errors, thereby increasing $\varepsilon^1_{\text{true}} - \varepsilon^{\min}_{\text{est}}$ and $|\text{Bias}(m,n)| = E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}] - E_{\mathcal{S}_n}[\varepsilon^{\min}_{\text{est}}]$. Second, as the sample size n is increased, under the typical condition that the variance of the error estimator decreases with increasing sample size, $E_{\mathcal{S}_n}[\varepsilon^{\min}_{\text{est}}]$ increases and $|\text{Bias}(m,n)|$ decreases.

Bias is only one factor in estimating $E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}]$ by $\varepsilon^{\min}_{\text{est}}$. Another is the variance of $\varepsilon^{\min}_{\text{est}}$. In fact, one should consider the *root-mean-square (RMS) error* for $\varepsilon^{\min}_{\text{est}}$ as an estimator of $E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}]$, which is given by

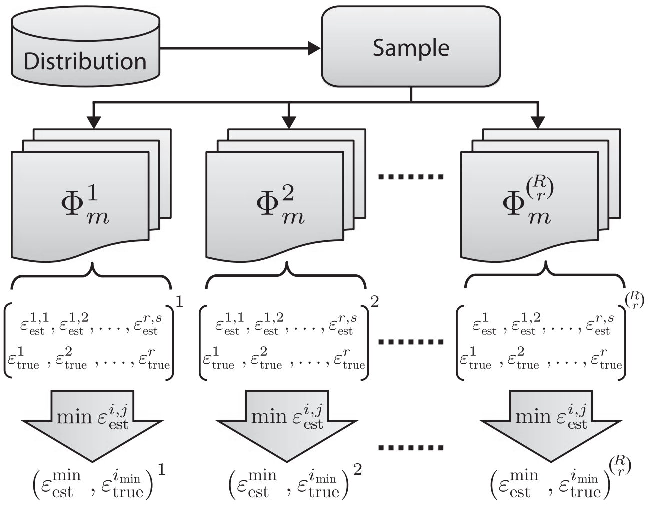
Distribution
Sample
Φ_m^1
Φ_m^2
$\Phi_m^{\binom{R}{r}}$
$\begin{pmatrix}\varepsilon^{1,1}_{\text{est}}, \varepsilon^{1,2}_{\text{est}}, \ldots, \varepsilon^{r,s}_{\text{est}} \\ \varepsilon^{1}_{\text{true}}, \varepsilon^{2}_{\text{true}}, \ldots, \varepsilon^{r}_{\text{true}}\end{pmatrix}^1$
$\begin{pmatrix}\varepsilon^{1,1}_{\text{est}}, \varepsilon^{1,2}_{\text{est}}, \ldots, \varepsilon^{r,s}_{\text{est}} \\ \varepsilon^{1}_{\text{true}}, \varepsilon^{2}_{\text{true}}, \ldots, \varepsilon^{r}_{\text{true}}\end{pmatrix}^2$
$\begin{pmatrix}\varepsilon^{1}_{\text{est}}, \varepsilon^{1,2}_{\text{est}}, \ldots, \varepsilon^{r,s}_{\text{est}} \\ \varepsilon^{1}_{\text{true}}, \varepsilon^{2}_{\text{true}}, \ldots, \varepsilon^{r}_{\text{true}}\end{pmatrix}^{\binom{R}{r}}$
$\min \varepsilon^{i,j}_{\text{est}}$
$\min \varepsilon^{i,j}_{\text{est}}$
$\min \varepsilon^{i,j}_{\text{est}}$
$\left(\varepsilon^{\min}_{\text{est}}, \varepsilon^{i_{\min}}_{\text{true}}\right)^1$
$\left(\varepsilon^{\min}_{\text{est}}, \varepsilon^{i_{\min}}_{\text{true}}\right)^2$
$\left(\varepsilon^{\min}_{\text{est}}, \varepsilon^{i_{\min}}_{\text{true}}\right)^{\binom{R}{r}}$

Fig. 1. Multiple-rule testing procedure on a single sample.

$$\text{RMS}(m,n) = \sqrt{E_{\mathcal{S}_n}[\Delta^2]} = \sqrt{\text{Bias}(m,n)^2 + \text{Var}_{\mathcal{S}_n}[\Delta]}. \tag{3}$$

Even should the bias be small, the estimation will not be accurate if the deviation variance is large.

As discussed thus far, the classification rules are fixed; however, since our interest is in the number of classification rules (and error estimators), not any particular rule or estimator, we assume there is a total of R classification rules from which we randomly choose r. This corresponds to a situation where a researcher selects r from among a large collection (R) of potential classification rules and applies s error estimators to each selected rule. Given r, there are $\binom{R}{r}$ possible collections of classification rules to be combined with s error estimators to form $\binom{R}{r}$ possible collections of classifier rule models of size m. Hence, the number m of classifier rule models increments according to $s, 2s, \ldots, Rs$. We denote the collection of classifier models of size m by Φ_m. To get the distribution of errors, one needs to generate independent samples from the same feature-label distribution and apply the procedure shown in Figure 1.

The previously discussed performance measures must be adjusted to take model randomization into account. Given a sample $\mathcal{S}_n$, for a realization of Φ_m we find an expected deviation according to Equation (2), but now we have a random process generating the realizations so we have to take the expectation over that process to obtain the expected bias,

$$\text{Bias}_{\text{Av}}(m,n) = E_{\Phi_m}\left[E_{\mathcal{S}_n}[\Delta|\Phi_m]\right], \tag{4}$$

where the subscript indicates the average over Φ_m. A similar averaging arises with the deviation variance to yield $\text{Var}_{\text{Av}}(m,n) = E_{\Phi_m}[\text{Var}_{\mathcal{S}_n}[\Delta|\Phi_m]]$. The RMS now takes the form

$$\text{RMS}_{\text{Av}}(m,n) = E_{\Phi_m}\left[\sqrt{E_{\mathcal{S}_n}[\Delta^2|\Phi_m]}\right]. \tag{5}$$

Having discussed the performance measures relating to estimating $E_{\mathcal{S}_n}[\varepsilon^{i_{\min}}_{\text{true}}]$ by $\varepsilon^{\min}_{\text{est}}$, we now consider the comparative advantage of the classification rule $\Psi_{i_{\min}}$. Its *true comparative advantage* with respect to the other considered classification rules is

$$A_{\text{true}} = E_{\mathcal{S}_n}\left[\varepsilon^{i_{\min}}_{\text{true}}\right] - \frac{1}{r-1}\sum_{i \neq i_{\min}} E_{\mathcal{S}_n}\left[\varepsilon^i_{\text{true}}\right]. \tag{6}$$

Its *estimated comparative advantage* is given by

$$A_{\text{est}} = \varepsilon^{\min}_{\text{est}} - \frac{1}{r-1}\sum_{i \neq i_{\min}} \varepsilon^{i, j_{\min}}_{\text{est}}, \tag{7}$$

where we use the error estimator associated with the pair $(i_{\min}, j_{\min})$ for which the minimum estimated error is obtained (assuming that this would be the error estimator chosen by a researcher for the sake of consistency). The expectation, $E_{\mathcal{S}_n}[A_{\text{est}}]$, of this distribution gives the *mean estimated comparative advantage* of $\Psi_{i_{\min}}$ with respect to the collection. A key bias to be considered is $C_{\text{bias}} = E_{\mathcal{S}_n}[A_{\text{est}}] - A_{\text{true}}$ because it measures over optimism in comparative performance. In the case of randomization, the true comparative advantage becomes $E_{\Phi_m}[A_{\text{true}}|\Phi_m]$ and the mean estimated comparative advantage becomes $E_{\Phi_m}[E_{\mathcal{S}_n}[A_{\text{est}}|\Phi_m]]$.

2.1 Simulation design

We use a general model based on multivariate Gaussian distributions with a blocked covariance structure that conforms to various observations made in microarray expression-based studies (Hua *et al.*, 2009). A battery of distribution models can be constructed by changing model parameters to generate different synthetic data samples. We also consider four real datasets.

2.1.1 Synthetic data In microarray studies, assuming a blocked covariance matrix is a way of modeling groups of interacting genes where there is negligible interaction between the groups. It has been used in genomic classification to model genes collected into distinct pathways, each pathway being represented by a block (Dougherty *et al.*, 2007; Shmulevich and Dougherty, 2007). As explained in Hua *et al.* (2009), although the model does not embrace all details of the experimental procedures, it is general enough to include major aspects and various complexity levels suitable for simulation of real-world scenarios. Sample points are taken from two equally likely classes, C_0 and C_1, having D features. Furthermore, by putting c equally likely subclasses in C_1, each having its own distribution, one can model cases like different stages or subtypes of a cancer. Each sample point in C_1 belongs to one and only one of these subclasses.

Features are categorized into two major groups, markers and non-markers. Markers resemble genes causing disease or susceptibility to disease. The groups have different class-conditional distributions for the two classes. They can be further categorized into two different types: global and heterogeneous markers. Global markers are homogeneously distributed among the two classes with D_{gm}-dimensional Gaussian distributions and parameters $(\mu_0^{\text{gm}}, \Sigma_0^{\text{gm}})$ for class 0 and $(\mu_1^{\text{gm}}, \Sigma_1^{\text{gm}})$ for class 1, where D_{gm} is the total number of global markers. Heterogeneous markers are divided into c subclasses within class 1. Each subclass is associated with D_{hm} mutually exclusive heterogeneous markers having D_{hm}-dimensional Gaussian distributions with parameters $(\mu_1^{\text{hm}}, \Sigma_1^{\text{hm}})$. The sample points not belonging to this particular subclass are considered to have D_{hm}-dimensional Gaussian class-conditional distributions with parameters $(\mu_0^{\text{hm}}, \Sigma_0^{\text{hm}})$.

Assuming that global and heterogeneous markers possess identical covariance structures, we use $\{\Sigma_0, \Sigma_1\}$ instead of $\{\Sigma_0^{\text{gm}}, \Sigma_1^{\text{gm}}\}$ and $\{\Sigma_0^{\text{hm}}, \Sigma_1^{\text{hm}}\}$. We assume that $\Sigma_0 = \sigma_0^2 \Sigma$ and $\Sigma_1 = \sigma_1^2 \Sigma$, where σ_0^2 and σ_1^2 can be different, and that Σ has the following block structure:

$$\Sigma = \begin{bmatrix} \Sigma_\rho & 0 & 0 & \dots & 0 \\ 0 & \Sigma_\rho & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \Sigma_\rho \end{bmatrix},$$

where Σ_ρ is a $l \times l$ matrix, with 1 on the diagonal and ρ off the diagonal. In the block-based covariance structure, the markers are divided into equal-size blocks of size l. Markers of different blocks are uncorrelated, while all the markers in the same block are correlated to each other with correlation coefficient ρ.

A consequence of having unequal variances and subclasses in the class-conditional distributions is to introduce non-linearities in the decision boundaries for the model, where less global markers and larger difference in the variances lead to a more non-linear decision boundary. Because the global markers and the heterogeneous markers possess the same structure, we can assume the same mean vectors $\{\mu_0, \mu_1\}$ for both groups, $\{\mu_0^{\text{gm}}, \mu_1^{\text{gm}}\}$ and $\{\mu_0^{\text{hm}}, \mu_1^{\text{hm}}\}$, as we did for the covariance matrices. Furthermore, we use the same structure for μ_0 and μ_1 in the form of $m_0 \times (1,1,\dots,1)$ and $m_1 \times (1,1,\dots,1)$, respectively, where m_0 and m_1 are scalars (Hua *et al.*, 2009; Yousefi *et al.*, 2010).

Table 1. Distribution model parameters

Parameters	Values/description
Mean	$m_0 = 0.23, m_1 = 0.8$ (equal variance) $m_0 = 0.11, m_1 = 0.9$ (unequal variance)
Variances	$\sigma_0^2 = 0.6^2, \sigma_1^2 = 0.6^2$ (equal variance) $\sigma_0^2 = 0.6^2, \sigma_1^2 = 1.2^2$ (unequal variance)
Block size	$l = 5$
Features	$D = 20000$
Feature block correlation	$\rho = 0.8$
Subclasses	$c = 2$
Global markers	$D_{gm} = 20$
Heterogeneous markers	$D_{hm} = 50$
High-variance non-markers	$D_{hv} = 2000$
Low-variance non-markers	$D_{lv} = 17880$

Table 2. A summary of the real datasets used in this study

Dataset	Dataset type	Feature|sample size
Yeoh *et al.* (2002)	Pediatric ALL	5077|149/99
Zhan *et al.* (2006)	Multiple myeloma	54613|156/78
Chen *et al.* (2004)	HCC	10237|75/82
Natsoulis *et al.* (2005)	Drugs response on rats	8491|120/61

Similar to the global markers, there are two types of non-markers: high-variance and low-variance non-markers. The D_{hv} features belonging to the former group are uncorrelated and their distributions are described by $pN(m_0, \sigma_0^2) + (1-p)N(m_1, \sigma_1^2)$, where m_0, m_1, σ_0^2 and σ_1^2 take values equal to the means and variances of the markers, respectively, and p is a random value uniformly distributed over $[0,1]$. The D_{lv} remaining features are uncorrelated low-variance non-markers, each having a Gaussian distribution with parameters (m_0, σ_0^2).

A typical microarray experiment usually contains tens of thousands of probes (genes) but a small number of sample points, typically less than 200. We choose the total number of features to be $D = 20000$ and the number of sample points to be 60 and 120. Two variance settings are considered: equal variances $\{\sigma_0^2 = 0.6^2, \sigma_1^2 = 0.6^2\}$ and unequal variances $\{\sigma_0^2 = 0.6^2, \sigma_1^2 = 1.2^2\}$. For the blocked covariance matrix, we choose block size $l = 5$ and correlation coefficient $\rho = 0.8$, giving relatively tight correlation within a block, which would be expected for a pathway. We do not choose model parameters in accordance with the Bayes errors or the estimated errors; rather, we choose them in accordance with achievable true errors seen in real problems. Table 1 shows the parameters of the distribution models used in this study. See Yousefi *et al.* (2010) for more details about the choice of parameters.

2.1.2 Real data This study uses four real datasets from microarray experiments consisting of more than 150 arrays: pediatric acute lymphoblastic leukemia (ALL) (Yeoh *et al.*, 2002), multiple myeloma (Zhan *et al.*, 2006), hepatocellular carcinoma (HCC) (Chen *et al.*, 2004) and drugs and toxicants response on rats dataset (Natsoulis *et al.*, 2005). We use $n = 60$ sample points for training. The remaining sample points are held-out for computing the true error. To the extent possible, we try to maintain the original labeling and follow the data preparation directions used in the papers reporting these datasets. Table 2 shows a summary of the four real datasets. Full descriptions are presented in the Supplementary Materials.

Table 3. Classifier rule models considered in this study

Classification rule	Feature selection	Error estimation
3NN	t-test	5-fold cross-validation
LDA	t-test + SFS	10-fold cross-validation
DLDA		LOO
NMC		
L-SVM		
RBF-SVM		

2.2 Classification schemes

Six classification rules are considered: 3-nearest neighbors (3NN), linear discriminant analysis (LDA), diagonal LDA (DLDA), nearest-mean classifier (NMC), linear support vector machine (L-SVM) and radial basis function SVM (RBF-SVM). Two different feature selection methods are considered: t-test and t-test followed by sequential forward search (t-test + SFS). One-stage feature selection uses the t-test and five features are selected. For two-stage feature selection, the number of features is reduced to 500 in the first stage (t-test) and then to 5 by SFS. Three cross-validation error estimation methods are considered: 5-fold, 10-fold and leave-one-out (LOO). Combining six classification rules with two feature selection methods results in $R=12$ classification rules from which we randomly choose $r=1,2,\dots,R$ and design the classifiers on one sample. Note that with three error estimators, there is a maximum of 36 different classifier rule models. Table 3 lists the classification rules, feature selection methods and error estimation procedures utilized.

To illustrate the definitions, let us suppose we are only considering two classification rules, LDA and 3NN, without feature selection and two error estimators, LOO and CV5 (5-fold cross-validation). LDA is based on the discriminant for the optimal classifier in a Gaussian model (Gaussian class-conditional densities) with common covariance matrix by plugging the sample means and pooled sample covariance matrix obtained from the data into the discriminant. Assuming equally likely classes, LDA assigns $\mathbf{x}$ to class 1 if and only if

$$(\mathbf{x}-\bar{x}_1)^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\bar{x}_1)\leq(\mathbf{x}-\bar{x}_0)^T\widehat{\Sigma}^{-1}(\mathbf{x}-\bar{x}_0), \tag{8}$$

where $\bar{x}_u$ is the sample mean for class u, $u=0,1$, and $\widehat{\Sigma}$ is the pooled sample covariance matrix. While derived under the Gaussian assumption with common covariance matrix, LDA can provide good results when these assumptions are mildly violated. For the 3NN rule, the designed classifier is defined to be 0 or 1 at $\mathbf{x}$ according to which is the majority among the labels of the 3 points closest to $\mathbf{x}$.

In k-fold cross-validation, the given sample $\mathcal{S}_n$ is randomly partitioned into k folds (subsets) $\mathcal{S}_{(i)}$, for $i=1,2,\dots,k$. Each fold is left out of the design process, a (surrogate) classifier, $\psi_{n,i}$, is designed on $\mathcal{S}_n-\mathcal{S}_{(i)}$, the error of $\psi_{n,i}$ is estimated as the counting error $\psi_{n,i}$ makes on $\mathcal{S}_{(i)}$, and the cross-validation estimate is the average error committed on all folds. If there is feature selection, then it must be redone for every fold because it is part of the classification rule. In leave-one-out cross-validation, each fold consists of a single point. Owing to computational requirements, k-fold cross-validation, $k<n$, usually involves a random selection of partitions. In general, the bias of cross-validation is typically slightly pessimistic, provided that the number of folds is not too small. The problem with cross-validation is that it tends to be inaccurate for small samples because, for these, it has large variance (Braga-Neto and Dougherty, 2004) and is poorly correlated with the true error (Hanczar *et al*., 2007). For all error estimators, there is variation resulting from the sampling process. For randomized cross-validation, a second variance contribution, referred to as 'internal variance', arises from the random selection of the partitions (see Hanczar and Dougherty, 2010). The latter does not apply to LOO because only a single set of folds is possible.

Relative to the definitions in the Section 2, if we let Ψ_1 be LDA, Ψ_2 be 3NN, Ξ_1 be LOO and Ξ_2 be CV5, then there are four classifier rule models: (LDA, LOO), (LDA, CV5), (3NN, LOO), (3NN, CV5). There are two true errors, $\varepsilon_{\text{true}}^{\text{LDA}}$ and $\varepsilon_{\text{true}}^{\text{3NN}}$, four estimated errors, $\varepsilon_{\text{est}}^{\text{LDA,LOO}}$, $\varepsilon_{\text{est}}^{\text{LDA,CV5}}$, $\varepsilon_{\text{est}}^{\text{3NN,LOO}}$ and $\varepsilon_{\text{est}}^{\text{3NN,CV5}}$, and $\varepsilon_{\text{est}}^{\min}$ is the minimum of the four estimated errors.

2.3 Implementation

The raw output of the synthetic data simulation consists of the true and estimated error pairs resulting from applying the 36 different classification schemes on 10000 independent random samples drawn from the aforementioned four different distribution models. We approximate the expected true error by taking the average of true errors of each classification rule over the samples. We generate all $\binom{R}{r}$ possible collections of classification rules of size r, each associated with s error estimation rules, resulting in $\binom{R}{r}$ collections of classifier rule models of size m. For each collection, we find the true and estimated error pairs from the raw output data. Then, for each sample, we find the classifier model (including the classification and error estimation rules) in the collection that gives the minimum estimated error. We record the estimated error, its corresponding true error and the classification and error estimation rules. Given the collection, Φ_m, we compute Δ and A_{est} for each sample. Then, we approximate $E_{\mathcal{S}_n}[A_{\text{est}}|\Phi_m]$, $\text{Var}_{\mathcal{S}_n}[\Delta|\Phi_m]$, $\sqrt{E_{\mathcal{S}_n}[\Delta^2|\Phi_m]}$, $E_{\mathcal{S}_n}[\Delta|\Phi_m]$ and $A_{\text{true}}|\Phi_m$, by taking the average over all the samples. Finally, we approximate $E_{\Phi_m}[E_{\mathcal{S}_n}[\Delta|\Phi_m]]$, $E_{\Phi_m}[\text{Var}_{\mathcal{S}_n}[\Delta|\Phi_m]]$, $E_{\Phi_m}[\sqrt{E_{\mathcal{S}_n}[\Delta^2|\Phi_m]}]$, $E_{\Phi_m}[E_{\mathcal{S}_n}[A_{\text{est}}|\Phi_m]]$ and $E_{\Phi_m}[A_{\text{true}}|\Phi_m]$ by taking the average over the generated collections.

Real data simulations differ somewhat from the synthetic data in the way that the true and estimated errors are computed. For the synthetic data, we generate 10000 pairs of samples (training set of size 60 or 120 and test set of size 5000) from the assumed distribution model. But for a real dataset, we randomly pick 60 sample points for training (classifier design and error estimation). The remaining held-out sample points are used to calculate the true error. We repeat this process 10000 times.

3 RESULTS AND DISCUSSION

The full set of results appears in the Supplementary Material. In the article, we provide representative examples for each issue. We consider two cases regarding the error estimators: multiple error estimators and a single error estimator. For multiple error estimators, $s=3$, we consider all three error estimators at once, and $m=3,6,\dots,36$. For a single error estimator, $s=1$, we have three difference cases, depending on which error estimator is used, and $m=1,2,\dots,12$ for each error estimator. For $s=3$, keep in mind that the simulations are incremented in steps of three, $3,6,\dots,36$, because each classification rule is evaluated with all three error estimators, as would be the case in practice if an investigator were to consider three error estimators. For a single error estimator, we show LOO in the article and leave the others to the Supplementary Material.

3.1 Joint distributions

For the synthetic data, the joint distributions are estimated with a bivariate Gaussian-kernel density estimation method. The first set of results (Supplementary Figures s1–s16) show joint distributions between the minimum estimated errors and their corresponding true errors, $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{i_{\min}}$, for multiple and single error estimators, different sample sizes and variances. Each plot includes the regression line and a small circle showing the pair of sample means. As m increases, the distributions tend to be more circular (indicating less correlation) and also more compact (indicating smaller variance). Furthermore, as m increases, the distributions

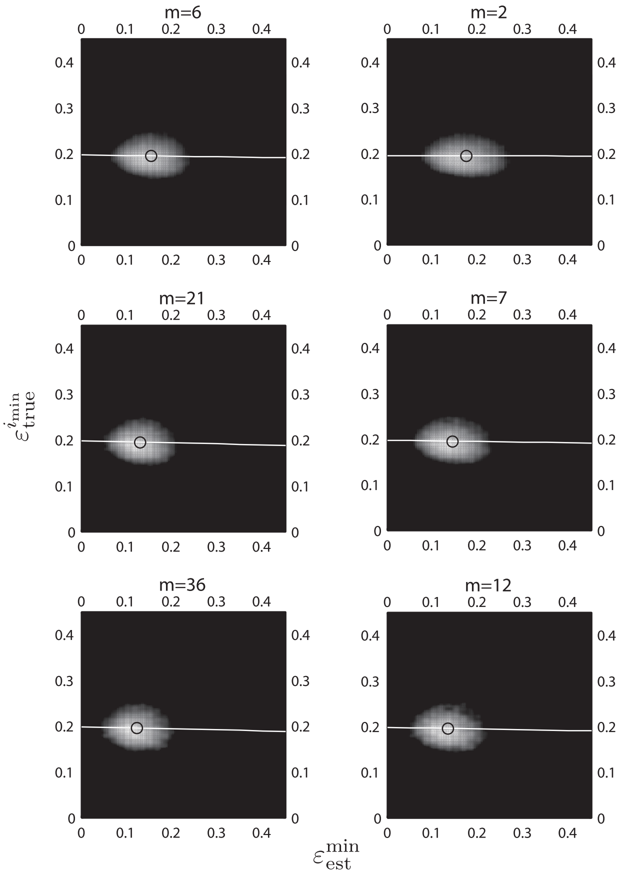

Fig. 2. The joint distributions between $\varepsilon_{\mathrm{est}}^{\min}$ and $\varepsilon_{\mathrm{true}}^{i_{\min}}$ with respect to the collection size m, for all classifier rule models for $m=6,21,36$ (left column) and for single LOO error estimation for $m=2,7,12$ (right column). The real dataset is multiple myeloma by Zhan *et al*. (2006). The white line shows the regression line and the circle indicates the sample mean of the joint distribution.

move to the left, thereby demonstrating greater bias. Note the smaller variation for sample size 120.

For the real data, the joint distributions are again estimated with a bivariate Gaussian-kernel density estimation method. Supplementary Figures s25–s40 show the joint distribution of $(\varepsilon_{\mathrm{est}}^{\min}, \varepsilon_{\mathrm{true}}^{i_{\min}})$ for multiple and single error estimators, and for different real datasets. Figure 2 shows the distributions and regression lines for the myeloma data: the left column is for multiple error estimators and shows $m=6,21,36$; the right column is for the single error estimator LOO and shows $m=2,7,12$. Similar to the synthetic data, as m increases, the distributions tend to be more circular, have smaller variances and move to the left. What is most striking is the absence of regression between $\varepsilon_{\mathrm{est}}^{\min}$ and $\varepsilon_{\mathrm{true}}^{i_{\min}}$. This lack of regression is consistent with what has been observed in other settings when cross-validation is used to estimate the true error (Hanczar *et al*., 2007, 2010).

3.2 Moments and comparative performance

For the synthetic data and the multiple error estimator case, Figure 3 shows: (a) the expected bias, $\mathrm{Bias}_{\mathrm{Av}}$; (b) the expected variance, $\mathrm{Var}_{\mathrm{Av}}$; (c) the expected RMS, $\mathrm{RMS}_{\mathrm{Av}}$; and (d) the expected comparative performance bias, C_{bias}. Note that Figure 3d does not graph $m=s$ because the comparative advantages are not defined when $r=1$. The same applies for analogous subfigures in the rest of the article. For increasing m, the bias and RMS get worse, but even with $m=3$, the RMS is about 0.1 for sample size 60. For this sample size and $m=36$, the comparative-performance bias has reached -0.1. Figure 4 shows corresponding results for a single error estimator, LOO. These too are especially alarming for $n=60$. Note that the RMS, actually, has a temporary small dip at $m=2$, which is a result of steep decline in variance between $m=1$ and $m=2$.

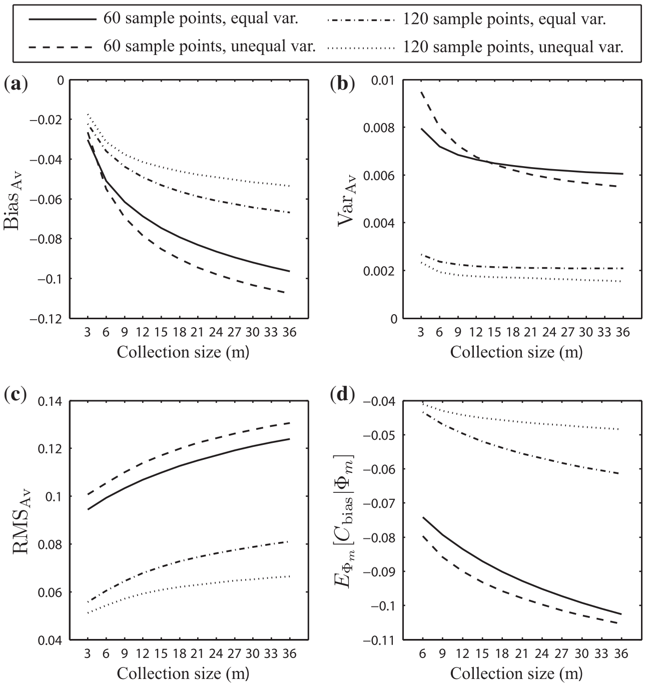

Fig. 3. (**a**) The expected bias, $\mathrm{Bias}_{\mathrm{Av}}$; (**b**) the expected variance, $\mathrm{Var}_{\mathrm{Av}}$; (**c**) the expected RMS, $\mathrm{RMS}_{\mathrm{Av}}$; and (**d**) the expected comparative performance bias, $E_{\Phi_m}[C_{\mathrm{bias}}|\Phi_m]$: resulted from the distributions of $\varepsilon_{\mathrm{est}}^{\min}$ and $\varepsilon_{\mathrm{true}}^{i_{\min}}$ on the synthetic data for all 36 classification models, with respect to the collection size m.

For the real data, Figures 5 and 6 show corresponding results to Figures 3 and 4, respectively (ignore for the moment the 'average' curves, which will soon be discussed). Note the widely different behaviors between the different datasets. Since we are using the full dataset as an empirical distribution to serve as an approximation of the underlying feature-label distribution and are sampling from the empirical distribution, the different biases result from different behaviors of the error estimators on the different distributions. In practice, given a single sample dataset, one would have no idea of what kind of biases and RMS deviations to expect. This uncertainty exemplifies the standard conundrum faced when one lacks prior information regarding the feature-label distribution. In our case, the problem is that error estimator performance is heavily dependent on the underlying feature-label distribution.

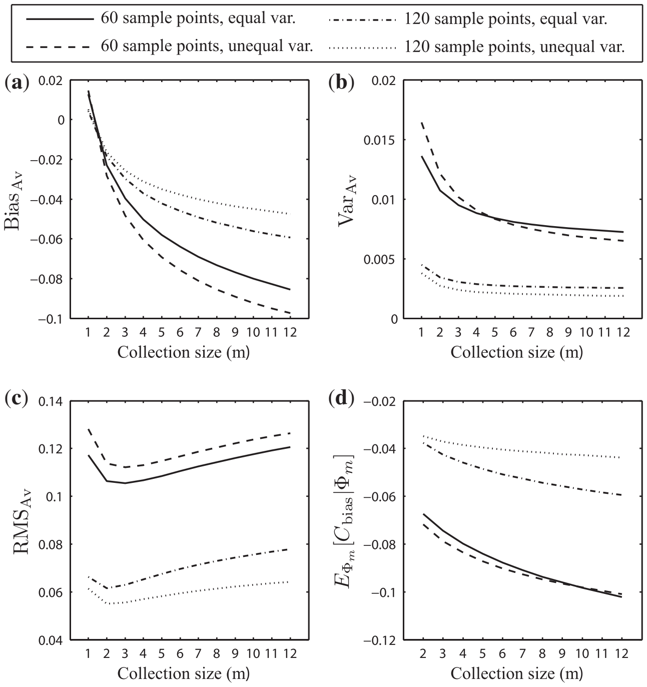

Fig. 4. (**a**) The expected bias, Bias$_{\text{Av}}$; (**b**) the expected variance, Var$_{\text{Av}}$; (**c**) the expected RMS, RMS$_{\text{Av}}$; and (**d**) the expected comparative performance bias, $E_{\Phi_m}[C_{\text{bias}}|\Phi_m]$: resulted from the distributions of $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{i_{\min}}$ on the synthetic data for single LOO error estimation, with respect to the collection size m.

3.3 Averaging over different populations

The problem of multiple-rule bias is exacerbated if one combines it with applying the multiple rules across multiple datasets (Yousefi *et al*., 2010) and then minimizes over both the classifier models and datasets; however, using multiple datasets can mitigate multiple-rule bias if performances are averaged over the datasets. In this case, each dataset is a sample from a feature-label distribution F_k, $k=1,2,\ldots,K$, and our concern is with average performance over the K feature-label distributions. Assuming multiple error estimators, there are m classification rules being considered over the K feature-label distributions. Our interest is now with

$$\varepsilon_{\text{est}}^{\min}(K)=\min\left\{\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{1,1,k},\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{1,2,k},\ldots,\right.$$
$$\left.\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{1,s,k},\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{2,1,k},\ldots,\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{r,s,k}\right\} \quad (9)$$

where $\varepsilon_{\text{est}}^{i,j,k}$ is the estimated error of classifier ψ_i and the error estimation rule Ξ_j on the dataset from feature-label distribution F_k. The bias takes the form

$$B(m,n,K)=E\left[\varepsilon_{\text{est}}^{\min}(K)\right]-\frac{1}{K}\sum_{k=1}^{K}E\left[\varepsilon_{\text{true}}^{i_{\min},k}\right] \quad (10)$$

where $\varepsilon_{\text{true}}^{i_{\min},k}$ is the true error of classifier $\psi_{i_{\min}}$ on the dataset from feature-label distribution F_k.

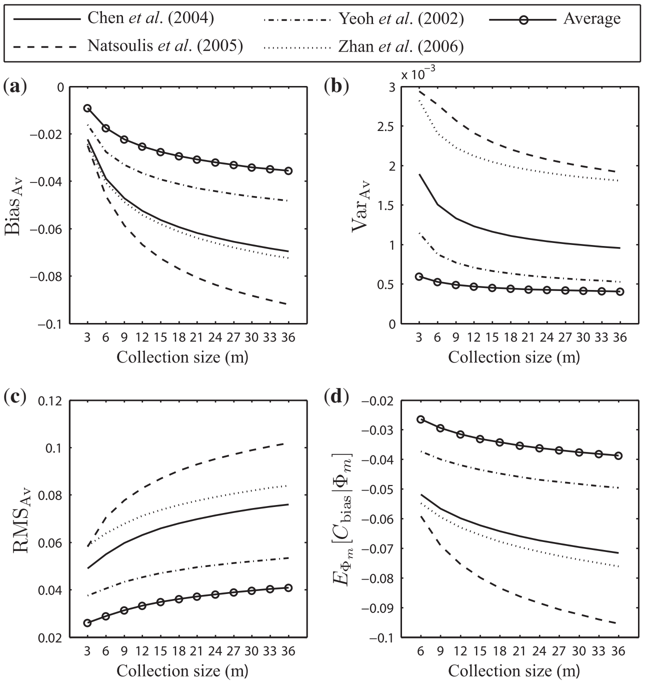

Fig. 5. (**a**) The expected bias, Bias$_{\text{Av}}$; (**b**) the expected variance, Var$_{\text{Av}}$; (**c**) the expected RMS, RMS$_{\text{Av}}$; and (**d**) the expected comparative performance bias, $E_{\Phi_m}[C_{\text{bias}}|\Phi_m]$: resulted from the distributions of $\varepsilon_{\text{est}}^{\min}$ and $\varepsilon_{\text{true}}^{i_{\min}}$ on the real data for all 36 classification models, with respect to the collection size m.

The 'average' curves in Figures 5 and 6 illustrate the effects of averaging. They show less estimation bias, less variance, smaller RMS and less comparative-performance bias when averaging is employed, as opposed to using the datasets individually. The situation is similar, albeit a bit more complicated, than the argument in Yousefi *et al.* (2010) for averaging results over a large number of datasets. Here, the averaging is done to mitigate multiple-rule bias.

It is possible to obtain theoretical results regarding the effect of averaging on the bias of Equation (10). In particular, if all error estimators are unbiased, then we prove in the Appendix A that $\lim_{K\to\infty}B(m,n,K)=0$. If one looks closely at the proof, it is clear that the unbiasedness assumption can be relaxed in each of the lemmas. In the first lemma, which shows that $B(m,n,K)\leq 0$, we need only assume that none of the error estimators is pessimistically biased. In the second lemma, which shows that $\lim_{K\to\infty}B(m,n,K)\geq 0$, unbiasedness can be relaxed to weaker conditions regarding the expectations of the terms making up the minimum in Equation (9); however, these conditions are rather arcane and do not add much practical insight. Moreover, we are interested in close-to-unbiased error estimators, specifically, the cross-validation estimators, so that we can expect $|B(m,n,K)|$ to diminish with averaging, even if the limit of $B(m,n,K)$ does not actually converge to 0.

3.4 Concluding remarks

From a practical standpoint, the results obtained in the present paper quantitatively demonstrate the large degree of overoptimism that results from comparing classifier rule models via their performances on a small dataset owing to the inaccuracy of error estimation on

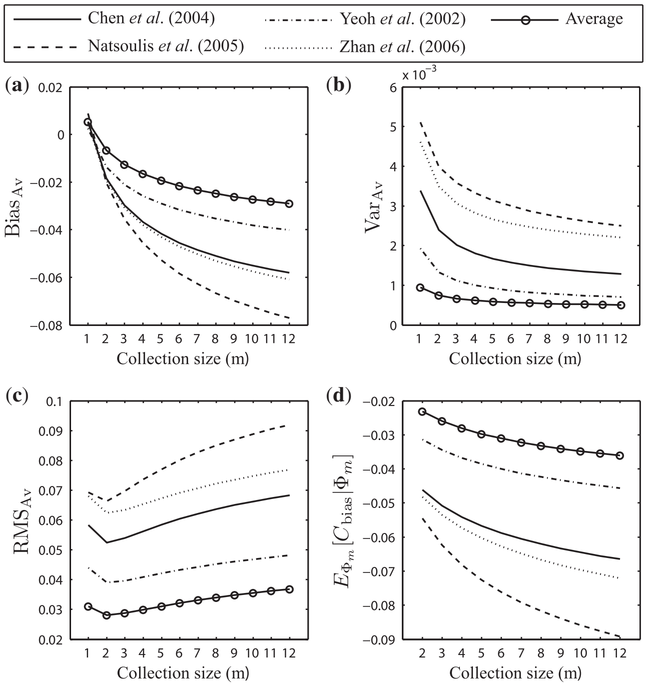

Fig. 6. (**a**) The expected bias, $\mathrm{Bias}_{\mathrm{Av}}$; (**b**) the expected variance, $\mathrm{Var}_{\mathrm{Av}}$; (**c**) the expected RMS, $\mathrm{RMS}_{\mathrm{Av}}$; and (**d**) the expected comparative performance bias, $E_{\Phi_m}[C_{\mathrm{bias}}|\Phi_m]$: resulted from the distributions of $\varepsilon_{\mathrm{est}}^{\min}$ and $\varepsilon_{\mathrm{true}}^{i_{\min}}$ on the real data for single LOO error estimation, with respect to the collection size *m*.

small samples. As the array of simulations show, optimistic bias accrues rapidly with even a small number of models being compared. We have observed from both simulations and theoretical analysis that the problem can be mitigated by averaging performances across a family of datasets; indeed, this is the recommendation that we put forth. The downside is that averaging eliminates the possibility of comparing classification rule performances on a single population. In fact, the latter possibility has been precluded at the outset by the experimental design: too small of a sample to obtain accurate error estimates. If there is only a single small sample, then the multiple-rule bias precludes any conclusions whatsoever, whereas at least if a collection of datasets are employed, then one may be able to make a conclusion relative to the collection of populations (depending on the accuracy of the error estimator). Even with averaging, we must offer a word of caution. While the proposition we have proven regarding the convergence of $B(m,n,K)$ is promising, like most distribution-free results it leaves open the rate of convergence, which in practice determines the number of datasets one must utilize to reduce the bias to some predetermined level. This leads us to some final comments.

The concerns expressed regarding the difficulty of establishing performance advantages for proposed classification rules (Boulesteix, 2010; Jelizarow *et al.*, 2010; Rocke *et al.*, 2009) reflect fundamental epistemological issues confronting bioinformatics as it addresses the high-throughput environment with limited sample sizes and limited statistical knowledge of how to deal with this new world (Dougherty and Braga-Neto, 2006; Dougherty, 2008; Mehta *et al.*, 2004). The problem addressed in this article arises from the bias and variance, and therefore the RMS, of error estimators. Very little is known about the performance of common error estimators, in particular, cross-validation. To take a salient example: LOO. Prior to 2009, all that was known about LOO for LDA and Gaussian class-conditional distributions were asymptotic expressions for the expectation and variance of the estimator in one dimension (Davison and Hall, 1992). In 2009, the distribution of the LOO estimator was discovered in this model for an arbitrary dimension *m* without assuming a common variance for $m=1$ and assuming a common covariance matrix with $m>1$ (Zollanvari *et al.*, 2009). Still, none of these results treated the joint distribution of the estimated and true errors, nor, in particular, the RMS. In 2010, the joint distribution was found exactly for $m=1$ without assuming a common variance and an approximation was found for $m>1$ assuming a common covariance matrix (Zollanvari *et al.*, 2010). Besides the joint distribution via complete enumeration (Xu *et al.*, 2006) and the correlation (Braga-Neto and Dougherty, 2010) for multinomial discrimination, there are no other analytic results regarding the joint behavior of LOO with the true error. This dearth of results is striking considering that LOO was first proposed in 1968 (Lachenbruch and Mickey, 1968), it has been used extensively, and the variance problems of LOO have been known from at least 1978 (Glick, 1978). As for more complicated cross-validation estimators that require random resampling, essentially nothing is known. If lamentations regarding the lack of performance characterization in bioinformatics are to be abated, then much greater knowledge regarding error estimation must be discovered.

ACKNOWLEDGEMENTS

We would like to thank the High-Performance Biocomputing Center of TGen for providing the clustered computing resources used in this study; this includes the Saguaro-2 cluster supercomputer, a collaborative effort between TGen and the ASU Fulton High Performance Computing Initiative.

Funding: National Institutes of Health grant (Saguaro-2 cluster, 1S10RR025056-01, in part) and the National Science Foundation (CCF-0634794).

Conflict of Interest: none declared.

REFERENCES

Boulesteix,A.-L. (2010) Over-optimism in bioinformatics research. *Bioinformatics*, **26**, 437–439.

Boulesteix,A.-L. and Strobl,C. (2009) Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Med. Res. Meth.*, **9**, 85.

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. (2006) Exact performance of error estimators for discrete classifiers. *Pattern Recognit.*, **38**, 1799–1814.

Braga-Neto,U.M. and Dougherty,E.R. (2010) Exact correlation between actual and estimated errors in discrete classification. *Pattern Recognit. Lett.*, **31**, 407–412.

Chen,X. *et al.* (2004) Novel endothelial cell markers in hepatocellular carcinoma. *Modern Pathol.*, **17**, 1198–1210.

Davison,A.C. and Hall,P. (1992) On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems. *Biometrika*, **79**, 279–284.

Dougherty,E.R. and Braga-Neto,U.M. (2006) Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity. *J. Biol. Syst.*, **14**, 65–90.

Dougherty,E.R. *et al.* (2007) Validation of computational methods in genomics. *Curr. Genomics*, **8**, 1–19.

Dougherty,E.R. (2008) On the epistemological crisis in genomics. *Curr. Genomics*, **9**, 69–79.

Glick,N. (1978) Additive estimators for probabilities of correct classification. *Pattern Recognit.*, **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.*, **2007**, 38473.

Hanczar,B. and Dougherty,E.R. (2010) On the comparison of classifiers for microarray data. *Curr. Bioinformatics*, **5**, 29–39.

Hanczar,B. *et al.* (2010) Small-sample precision of ROC-related estimates. *Bioinformatics*, **26**, 822–830.

Hua,J. *et al.* (2009) Performance of feature selection methods in the classification of high-dimensional data. *Pattern Recognit.*, **42**, 409–424.

Jelizarow,M. *et al.* (2010) Over-optimism in bioinformatics: an illustration. *Bioinformatics*, **26**, 1990–1998.

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

Mehta,T. *et al.* (2004) Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.*, **36**, 943–947.

Natsoulis,G. *et al.* (2005) Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.*, **15**, 724–736.

Rocke,D.M. *et al.* (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, **25**, 701–702.

Shmulevich,I. and Dougherty,E.R. (2007) *Genomic Signal Processing*. Princeton University Press, Princeton.

Xu,Q. *et al.* (2006) Confidence intervals for the true classification error conditioned on the estimated error. *Technol. Cancer Res. Treat.*, **5**, 579–589.

Yeoh,E.J. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.

Yousefi,M.R. *et al.* (2010) Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, **26**, 68–76.

Zhan,F. *et al.* (2006) The molecular classification of multiple myeloma. *Blood*, **108**, 2020–2028.

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

Zollanvari,A. *et al.* (2010) Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis. *IEEE Trans. Inform. Theory*, **56**, 784–804.

APPENDIX A

We prove that if all error estimators are unbiased, then $\lim_{K\to\infty} B(m,n,K)=0$.

LEMMA A.1. *If all error estimators are unbiased, then* $B(m,n,K)\le 0$.

PROOF. Define the set $\mathcal{S}_n=\{\mathcal{S}_n^1,\mathcal{S}_n^2,\dots,\mathcal{S}_n^K\}$, where $\mathcal{S}_n^k$, $k=1,2,\dots,K$ is a random sample taken from the distribution F_k for $k=1,2,\dots,K$. Also, we can rewrite Equation (9) as

$$\varepsilon_{\text{est}}^{\min}(K)=\min_{i,j}\left\{\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{i,j,k}\right\},\tag{A.1}$$

where $i=1,2,\dots,r$ and $j=1,2,\dots,s$. Owing to the unbiasedness of the error estimators, $E_{\mathcal{S}_n^k}[\varepsilon_{\text{est}}^{i,j,k}]=E_{\mathcal{S}_n^k}[\varepsilon_{\text{true}}^{i,k}]$. Referring to Equations (10) and (A.1), we have

$$\begin{aligned}
&B(m,n,K)\\
&\quad=E_{\mathcal{S}_n}[\varepsilon_{\text{est}}^{\min}(K)]-\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\right]\\
&\quad=E_{\mathcal{S}_n}\left[\min_{i,j}\left\{\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{i,j,k}\right\}\right]-\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\right]\\
&\quad\le\min_{i,j}\left\{E_{\mathcal{S}_n}\left[\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{i,j,k}\right]\right\}-\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\right]\\
&\quad=\min_{i,j}\left\{\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n}\left[\varepsilon_{\text{est}}^{i,j,k}\right]\right\}-\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\right]\\
&\quad=\min_{i,j}\left\{\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{est}}^{i,j,k}\right]\right\}-\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\right]\\
&\quad=\min_{i}\left\{\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i,k}\right]\right\}-\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\right]\\
&\quad\le 0.
\end{aligned}\tag{A.2}$$

where the relations in the third and sixth lines result from Jensen's inequality and unbiasedness of the error estimators, respectively. □

LEMMA A.2. *If all error estimators are unbiased, then* $\lim_{K\to\infty}B(m,n,K)\ge 0$.

PROOF. Let

$$A^{i,j}=\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{\text{est}}^{i,j,k},\qquad T^{i}=\frac{1}{K}\sum_{k=1}^{K}E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i,k}\right].\tag{A.3}$$

Owing to the unbiasedness of the error estimators, $E_{\mathcal{S}_n}[A^{i,j}]=T^i\le 1$. Without loss of generality, we assume $T^1\le T^2\le\dots\le T^r$. To avoid cumbersome notation, we will further assume that $T^1<T^2$ (with some adaptation, the proof goes through without this assumption). Let $2\delta=T^2-T^1$ and

$$\begin{aligned}
B_\delta=&\left(\bigcap_{j=1}^{s}\left(T^1-\delta\le A^{1,j}\le T^1+\delta\right)\right)\\
&\bigcap\left(\min_{i\ne 1,j}\left\{A^{i,j}\right\}>T^1+\delta\right).
\end{aligned}\tag{A.4}$$

Because $|\varepsilon_{\text{est}}^{i,j,k}|\le 1$, $\text{Var}_{\mathcal{S}_n}[A^{i,j}]\le 1/K$. hence, for $\tau>0$, there exists $K_{\delta,\tau}$ such that $K\ge K_{\delta,\tau}$ implies $P(B_\delta(K))>1-\tau$. Hence, referring to Equation (10), for $K\ge K_{\delta,\tau}$,

$$\begin{aligned}
E_{\mathcal{S}_n}\left[\varepsilon_{\text{est}}^{\min}(K)\right]=&E_{\mathcal{S}_n}\left[\varepsilon_{\text{est}}^{\min}(K)\,|\,B_\delta\right]P(B_\delta)\\
&+E_{\mathcal{S}_n}\left[\varepsilon_{\text{est}}^{\min}(K)\,|\,B_\delta^c\right]P(B_\delta^c)\\
\ge&E_{\mathcal{S}_n}\left[\varepsilon_{\text{est}}^{\min}(K)\,|\,B_\delta\right]P(B_\delta)\\
=&E_{\mathcal{S}_n}\left[\min{}_jA^{1,j}\right]P(B_\delta)\\
\ge&(T^1-\delta)(1-\tau).
\end{aligned}\tag{A.5}$$

Again referring to Equation (10) and recognizing that $i_{\min}=1$ in B_δ, for $K \geq K_{\delta,\tau}$,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\right] &= \frac{1}{K}\sum_{k=1}^{K}\Big(E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\,|\,B_\delta\right]P(B_\delta) \\ &\quad + E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{i_{\min},k}\,|\,B_\delta^c\right]P(B_\delta^c)\Big) \\ &\leq \frac{1}{K}\sum_{k=1}^{K}\Big(E_{\mathcal{S}_n^k}\left[\varepsilon_{\text{true}}^{1,k}\,|\,B_\delta\right]P(B_\delta) \\ &\quad + P(B_\delta^c)\Big) \\ &\leq T^1+\tau. \end{aligned} \tag{A.6}$$

Putting Equations (A.5) and (A.6) together and referring to Equation (10) yields, for $K \geq K_{\delta,\tau}$,

$$B(m,n,K) \geq (T^1-\delta)(1-\tau)-T^1-\tau \geq -(2\tau+\delta) \tag{A.7}$$

Since δ and τ are arbitrary positive numbers, this implies that for any $\eta>0$, there exists K_η such that $K \geq K_\eta$ implies $\lim_{K\to\infty} B(m,n,K) \geq 0$, which is precisely what we want to prove. □

Combining Lemmas A.1 and A.2, we have proven that $\lim_{K\to\infty} B(m,n,K)=0$ under the assumption that all the error estimators are unbiased.