Gene expression

Reporting bias when using real data sets to analyze classification performance

Mohammadmahdi R. Yousefi1, Jianping Hua2, Chao Sima2 and Edward R. Dougherty1,2,*

1Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 and 2Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA

Received on August 7, 2009; revised on October 16, 2009; accepted on October 17, 2009

Advance Access publication October 21, 2009

Associate Editor: Martin Bishop

1 INTRODUCTION

Classification plays a key role in bioinformatics, for instance, gene expression can be used to discriminate between different kinds of cancer, different stages of tumor development or many such differences. With microarray data, classifiers are designed from a sample of expression vectors, which requires assessing expression levels from RNA obtained from the different tissues with microarrays, determining genes whose expression levels can be used as classifier variables, and then applying some rule to design the classifier from the sample microarray data. It is common practice in the bioinformatics literature for someone to propose a new classification rule, perhaps in the form of a construction rule or a new feature-selection method, and support the proposed classification rule by showing its performance on some data sets, often by comparing its performance on these data sets with some other classification rules. Typically, these data sets represent classification problems for different populations, so that each represents a sample from a different population. Henceforth, we will refer to these real data sets as samples, keeping in mind that they arise from distinct populations. Our interest here is in studying the potential for ‘reporting bias’, by which we mean the optimistic bias that results if the paper does not provide performance results on all samples considered in the research. To wit, suppose the proposed classification rule has been tested on m samples but the paper only reports the results on k < m samples. If k is small in comparison with m, then there will be optimistic bias in the reported errors.

This bias has two inherent sources resulting from working with high-dimensional problems on small samples. First, when feature selection is involved, which is inevitable with high-throughput biological measurements, there is little correlation between the errors of the selected feature set and the best feature set, there being a wide dispersion of the errors of selected feature sets (Sima and Dougherty, 2006). The second problem, and one that influences the first, is the high variance of cross-validation error estimation when small samples are used (Braga-Neto and Dougherty, 2004; Glick, 1978); indeed, there is generally very little correlation between the true and estimated errors (Hanczar et al., 2007). The result is that, when one orders a list of estimated errors, the better estimates tend to be biased optimistically and the worst tend to be biased pessimistically. Thus, reporting the high end of the list creates reporting bias. The purpose of this article is to quantify that bias.

We will consider this issue from two perspectives. First, so that we know the feature-label distribution from which the samples are drawn and there is no problem with obtaining very precise estimates of the true error, we will assume a given feature-label model and consider m samples from the model. This synthetic approach has the added advantage that the expected errors for all samples are the same. Next, we will consider m real data samples, where we will have to handle the problem of estimating the true errors and the
fact that the samples represent different feature-label distributions so that the expected errors are different.

To characterize the issue of reporting bias, let us assume that all samples are from a single distribution $F$ and focus on the case for $k=1$, so that only the results with minimum estimated error are reported. Let $S = \{S_1, S_2, \ldots, S_m\}$ be a family of $m$ samples drawn from $F$, of identical sample size $n$. A classification and performance evaluation scheme is composed of $(\Omega, \Lambda, \Sigma)$, where $\Sigma$ is the feature-sequence part of the classification rule, $\Lambda$ is the classifier construction part of the classification rule and $\Omega$ is the error estimation procedure. The true error for the designed classifier, $\hat{\epsilon}_{\alpha}$, on $S_1$ is given by $P(\hat{\psi}_{\alpha} | X \neq Y)$, where $(X, Y)$ is the feature-label pair. The estimated error on $S_1$ is $\hat{\epsilon}_{\alpha} = E[\hat{\psi}_{\alpha} | X \neq Y]$. Assuming that the $m$ samples $S_i$ are i.i.d., then the $m$ estimated errors are i.i.d., as are the corresponding true errors. The minimum estimated error is then:

$$\hat{\epsilon}_{\alpha}^{\min} = \min_{i=1}^{m}[\hat{\epsilon}_{\alpha}^{i}]$$

whose distribution, $F_{\alpha}^{\min}(x) = P(\hat{\epsilon}_{\alpha}^{\min} \leq x)$, is given by:

$$F_{\alpha}^{\min}(x) = 1 - \left[1 - P(\hat{\epsilon}_{\alpha} \leq x)^m\right]$$

for any $i$. Letting $S_1^{\min}$ denote the sample on which the minimum error estimate occurs, $\hat{\epsilon}_{\alpha}^{\min} = \hat{\epsilon}_{\alpha}^{1}$, and the corresponding true error is $\epsilon_{\alpha}^{1}$. The potential for bias is seen in Equation (2). As long as $P(\hat{\epsilon}_{\alpha} \leq x) > 0$, $P(\hat{\epsilon}_{\alpha}^{\min} \leq x) \rightarrow 1$ as $m \rightarrow \infty$. Our interest in the present article is characterizing the resulting biases for certain models and for some instances of real data. In addition to the minimum estimated error, we consider the second minimum estimated error, which will still exhibit bias.

2 SYSTEMS AND METHODS

2.1 Performance measures

Continuing to focus on $k=1$, we investigate four deviation distributions that demonstrate performance relative to the bias issue. Since the true errors for the synthetic data are i.i.d. $E_1[\epsilon_1] = \cdots = E_1[\epsilon_m]$, the expectation being taken across all families $S$. We will write $E_1[\epsilon_{\alpha}]$ to denote this common expectation, which owing to i.i.d. gives the expected true error across the samples. Similarly, since the estimated errors are i.i.d. $E_1[\hat{\epsilon}_{\alpha}^i] = \cdots = E_1[\hat{\epsilon}_{\alpha}^m]$, and we write $E_1[\hat{\epsilon}_{\alpha}]$ to denote this common expectation, which gives the expected estimated error across the samples. These four deviation distributions are:

(1) The deviation distribution $\epsilon_1^{\alpha} - \hat{\epsilon}_{\alpha}^{\min}$ describes the effect of taking the estimated error on $S_1^{\alpha}$ as a characterization of its performance. The distribution is a function of both sample size $n$ and the number $m$ of samples. The mean of this distribution is $E_1[\epsilon_1^{\alpha}] - E_1[\hat{\epsilon}_{\alpha}^{\min}]$, where $E_1[\epsilon_1^{\alpha}]$ and $E_1[\hat{\epsilon}_{\alpha}^{\min}]$ are the expected estimated error and the expected true error for $S_1^{\alpha}$, respectively, across all families $S$. It gives the bias of the estimated error on $S_1^{\alpha}$, relative to the true error on $S_1^{\alpha}$. Thus, a claim based on $\epsilon_1^{\alpha}$ that a proposed classifier performs well on the kind of data represented by $S_1^{\alpha}$ would be unjustified.

(2) The deviation distribution $\epsilon_1^{\alpha} - E_1[\epsilon_{\alpha}]$ describes the difference between the minimum estimated error and the expected true error across the samples. The mean, $E_1[\epsilon_1^{\alpha}] - E_1[\epsilon_{\alpha}]$, of this distribution gives the bias of the minimum estimated error relative to the expected true error. Thus, if only the result for $S_1^{\alpha}$ is reported, not only would this not justify a claim of superior performance on the data represented by $S_1^{\alpha}$, but it would also be optimistically biased relative to the classifier performance across the various kinds of data represented by the samples, an important factor for bioinformaticians whose interests are with the different kinds of data tested, but not reported.

(3) The deviation distribution $\epsilon_1^{\alpha} - \frac{1}{m} \sum_{i=1}^{m} \epsilon_1^{i}$ describes the relation between the minimum estimated error and the average estimated error across the samples. This distribution is fully determined by the distribution of the individual estimated errors $\epsilon_1^{i}$. The mean, $E_1[\epsilon_1^{\alpha}] - E_1[\epsilon_{\alpha}]$, of the distribution gives the bias of the minimum estimated error relative to the mean estimated error. Thus, if only the result for $S_1^{\alpha}$ were reported, it would be optimistically biased relative to the overall tests performed on the classifier.

(4) The deviation distribution $\epsilon_1^{\alpha} - E_1[\epsilon_{\alpha}]$ describes the relation between the true error of $S_1^{\alpha}$ and the expected true error across the samples. The mean, $E_1[\epsilon_1^{\alpha}] - E_1[\epsilon_{\alpha}]$, of the distribution gives the bias of the minimum estimated error relative to the expected true error across the samples, this measure gives the bias of the true error corresponding to the minimum estimated error relative to the expected true error across the samples.

Analysis of the error rates for real data is, to some extent, different from synthetic data. For the synthetic data, the $m$ samples are i.i.d., in particular, $E_1[\epsilon_{\alpha}^{i}] = \cdots = E_1[\epsilon_{\alpha}^{m}]$. Thus, choosing the minimum estimated error is a reasonable way to measure the best performance across the samples. Since the real data are from different populations, the estimated and true errors are no longer i.i.d., in particular, the expected estimated and expected true errors are no longer identical. Thus, when using real data, we normalize the expected errors by subtracting the corresponding expected true errors and measure performance by:

$$\delta_{\alpha}^{est} = \frac{\epsilon_{\alpha}^{est} - E_1[\epsilon_{\alpha}]}{E_1[\epsilon_{\alpha}]}$$

The best performance is given by:

$$\delta_{\alpha}^{est} = \min_i(\delta_{\alpha}^{est i})$$

For the real data, $S_1^{\alpha}$ denotes the sample on which this minimum occurs and $\delta_{\alpha}^{est} = \min_i(\delta_{\alpha}^{est i})$. The true error is also normalized by $E_1[\epsilon_{\alpha}]$ and $\delta_{\alpha}^{true} = \frac{\epsilon_{\alpha}^{true} - E_1[\epsilon_{\alpha}]}{E_1[\epsilon_{\alpha}]}$ is used instead of $\epsilon_{\alpha}^{true}$. $\delta_{\alpha}^{true}$ corresponds to the normalized true error on $S_1^{\alpha}$.

For the real data, there are four deviation distributions corresponding to those for the synthetic data. Since the samples are not being drawn according to an i.i.d. random family $S = \{S_1, S_2, \ldots, S_m\}$, there is no direct correspondent to $E_1[\epsilon_{\alpha}]$. Thus, we extend the definitions of the second and fourth deviation distributions for the synthetic data by replacing $E_1[\epsilon_{\alpha}]$ by the average $\frac{1}{m} \sum_{i=1}^{m} \epsilon_{\alpha}^{i}$. The four deviation distributions for the real data are given by:

(6) $\delta_{\alpha}^{est} = \frac{\epsilon_{\alpha}^{est} - E_1[\epsilon_{\alpha}]}{E_1[\epsilon_{\alpha}]}$

$$\delta_{\alpha}^{est} = \frac{1}{m} \sum_{i=1}^{m} \epsilon_{\alpha}^{i} - E_1[\epsilon_{\alpha}]$$

(7) $\delta_{\alpha}^{true} = \frac{\epsilon_{\alpha}^{true} - E_1[\epsilon_{\alpha}]}{E_1[\epsilon_{\alpha}]}$

$$\delta_{\alpha}^{true} = \frac{1}{m} \sum_{i=1}^{m} \epsilon_{\alpha}^{i} - E_1[\epsilon_{\alpha}]$$

(8) $\delta_{\alpha}^{true} = \frac{\epsilon_{\alpha}^{true} - E_1[\epsilon_{\alpha}]}{E_1[\epsilon_{\alpha}]}$

$$\delta_{\alpha}^{true} = \frac{1}{m} \sum_{i=1}^{m} \epsilon_{\alpha}^{i} - E_1[\epsilon_{\alpha}]$$
This is a price paid for using real data, for which one cannot generate i.i.d. distribution to be considered, those of Equations (6) and (8).

δ

Taking expectations in each of these equations yields

E_δ[δ_{est} - δ_{true}] = E_δ(E_{true} - E_{est})

(10)

E_δ[δ_{est} - δ_{true}] = E_δ(E_{true} - E_{est})

(11)

E_δ[δ_{est} - δ_{true}] = E_δ(E_{true} - E_{est})

(12)

The mean for the basic δ-deviation of Equation (6) has the same mean [Equation (10)] as the basic δ-deviation (first deviation distribution) for the synthetic data. Since the δ-deviation of Equation (7) also has the same mean, we need not consider it. The mean of the δ-deviation of Equation (8) differs by subtracting the average of the error estimation biases across the real data samples [Equation (12)]. Since we will be using cross-validation for error estimation and cross-validation is pessimistically biased, the mean of the δ-deviation of Equation (8) will be reduced. Finally, the mean of the δ-deviation of Equation (9) is 0, so that this δ-deviation is not considered. Owing to these observations, there are only two real data deviation distributions to be considered, those of Equations (6) and (8).

This is a price paid for using real data, for which one cannot generate i.i.d. synthetic data. The expectation E_δ[δ_{true}] does not exist for the real data and replacing it by 1/n \sum_{i=1}^{n} δ_{true} yields useless δ-deviation expectations.

In the case of the real data, as is typically done, the training sample will be drawn from the full real-data set and the true error estimated by using larger portion of the data set not used for training. Note that the absence of normalization for the synthetic data is not because the data are synthetic but because, in the case of synthetic data, we are able to model the problem with \Sigma_{1}, \Sigma_{2}, ... , \Sigma_{n} being i.i.d.

2.2 Simulation design

For the synthetic data, we utilize a parametric distribution model whose properties reflect those of observations made in microarray experiments. Synthetic data samples are generated from the distribution model and tested on different classification schemes. Then, we conduct a simulation utilizing several microarray expression data sets.

2.2.1 Synthetic data

We employ a general model based on multivariate Gaussian distributions with a block-based covariance structure that conforms to various observations made in microarray expression-based studies (Hua et al., 2009). Whereas the current model was initially introduced to provide a model-based study of the peaking phenomenon in the kind of high-dimensional setting not envisioned in classical peak studies, blocking the covariance matrix is a standard tool to model groups of interacting variables where there is negligible interaction between the groups. The block-covariance model has been used extensively 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). A battery of distribution models is constructed by changing model parameters to generate different synthetic data samples. Sample points can be categorized primarily into two equally likely classes (C_0 versus C_1) with a feature-label distribution of feature size D. Furthermore, cases like different stages or subtypes of a cancer can be represented by ϵ equally likely subclasses in C_1, each having its own distribution. Each sample point in C_1 belongs to one and only one of these subclasses.

There are two major groups of features: markers and non-markers. Markers resemble genes causing disease or susceptibility to disease. Thus, they have different class-conditional distributions for the two classes. Non-markers correspond to genes having no discriminating ability across different classes. The markers can be further subgrouped into two different types: global markers and heterogeneous markers. The class-conditional distributions of the global markers in the two classes are assumed to be \Sigma_{gm}-dimension Gaussian distributions with parameters (\mu_{gm}^{0}, \Sigma_{gm}^{0}) for class 0 and (\mu_{gm}^{1}, \Sigma_{gm}^{1}) for class 1, where \Sigma_{gm} is the total number of global markers homogeneously distributed among the two classes.

Class 1 can have c subclasses within itself, each one associated with \Sigma_{hm} mutually exclusive heterogeneous markers having \Sigma_{hm}-dimension Gaussian class-conditional distributions with parameters (\mu_{hm}^{0}, \Sigma_{hm}^{0}) for class 0 and (\mu_{hm}^{1}, \Sigma_{hm}^{1}) for class 1, where \Sigma_{hm} is the total number of heterogeneous markers homogeneously distributed among the two classes.

Assuming that global and heterogeneous markers possess identical covariance structures, we use Σ_{1} = Σ_{2} instead of \Sigma_{gm}^{0}, \Sigma_{gm}^{1} and \Sigma_{hm}^{0}, \Sigma_{hm}^{1}, and we assume that Σ_{1} = σ_{1}^{2}Σ and Σ_{2} = σ_{2}^{2}Σ, where σ_{1}^{2} and σ_{2}^{2} can be different, and Σ has the following block structure:

\begin{equation}
\Sigma = \begin{bmatrix}
\Sigma_{1} & 0 & 0 & \cdots & 0 \\
0 & \Sigma_{2} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \Sigma_{1} & 0 \\
0 & 0 & \cdots & 0 & \Sigma_{1}
\end{bmatrix}
\end{equation}

where Σ is a k × k 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 k. 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 (Hua et al., 2009).

By assigning different values to mean vectors one can define different basic models. Because the global markers and the heterogeneous markers possess the same structure, we can assume the same mean vectors [\mu_{g}, \mu_{h}] for both groups [\mu_{g}^{0}, \mu_{h}^{0}] and [\mu_{g}^{1}, \mu_{h}^{1}], as we did for the covariance matrices. Furthermore, we use the same structure for \mu_{g} and \mu_{h} in the form
of $m_0 \times (1, \ldots, 1)$ and $m_1 \times (1, \ldots, 1)$, respectively, where $m_0$ and $m_1$ are scalars.

Like the global markers, two types of non-markers are defined: high-variance
non-markers and low-variance non-markers. The $D_m$ features belonging to the former are uncorrelated and their distributions are described by
$p(0, \sigma_2^m) = p(1, \sigma_1^m)$, where $\sigma_2^m$ and $\sigma_1^m$ take values equal to the
variances of the markers and $p$ is a random value uniformly distributed
over $[0, 1]$. The $D_m$ remaining features are uncorrelated low-variance
non-markers, each having Gaussian distribution with parameters $(0, \sigma_2^m)$.

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. A typical microarray experiment usually contains tens
of thousands of probes (genes). Hence, we choose the total number of
features to be $D = 20,000$. Also for variances $\sigma_2^m$ and $\sigma_1^m$, two cases are
considered: equal variances $[\sigma_2^m = 0.6, \sigma_1^m = 0.6]$ and unequal variances
$[\sigma_2^m = 0.6, \sigma_1^m = 1.2]$.

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. Given $m_0$ and $m_1$, lower error rates are
obtainable when the sample size is larger. We choose the linear discriminant
analysis (LDA) classifier, 120 sample points and the $t$-test for selecting five
features to design two experiments, that is, to define $m_0$ and $m_1$ for both equal
and unequal variances, with the average true errors close to 0.25. Preliminary
simulations show that to get the desired error rates, one can set $m_0 = 0.25$ and
$m_0 = 0.8$ when the variances are equal and $m_0 = 0.11$ and $m_0 = 0.9$ when they
are unequal.

For the blocked covariance matrix, we choose block size $k = 5$ and
correlation coefficient $\rho = 0.8$, giving relatively tight correlation within
a block, which would be expected for a pathway. Table 1 shows full details of
the distribution model.

Table 1. Distribution model parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$m_0 = 0.23$</td>
<td>$m_1 = 0.8$ (equal variances)</td>
</tr>
<tr>
<td></td>
<td>$m_0 = 0.11$</td>
<td>$m_1 = 0.9$ (unequal variances)</td>
</tr>
<tr>
<td>Variances</td>
<td>$\sigma_2^m = 0.6, \sigma_1^m = 0.6$</td>
<td>(equal variances)</td>
</tr>
<tr>
<td></td>
<td>$\sigma_2^m = 0.6, \sigma_1^m = 1.2$</td>
<td>(unequal variances)</td>
</tr>
<tr>
<td>Block size</td>
<td>$k = 5$</td>
<td>Features</td>
</tr>
<tr>
<td>Features</td>
<td>$D = 20000$</td>
<td>Feature block correlation $\rho = 0.8$</td>
</tr>
<tr>
<td>Subclasses</td>
<td>$c = 2$</td>
<td>Heterogeneous markers $D_{hm} = 0$</td>
</tr>
<tr>
<td>Classes</td>
<td>$D_{hm} = 50$</td>
<td>High-variance non-markers $D_{hm} = 2000$</td>
</tr>
<tr>
<td></td>
<td>$D_{lm} = 17980$, $D_{lm} = 17880$</td>
<td>Low-variance non-markers</td>
</tr>
</tbody>
</table>

2.2.2 Real data We assume there is a database containing $M$ real data sets,
each one drawn from a unique distribution. In each iteration of the simulation,
$n \leq M$ data sets are randomly drawn from the database and denoted as
$S_1^n, S_2^n, \ldots, S_M^n$. Based on a holdout scheme, for data set $S_m^n$, $n$ sample points are
drawn without replacement to form the training sample $S_m^n$, and the remaining
sample points in $S_m^n$ are used to compute the true error, $e_{true}$, where
the assumption is made that the holdout set is sufficiently large so that the
error on the holdout is close to the true error. As with the synthetic data,
the classification and performance evaluation scheme is $(\Omega, \Lambda, Z)$, where
$\Omega$ is the feature-selection part, $\Lambda$ is the classifier construction part, and $Z$
is the error-estimation procedure. The estimated error on sample $S_m^n$ is
$e_{est} = 20Q_A$.

This study uses $M = 12$ real data sets from microarray experiments, all
consisting of more than 150 arrays: Bhatnagar et al., 2000; Chen et al., 2001;
Dunn et al., 2001; Ewen et al., 2002; Foster et al., 2003; Gene Expression
Profiling; Eng et al., 2003; Furney et al., 2004; Golub et al., 2005; Haghighi et al., 2005; Redon et al., 2005; Rosenwald et al., 2002; Saal et al., 2001; Sall et al., 2002; van de Vijver et al., 2000; Wang et al., 2006; Yeoh et al., 2004; Zhan et al., 2005). Details are
given in the Supplementary Materials and the companion web site. To the
tent possible, we have tried to maintain the original labeling and follow
the data preparation directions used in the papers reporting these data sets;
however, in several cases we have relabeled sample points for reasons given
in the Supplementary Material. Table 2 presents a summary of the 12 real
data sets we have used in this study.

2.3 Classification schemes To obtain the distributions for performance analysis, a pre-defined
classification scheme is applied to $m$ random training samples of size $n$. A two-stage feature-selection procedure is used: a filter method in
the first stage reduces the number of features to $D$ and a wrapper method
in the second stage reduces these to a final feature set containing $d$ features.
A classifier is designed and the error estimated for the $d$ selected features.
For the synthetic data, the true error of the designed classifier is
computed on $n_l$ (large) independent sample points drawn from the same
distribution $F_l$ for the real data, it is computed on the holdout portion of
the sample.

Two classification rules, LDA and 3NN are considered in our simulation.
These have been chosen because they are well studied, the pair provides
linear and non-linear classification, and they are computationally tractable
for the extensive simulations required in the current study. For the two-
stage feature-selection scheme, we use the $t$-test in the first stage and
sequential forward search (SFS) in the second stage. We can simply have
one-stage feature selection by skipping one of the two stages. We only
consider skipping the second stage because using a wrapper method with
a very large number of features enormously increases the computational
time.

Bolstered resubstitution error estimation for the LDA classifier and semi-
bolstered resubstitution error estimation for the 3NN classifier are chosen to
enhance the accuracy level within the SFS feature-selection algorithm (Simu
et al., 2005). We employ 10-fold cross-validation for error estimation outside
the feature selection for LDA classifier, but use 5-fold cross-validation for
3NN classifier due to the fact that 3NN is an inherently slow classification
algorithm. The actual simulation for the synthetic data is done in the following
iterative process: first, 120,000 training samples of size $n_l = 60$ (or 120) and
test samples of size $n_l = 5000$ are generated from the assumed distribution
model using a set of parameters in Table 1 to compute 120,000 estimated
and true error pairs. In each iteration, $m$ error pairs are randomly selected from
these 120,000 pairs. Then, the minimum and the second minimum estimated
errors and their corresponding true errors are found, which results in the
empirical distribution of random variables defined in Section 1. $E_3(e_{true})$
and $E_3(e_{est})$ are computed by averaging the 120,000 pairs of estimated
and true errors.

To compute $E_3(e_{true})$ and $E_3(e_{est})$ for the real data, the same process
as for the synthetic data can be implemented here but only on individual
samples. Each sample $S_m^n$ has its own expected estimated and true errors.
In each iteration, $n_l = 60$ sample points are randomly drawn from $S_m^n$, a classifier
is designed, its estimated error computed via cross-validation, and its true error
taken to be the holdout error on the sample points not used for training. This
process is repeated 120,000 times with $E_3(e_{true})$ and $E_3(e_{est})$ computed by
averaging the estimated and true errors, respectively, for $S_m^n$. Details of the
classification schemes for the simulation on synthetic and real data are given
in Tables 3 and 4, respectively.

3 RESULTS AND DISCUSSION The complete set of results is available on the companion web site
for the article, including visualizations of the joint distributions for

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

<table>
<thead>
<tr>
<th>Data set</th>
<th>Data set type</th>
<th>Feature size</th>
<th>Sample size (Class 0/Class 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhattacharjee et al. (2001)</td>
<td>186 lung tumors and 17 normal lung</td>
<td>12460</td>
<td>139/64</td>
</tr>
<tr>
<td>Su et al. (2001)</td>
<td>11 different tumor samples</td>
<td>12533</td>
<td>83/91</td>
</tr>
<tr>
<td>van de Vijver et al. (2002)</td>
<td>Primary breast carcinomas</td>
<td>5003</td>
<td>70/196</td>
</tr>
<tr>
<td>Chen et al. (2004)</td>
<td>Hepatocellular carcinoma</td>
<td>10237</td>
<td>75/82</td>
</tr>
<tr>
<td>Natsoulis et al. (2005)</td>
<td>Drugs and toxicants response on rats data set</td>
<td>8491</td>
<td>120/61</td>
</tr>
<tr>
<td>Wang et al. (2005)</td>
<td>Lymph-node-negative breast cancer</td>
<td>22215</td>
<td>183/93</td>
</tr>
<tr>
<td>Potti et al. (2006)</td>
<td>Non-small-cell lung cancer</td>
<td>54613</td>
<td>156/78</td>
</tr>
<tr>
<td>Desmedt et al. (2007)</td>
<td>Node-negative breast cancer</td>
<td>22215</td>
<td>98/77</td>
</tr>
</tbody>
</table>

Table 3. Different classification schemes for the simulation of synthetic data

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature selection</th>
<th>Error estimation</th>
<th>Size of data</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st stage</td>
<td>2nd stage</td>
<td>Inside</td>
<td>Outside</td>
</tr>
<tr>
<td>Synthetic data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>t-test</td>
<td>1000</td>
<td>SFS</td>
<td>5 and 10</td>
</tr>
<tr>
<td>3NN</td>
<td>t-test</td>
<td>500</td>
<td>SFS</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4. Different classification schemes for the simulation of real data

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature selection</th>
<th>Error estimation</th>
<th>Size of data</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st stage</td>
<td>2nd stage</td>
<td>Inside</td>
<td>Outside</td>
</tr>
<tr>
<td>Real data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>t-test</td>
<td>1000</td>
<td>SFS</td>
<td>5 and 10</td>
</tr>
<tr>
<td>3NN</td>
<td>t-test</td>
<td>500</td>
<td>SFS</td>
<td>5</td>
</tr>
</tbody>
</table>

Here, we report some specific results that are indicative of the general trends and show how reporting bias can manifest itself. For LDA classification with 10-fold cross-validation, two feature-selection schemes and sample sizes 60 and 120, the means and variances of the four deviation distributions considered for the synthetic data are shown in Figure 2. The x-axis gives the number of samples over which the classifier has been tested, the y-axis gives the mean of the deviation distribution, $\mu_{\text{dev}}$, as a function of $m$ based on the 120 000 estimated and true error pairs, and the y-axis in the right column gives the variance of the deviation distribution, $\sigma^2_{\text{dev}}$, based on the 120 000 estimated and true error pairs.

The graph for $e_{\text{est}} - e_{\text{true}}$ shows that, if one selects the best performing sample, the error for the classifier designed from that sample is underestimated for $m \geq 2$, where for sample size 60, this bias is $-0.05$ with $m=4$, reaching $-0.1$ for $m=15$, the feature-selection scheme not making much difference. The low bias is about half if 120 sample points are used. This means that the reported error for the reported feature set is significantly optimistic. In particular, it is not uncommon to see samples of size 60 or less used to support a proposed classification rule and given the many existing classification rules, a performance improvement of 0.05 would be considered outstanding; yet with only four tested samples, the best would have a bias of $-0.05$, thereby negating any performance claim with regard to the reported sample if only the performing one were reported. Even with 120 sample points, which is quite large relative to the typical samples being used, at $m=4$ the negative bias exceeds $-0.03$, which in the scheme of things is significant because an improvement of 0.03 in classification accuracy would be considered important. Note that, because in general cross-validation is high-biased, for $m=1$ the bias is positive. Also note that the variance of $e_{\text{est}} - e_{\text{true}}$ decreases with increasing $m$. 

72
worse for about classifier performance. Here the low bias is even greater, reaching the tested samples, so that this statistic is more reflective of overall which characterizes the performance of the classification rule across gives the expected average true performance across all samples, Means (left column) and variances (right column) of four deviation distributions for the first and second minimum error estimates across the samples. One might use a family of the expected true error over all samples.

The bias curves for $\epsilon_{\text{est}}^\text{true} - E_0[\epsilon_{\text{est}}]$ are worse than those for $\epsilon_{\text{est}}^\text{true} - E_S[\epsilon_{\text{true}}]$: indeed, owing to the high bias of cross-validation,

$$E_0\left[\frac{1}{m} \sum_{i=1}^{m} \epsilon_{i,\text{est}}\right] > E_0[\epsilon_{\text{true}}]$$

and therefore

$$E_S[\epsilon_{\text{true}}] - E_0\left[\frac{1}{m} \sum_{i=1}^{m} \epsilon_{i,\text{est}}\right] < E_0[\epsilon_{\text{true}}] - E_S[\epsilon_{\text{true}}].$$

Note that the graphs for this distribution begin at $m=2$ because the expression reduces to 0 when $m=1$.

Unlike $\epsilon_{\text{est}}^\text{true} - E_0[\epsilon_{\text{true}}]$, the final deviation distribution, $\epsilon_{\text{true}} - E_S[\epsilon_{\text{true}}]$, is not affected by estimating the error for $S_0^{-m}$, so that its low bias is not as bad. The equality

$$\epsilon_{\text{true}}^\text{true} - E_S[\epsilon_{\text{true}}] = \epsilon_{\text{true}}^\text{true} - E_S[\epsilon_{\text{true}}]$$

clearly shows the relationship between $\epsilon_{\text{true}}^\text{true} - E_S[\epsilon_{\text{true}}]$ and $\epsilon_{\text{true}} - E_S[\epsilon_{\text{true}}]$, $E_0$. Nonetheless, the low bias of the first term, $\epsilon_{\text{true}}^\text{true} - E_S[\epsilon_{\text{true}}]$, demonstrates the reporting optimism of the true error resulting from selecting the best performing sample relative to the expected true error over all samples.

Figure 3 shows results analogous to Figure 2, except that it is for the second minimum estimated error. Although the optimistic bias for increasing $m$ is not as extreme, it is still substantial at modest numbers of samples. For sample size 60 with $t$-test feature selection, the biases reach $-0.05$ and $-0.075$ at $m=9$, respectively, for $\epsilon_{\text{est}}^\text{true}$ and $\epsilon_{\text{true}} - E_S[\epsilon_{\text{true}}]$, where $\epsilon_{\text{est}}^\text{true}$ and $\epsilon_{\text{true}}$ denote the estimated and true errors for the sample, $S_0^{-m}$, having the second lowest estimated error. Thus, if one tests a proposed classification rule on nine samples and reports the two best results, the reported results will possess strong optimistic bias. Note that bias $-0.075$ is reached at $m=9$, even though at $m=2$ the bias is positive on account of the positive bias of cross-validation.

Figure 4 considers 3NN classification for the synthetic data with 5-fold cross-validation, two feature-selection schemes and sample sizes 60 and 120. The left and right columns show the means of the four deviation distributions for the first and second minimum estimated error, respectively. Variance curves are on the companion web site. The trends are similar to those for LDA.

Figure 5 again considers LDA classification for the synthetic data with 10-fold cross-validation, two feature-selection schemes and sample sizes 60 and 120, except now the heterogeneous markers have been removed. The left and right columns show the means of the four deviation distributions for the first and second minimum estimated error, respectively. Once again, the trends are similar.

Figure 6 shows results analogous to those in Figures 2 and 3 for the real data using the two error-normalized deviation distributions, with LDA and 10-fold cross-validation. Similar results are observed. Finally, Figure 7 shows results for the real data for 3NN.

Reporting bias results from the imprecision of the individual error estimates across the samples. One might use a family of
samples with the idea that the analysis will show that a proposed classifier is superior for some populations out of those considered, but this approach leads to reporting bias because error estimation imprecision precludes any conclusions pertaining to individual samples. In particular, if a researcher only applies a proposed classification rule to a single small sample from a single population and obtains a superior error estimate, even if the estimator is unbiased, the root mean square (RMS) error between the true error and the error estimate is large (and typically unknown), which precludes any conclusion regarding the performance of the classifier. Letting $\epsilon_0$ and $\hat{\epsilon}_0$ denote the true and estimated errors, respectively, the RMS is given by $\sqrt{E[(\epsilon_0 - \hat{\epsilon}_0)^2]}$. The RMS can be decomposed in terms of the bias of the error estimator and the variance of the deviation, $\text{Var}[\epsilon_0 - \hat{\epsilon}_0]$. If we assume that the error estimator is
**Fig. 5.** Left column: means of four deviation distributions resulted from the distributions of $\epsilon_{i,\min}^{est}$ and $\epsilon_{i,\min}^{true}$; right column: means of four deviation distributions resulted from the distributions of $\epsilon_{i,\min}^{(2)}$ and $\epsilon_{i,\min}^{(2)}$. There are no heterogeneous markers in the distribution model. The classification scheme is composed of LDA classifier, $t$-test and $t$-test+SFS feature-selection methods and 10-fold cross-validation error estimation. The results are shown for five selected features.

approximately unbiased, then the problem with small samples is that the deviation variance is too large.

If one is going to use real data and samples are not large, then it is incumbent to apply the classification rule to samples from as many different populations as possible. This will not allow a claim of superiority on any particular population, but can lead to a claim of superiority across the populations. There is no reporting bias because

**Fig. 6.** Means of the two deviation distributions for the real data: $\delta_{i,\min}^{est}$ and $\delta_{i,\min}^{true}$ (left column); $\delta_{i,\min}^{(2)}$ and $\delta_{i,\min}^{(2)}$ (right column). The classification scheme is composed of LDA classifier, $t$-test and $t$-test+SFS feature-selection methods and 10-fold cross-validation error estimation. The results are shown for five selected features.

**Fig. 7.** Means of the two deviation distributions for the real data: $\delta_{i,\min}^{est}$ and $\delta_{i,\min}^{true}$ (left column); $\delta_{i,\min}^{(2)}$ and $\delta_{i,\min}^{(2)}$ (right column). The classification scheme is composed of 3NN classifier, $t$-test and $t$-test+SFS feature-selection methods and 5-fold cross-validation error estimation. The results are shown for five selected features.
there is no sample selection according to performance. Specifically, consider \( m \) population distributions \( F_1, F_2, \ldots, F_m \) corresponding to samples \( S_1, S_2, \ldots, S_m \) of sample sizes \( n_1, n_2, \ldots, n_m \), respectively. Let the classification errors on \( F_1, F_2, \ldots, F_m \) be \( e_1, e_2, \ldots, e_m \) and the error estimates obtained from the samples be \( \hat{e}_1, \hat{e}_2, \ldots, \hat{e}_m \). Rather than considering the RMS of each estimate individually, we can consider the RMS of the average estimate, 
\[
\hat{\epsilon} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \hat{e}_i^2},
\]
and the estimate, 
\[
\hat{\epsilon}_m = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \epsilon_i^2},
\]
of the average error. If we assume the estimates are approximately unbiased (and independent), then the RMS of the average is determined by variance of the deviation 
\[
\hat{\epsilon} = \epsilon,
\]
which is given by 
\[
\text{Var}[\hat{\epsilon} - \epsilon] = \frac{1}{m^2} \sum_{i=1}^{m} \text{Var}[\hat{e}_i - \epsilon_i].
\]
This variance is substantially less than the individual deviation variances. Regarding Equation (17), note that simply throwing any available sample into the mix may not be beneficial. Each individual variance depends on the sample size and if one has a family \( \mathcal{G} \) of fairly large samples and one small sample \( \mathcal{T} \), the deviation variance for \( \mathcal{G} \) may be less than that for \( \mathcal{G} \cup \{\mathcal{T}\} \).

While using a large number of real data samples that yields a large optimistic bias if only the best performing samples are selected, Equation (17) shows that using a large number of samples can lead to a worthwhile conclusion regarding a family of populations. No doubt this makes the entire enterprise computationally burdensome, this is a price that must be paid. While at one time there were not many publicly available data sets, that situation has changed to the point where a journal could make available a standardized family of data sets and require that any real data classifier comparisons be made using this family. It may be that the measurement quality across the samples is not uniform; nevertheless, this is the kind of data that are currently available and it is appropriate to evaluate proposed classifiers on available data. The weakness of averaging across samples is that some classification rule might work better than another on individual populations, but without large samples, this kind of conclusion is precluded in any case by the small samples. This is an inherent weakness of using real data samples.

We have restricted ourselves to standard classifiers and feature-selection algorithms so that the experimental results are within the mainstream and not the result of some complex methods whose complexity results in extreme overfitting, whose properties have not been extensively studied, or whose computational complexity makes them problematic for the extensive simulations carried out in this study. Nonetheless, given the consistency of the results across the simulations, reporting bias must be taken into account unless otherwise demonstrated. One interested in the performance of a particular classification rule can apply the experimental methodology to that rule.

**REFERENCES**


