BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 23 2014, pages 3349–3355
doi:10.1093/bioinformatics/btu527

Data and text mining

Advance Access publication August 13, 2014

Cross-validation under separate sampling: strong bias and how to correct it

Ulisses M. Braga-Neto[1,2], Amin Zollanvari[1,3] and Edward R. Dougherty[1,2,*]

[1]Department of Electrical and Computer Engineering, [2]Center for Bioinformatics and Genomic Systems Engineering and [3]Department of Statistics, Texas A&M University, College Station, TX, 77843, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: It is commonly assumed in pattern recognition that cross-validation error estimation is ‘almost unbiased’ as long as the number of folds is not too small. While this is true for random sampling, it is not true with separate sampling, where the populations are independently sampled, which is a common situation in bioinformatics.

Results: We demonstrate, via analytical and numerical methods, that classical cross-validation can have strong bias under separate sampling, depending on the difference between the sampling ratios and the true population probabilities. We propose a new separate-sampling cross-validation error estimator, and prove that it satisfies an ‘almost unbiased’ theorem similar to that of random-sampling cross-validation. We present two case studies with previously published data, which show that the results can change drastically if the correct form of cross-validation is used.

Availability and implementation: The source code in C++, along with the Supplementary Materials, is available at: http://gsp.tamu.edu/Publications/supplementary/zollanvari13/.

Contact: ulisses@ece.tamu.edu

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

Received on December 30, 2013; revised on July 23, 2014; accepted on July 30, 2014

1 INTRODUCTION

The most important property of a classifier is its error rate (probability of misclassification) because the error rate quantifies the predictive capacity of the classifier. If the feature-label distribution is known, then the true error can be found exactly; however, in practice, the feature-label distribution is unknown and the error must be estimated. If the sample is small, then the estimation must be computed using the same data as that used for training the classifier. Perhaps the most commonly used training data-based classification error estimator is cross-validation. It has a long history going back to 1968 (Lachenbruch and Mickey, 1968). In its most basic form, the k-fold cross-validation error estimate, $\hat{\varepsilon}_n^{\mathrm{cv}(k)}$, for a sample of size n (it is assumed that k divides n) is computed by selecting randomly a partition of the sample into k data ‘folds’ (subsets), for each fold applying the classification rule on the data not in the fold, computing the error rate of the designed classifier on the left-out fold, and then averaging the resulting k error rates. When $k = n$, one gets the leave-one-out estimator, $\hat{\varepsilon}_n^{l}$.

Cross-validation’s salient good property is that, under random sampling, it can be proved (see Devroye *et al.*, 1996) that it is ‘almost unbiased’, in the sense that

$$E[\hat{\varepsilon}_n^{\mathrm{cv}(k)}] = E[\varepsilon_{n-n/k}], \tag{1}$$

where ε_n is the true error (probability of misclassification) of a classifier designed on a sample of size n. Hence, the bias is not too great as long as n/k is small. For leave-one-out, $E[\hat{\varepsilon}_n^{l}] = E[\varepsilon_{n-1}]$, and the estimator is essentially unbiased. The salient point motivating the present article is that (1) depends on the sampling being random, and that when sampling is not random, there can be severe bias.

The importance of bias for an arbitrary error estimator $\hat{\varepsilon}_n$ can also be gleaned from its role in the estimator *root-mean-square error*: $\mathrm{RMS}[\hat{\varepsilon}_n] = E[(\hat{\varepsilon}_n - \varepsilon_n)^2]^{1/2} = \sqrt{\mathrm{Bias}[\hat{\varepsilon}_n]^2 + \mathrm{Var}_{\mathrm{dev}}[\hat{\varepsilon}_n]}$, where $\mathrm{Bias}[\hat{\varepsilon}_n] = E[\hat{\varepsilon}_n - \varepsilon_n]$ and $\mathrm{Var}_{\mathrm{dev}}[\hat{\varepsilon}_n] = \mathrm{Var}[\hat{\varepsilon}_n - \varepsilon_n]$ (Braga-Neto and Dougherty, 2004). As mentioned previously, for classical cross-validation under random sampling, it follows from (1) that, if n/k is small, then $\mathrm{Bias}[\hat{\varepsilon}_n] \approx 0$, in which case $\mathrm{RMS}[\hat{\varepsilon}_n] \approx \mathrm{Var}_{\mathrm{dev}}^{1/2}[\hat{\varepsilon}_n]$. While the variance of CV is known to be large in small-sample cases (Braga-Neto and Dougherty, 2004; Glick, 1973), it will typically reduce to zero as $n \to \infty$ (Devroye *et al.*, 1996). However, the bias introduced by application of the classical CV estimator under nonrandom sampling will generally not approach zero as $n \to \infty$. The result is an inconsistent estimator, which is imprecise under arbitrarily large sample sizes.

Under random sampling, an independent and identically distributed (i.i.d.) sample S is drawn from the *mixture* of the populations Π_0 and Π_1. This means that if a sample of size n is drawn for binary classification, then the numbers of sample points n_0 and n_1 drawn from the populations Π_0 and Π_1, respectively, are random variables $n_0 \sim \mathrm{Binomial}(n, c)$ and $n_1 \sim \mathrm{Binomial}(n, 1-c)$, where $c = P(Y=0)$ is the a priori probability that the label Y is zero, i.e. the sample point comes from population Π_0. This random-sampling assumption is so pervasive that it is usually assumed without mention and in books is often stated at the outset and then forgotten. For instance, Duda *et al.* (2000) state, ‘In typical supervised pattern classification problems, the estimation of the prior probabilities presents no serious difficulties’. They are referring to the fact that the prior probability $c = \Pr(Y=0)$ can be consistently estimated by the sampling ratio, $\hat{c} = \frac{n_0}{n}$. This is simply Bernoulli’s Law of Large

*To whom correspondence should be addressed.

Numbers: $\frac{n_0}{n} \rightarrow c$ in probability. However, suppose the sampling is not random, in the sense that the ratios $r = \frac{n_0}{n}$ and $1 - r = \frac{n_1}{n}$ are chosen before the sampling procedure. In this *separate-sampling* case, $S = S_0 \cup S_1$, where the sample points in S_0 and S_1 are selected randomly from Π_0 and Π_1, but given n, the individual class counts n_0 and n_1 are not determined by the sampling procedure. With separate sampling, we have no sensible estimate of c. Recognition of this particular problem of estimating the prior probability when sampling is separate and its effect on linear discriminant analysis (LDA) goes back to 1951 (Anderson, 1951). Often, one says that for separate sampling the ratios $r = \frac{n_0}{n}$ and $1 - r = \frac{n_1}{n}$ are chosen 'prior to' the sampling procedure. But there is in fact no temporal meaning to this. For instance, one could simply separately randomly sample Π_0 and Π_1 with n_0 and n_1 being randomly selected by a process independent of the sampling procedure, and the sampling would still be separate. The point is that r cannot be reasonably used as an estimate of c.

Figure 1 (taken from Esfahani and Dougherty, 2014) illustrates the effects of separate sampling on the expected true classifier error for two classification rules and multivariate Gaussian distributions of equal and unequal covariance structures and dimensionality $d = 3$. For a given sample size n, sampling ratio r, and classification rule, the expected true error rate $E[\varepsilon_n|r]$ is plotted for different class prior probabilities c, for LDA and a non-linear radial basis function support vector machine (RBF-SVM). For each r and n, n_0 is determined as $n_0 = \lceil nr \rceil$. We observe that the expected error is close to minimal when $r = c$ and that it can greatly increase when $r \neq c$. This kind of poor performance for separate sampling ratios not close to c is commonplace (Esfahani and Dougherty, 2014).

In this article, we investigate the effect of separate sampling on cross-validation error estimation. We will see that for a separate-sampling ratio r not close to c there can be large bias, optimistic or pessimistic. A serious consequence of this behavior can be ascertained by looking at Figure 1. Whereas the expected true error of the designed classifier grows large when r greatly deviates from c, a large optimistic cross-validation bias when r is far from c can obscure the large error and leave one with the illusion of good performance—and this illusion is not mitigated by large samples! To overcome the bias problem for classical cross-validation with separate sampling, we introduce a new cross-validation estimator designed for separate sampling and prove that it satisfies a bias property analogous to (1).

2 SYSTEMS AND METHODS

2.1 Discriminant analysis

We treat classification via discriminants to facilitate demarcation of the individual contributions of the class-conditional distributions to the error analysis. A sample-based *discriminant* is defined as a (measurable) function $W_n : S \mapsto \Re$, where from the definition, we see that we actually have a *family* of discriminants indexed by n. A discriminant W_n defines a classification rule via

$$\Psi_n(S)(X) = \begin{cases} 1, & W_n(S, X) \leq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

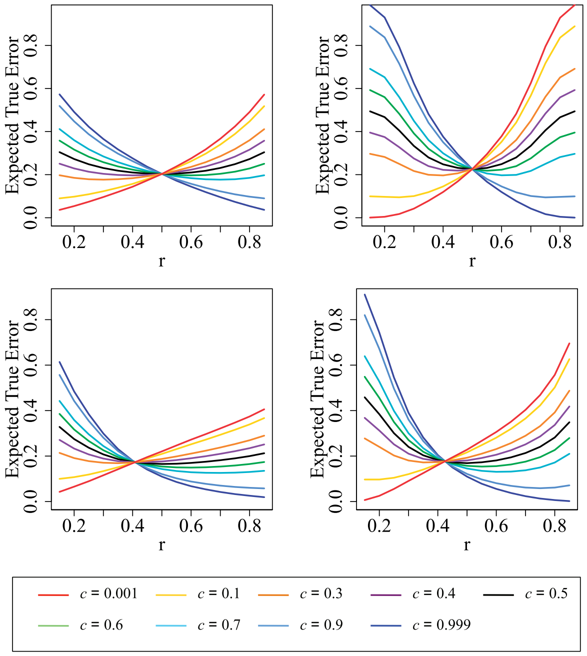

Fig. 1. Expected true error, $E[\varepsilon_n|r]$ as a function of r for LDA (left column) and an RBF-SVM classifier (right column) using synthetic data. Top row: $n = n_0 + n_1 = 80$ and equal covariance matrices; bottom row: $n = n_0 + n_1 = 80$ and unequal covariance matrices

where X comes from either Π_0 or Π_1. Because any classification rule Ψ_n can be expressed as a discriminant via $W_n(S, X) = I_{\Psi_n(S)(X)=0} - I_{\Psi_n(S)(X)=1}$, where I_A is the indicator function, discriminant analysis is completely general. We assume a common sense property of discriminants, that the order of the sample points within a sample does not matter.

With separate sampling, there are two *separate* samples $S_0 = \{X_1, \ldots, X_{n_0}\}$ and $S_1 = \{X_{n_0+1}, \ldots, X_{n_0+n_1}\}$ from populations Π_0 and Π_1, respectively. To demarcate the separate-sampling case from the random-sampling case, we will write the discriminant and corresponding classifier by $W_{n_0,n_1}(S_0, S_1, X)$ and $\Psi_{n_0,n_1}(S_0, S_1, X)$, respectively, with the latter defined in the same manner as (2) with $W_{n_0,n_1}(S_0, S_1, X)$ replacing $W_n(S, X)$.

The true classification error with random sampling is given by

$$\varepsilon_n = c\varepsilon_n^0 + (1 - c)\varepsilon_n^1, \tag{3}$$

where

$$\begin{aligned} \varepsilon_n^0 &= P(W_n(S, X) \leq 0 | X \in \Pi_0, S), \\ \varepsilon_n^1 &= P(W_n(S, X) > 0 | X \in \Pi_1, S), \end{aligned} \tag{4}$$

are the population-specific error rates. For separate sampling, the classification error is given by

$$\varepsilon_{n_0,n_1} = c\varepsilon_{n_0,n_1}^0 + (1 - c)\varepsilon_{n_0,n_1}^1, \tag{5}$$

where

$$\begin{aligned} \varepsilon_{n_0,n_1}^0 &= P(W_{n_0,n_1}(S_0, S_1, X) \leq 0 | X \in \Pi_0, S_0, S_1), \\ \varepsilon_{n_0,n_1}^1 &= P(W_{n_0,n_1}(S_0, S_1, X) > 0 | X \in \Pi_1, S_0, S_1). \end{aligned} \tag{6}$$

are the population-specific error rates.

2.2 Classical cross-validation error estimation

For $U \subset \{1, \dots, n\}$, let $S^{(U)}$ denote the sample S with the points indexed by U deleted, and define

$$W_n^{(U)}(S, X) = W_{n-m}(S^{(U)}, X)\,, \tag{7}$$

where $|U| = m$ is the size of U. Now let k divide n and consider a (random) partition $\{U_i; i = 1, \dots, k\}$ of $\{1, \dots, n\}$. Then the classical k-fold cross-validation estimator is given by

$$\hat{\varepsilon}_n^{\,\mathrm{cv}(k)} = \frac{1}{n} \sum_{i=1}^{k} \sum_{q \in U_i} (I_{W_n^{(U_i)}(S, X_q) \le 0} I_{Y_q = 0} + I_{W_n^{(U_i)}(S, X_q) > 0} I_{Y_q = 1})\,. \tag{8}$$

If $k = n$, this reduces to the leave-one-out estimator

$$\hat{\varepsilon}_n^{\,l} = \frac{1}{n} \sum_{i=1}^{n} (I_{W_n^{(i)}(S, X_i) \le 0}\, I_{Y_i = 0} + I_{W_n^{(i)}(S, X_i) > 0}\, I_{Y_i = 1}), \tag{9}$$

where we have omitted the braces around the singleton index set $\{i\}$.

Using the classical definition of cross-validation, (1) does not hold with separate sampling, in general. To demonstrate this, let $N_0 = \sum_{i=1}^{n} I_{Y_i = 0}$ be the (random) number of points from population Π_0 in the sample S; the expected cross-validation error rate under separate sampling is $E[\hat{\varepsilon}_n^{\,\mathrm{cv}(k)} | N_0 = n_0]$. For simplicity, we consider leave-one-out cross-validation. From (9),

$$\begin{aligned} E[\hat{\varepsilon}_n^{\,l} | N_0 = n_0] = {} & \frac{n_0}{n}\, P(W_n^{(1)}(S, X_1) \le 0 | Y_1 = 0, N_0 = n_0) \\ & + \frac{n_1}{n}\, P(W_n^{(1)}(S, X_1) > 0 | Y_1 = 1, N_0 = n_0) \\ = {} & \frac{n_0}{n}\, P(W_{n_0-1,n_1}(S_0^{(1)}, S_1, X) \le 0 | X \in \Pi_0) \\ & + \frac{n_1}{n}\, P(W_{n_0,n_1-1}(S_0, S_1^{(n_0+1)}, X) > 0 | X \in \Pi_1) \\ = {} & \frac{n_0}{n}\, E[\varepsilon^0_{n_0-1,n_1}] + \frac{n_1}{n}\, E[\varepsilon^1_{n_0,n_1-1}]\,. \end{aligned} \tag{10}$$

On the other hand, it follows from (3) that

$$\begin{aligned} E[\varepsilon_{n-1} | N_0 = n_0] &= c\, E[\varepsilon^0_{n-1} | N_0 = n_0] \\ &\quad + (1-c) E[\varepsilon^1_{n-1} | N_0 = n_0] \\ &= c\, E[\varepsilon^0_{n_0,n_1-1}] + (1-c) E[\varepsilon^1_{n_0,n_1-1}]\,. \end{aligned} \tag{11}$$

2.3 Cross-validation for separate sampling

To adapt cross-validation to separate sampling, let $U \subset \{1, \dots, n_0\}$, let $V \subset \{n_0+1, \dots, n_0+n_1\}$, let $S_0^{(U)}$ and $S_1^{(V)}$ denote the samples S_0 and S_1, with the points indexed by U and V deleted, respectively, and define

$$W_{n_0,n_1}^{(U,V)}(S_0, S_1, X) = W_{n_0-m,n_1-l}(S^{(U)}, S^{(V)}, X)\,. \tag{12}$$

where $|U| = m$ and $|V| = l$ are the sizes of U and V, respectively. Now let k_0 divide n_0 and k_1 divide n_1, and consider (random) partitions $\{U_i; i = 1, \dots, k_0\}$ of $\{1, \dots, n_0\}$ and $\{V_i; i = 1, \dots, k_1\}$ of $\{n_0+1, \dots, n_0+n_1\}$. *Separate-sampling* (k_0, k_1)*-fold cross-validation estimators* are defined by

$$\begin{aligned} \hat{\varepsilon}_{n_0,n_1}^{\,\mathrm{cv}(k_0,k_1),0} &= \frac{1}{n_0\, k_1} \sum_{i=1}^{k_0} \sum_{j=1}^{k_1} \sum_{r \in U_i} I_{W_{n_0,n_1}^{(U_i,V_j)}(S_0, S_1, X_r) \le 0}\,, \\ \hat{\varepsilon}_{n_0,n_1}^{\,\mathrm{cv}(k_0,k_1),1} &= \frac{1}{n_1 k_0} \sum_{i=1}^{k_0} \sum_{j=1}^{k_1} \sum_{r \in V_j} I_{W_{n_0,n_1}^{(U_i,V_j)}(S_0, S_1, X_r) > 0}\,. \end{aligned} \tag{13}$$

These are estimators of the population-specific true errors $\varepsilon^0_{n_0,n_1}$ and $\varepsilon^1_{n_0,n_1}$, respectively. One may use a convex combination of the previous estimators to yield a separate-sampling cross-validation estimator of the overall true error rate:

$$\hat{\varepsilon}_{n_0,n_1}^{\,\mathrm{cv}(k_0,k_1)} = c\, \hat{\varepsilon}_{n_0,n_1}^{\,\mathrm{cv}(k_0,k_1),0} + (1-c)\, \hat{\varepsilon}_{n_0,n_1}^{\,\mathrm{cv}(k_0,k_1),1}\,. \tag{14}$$

If c is known (or known to a high degree of accuracy), then one can use it in (14). If c is unknown, then there is no proper cross-validation estimator of the overall error rate.

If $k_0 = n_0$ and $k_1 = n_1$, then the (k_0, k_1)-fold cross-validation estimators defined previously reduce to *separate-sampling leave-one-out estimators*:

$$\begin{aligned} \hat{\varepsilon}_{n_0,n_1}^{\,l,0} &= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I_{W_{n_0,n_1}^{(i,j)}(S_0, S_1, X_i) \le 0}\,, \\ \hat{\varepsilon}_{n_0,n_1}^{\,l,1} &= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I_{W_{n_0,n_1}^{(i,j)}(S_0, S_1, X_{n_0+j}) > 0}\,. \end{aligned} \tag{15}$$

A convex combination of these yields a separate-sampling leave-one-out estimator of the overall true error rate:

$$\hat{\varepsilon}_{n_0,n_1}^{\,l} = c\, \hat{\varepsilon}_{n_0,n_1}^{\,l,0} + (1-c)\, \hat{\varepsilon}_{n_0,n_1}^{\,l,1}\,. \tag{16}$$

Again, in the absence of knowledge of c, no proper estimator of the overall error rate is possible.

We now show that a version of (1) holds for the separate-sampling cross-validation estimator.

Theorem. The cross-validation estimator in (14) satisfies

$$E[\hat{\varepsilon}_{n_0,n_1}^{\,\mathrm{cv}(k_0,k_1)}] = E[\varepsilon_{n_0-n_0/k_0,n_1-n_1/k_1}]\,. \tag{17}$$

Proof. First notice that if $U \subset \{1, \dots, n_0\}$ and $V \subset \{n_0+1, \dots, n_0+n_1\}$, with $|U| = m$ and $|V| = l$, then

$$\begin{aligned} & W_{n_0,n_1}^{(U,V)}(S_0, S_1, X_i) \sim \\ & W_{n_0-m,n_1-l}(S_0, S_1, X) | X \in \Pi_0,\ i \in U, \\ & W_{n_0,n_1}^{(U,V)}(S_0, S_1, X_i) \sim \\ & W_{n_0-m,n_1-l}(S_0, S_1, X) | X \in \Pi_1,\ i \in V. \end{aligned}$$

Therefore, from (13),

$$\begin{aligned} E[\hat{\varepsilon}_{n_0,n_1}^{\,cv(k_0,k_1),0}] &= E[\varepsilon^0_{n_0-n_0/k_0,n_1-n_1/k_1}], \\ E[\hat{\varepsilon}_{n_0,n_1}^{\,cv(k_0,k_1),1}] &= E[\varepsilon^1_{n_0-n_0/k_0,n_1-n_1/k_1}]\,. \end{aligned} \tag{18}$$

Finally,

$$\begin{aligned} E[\hat{\varepsilon}_{n_0,n_1}^{\,cv(k_0,k_1)}] &= c E[\hat{\epsilon}_{n_0,n_1}^{\,cv(k_0,k_1),0}] + (1-c) E[\hat{\epsilon}_{n_0,n_1}^{\,cv(k_0,k_1),1}] \\ &= c\, E[\varepsilon^0_{n_0-n_0/k_0,n_1-m_1/k_1}] + (1-c) E[\varepsilon^1_{n_0-n_0/k_0,n_1-n_1/k_1}] \\ &= E[\varepsilon_{n_0-n_0/k_0,n_1-m_1/k_1}]\,. \qquad \square \end{aligned}$$

In the case of the separate-sampling leave-one-out estimator defined in (16), the preceding theorem reduces to

$$E[\hat{\varepsilon}_{n_0,n_1}^{\,l}] = E[\varepsilon_{n_0-1,n_1-1}]\,. \tag{19}$$

3 RESULTS AND DISCUSSION

3.1 Simulation study with synthetic and real data

We have performed a set of experiments using both synthetic models and real data to examine the behavior of classical and separate-sampling cross-validation under separate sampling. Throughout we use 5-fold cross-validation. We consider four well-known classification rules: LDA, Quadratic Discriminant Analysis (QDA), Linear Support Vector Machine (L-SVM) and RBF-SVM (see the Supplementary Material for definitions of these classification rules).

To generate synthetic data, we use a model with class-conditional 3-dimensional Gaussian distributions, $N(\boldsymbol{\mu}_y, \Sigma_y)$, $y = 0, 1$, where $\boldsymbol{\mu}_0 = [0, 0, \ldots, 0, 0]$, $\boldsymbol{\mu}_1 = [0, 0, \ldots, 0, \theta]$ and Σ_y has σ^2 on the diagonal and ρ_y off the diagonal. The pair (ρ_0, ρ_1) can take on the values (0.8, 0.8) or (0.8, 0.4). We set θ so that the Mahalanobis distance between the classes for equal covariance matrices and the Bhattacharyya distance between the classes for unequal covariance matrices is 3. We consider $n = 80$ and $n = 1000$, so that we can compare small-sample and large-sample results.

We consider four public microarray real datasets: pediatric acute lymphoblastic leukemia (ALL; Yeoh *et al.*, 2002), acute myeloid leukemia (AML; Valk *et al.*, 2004), multiple myeloma (Zhan *et al.*, 2006) and breast cancer (Desmedt *et al.*, 2007). Table 1 provides a summary of these real datasets, including the total number of features and sample size. For a detailed description of the data preparation, the readers are referred to the Supplementary Materials. The experiments on real data are essentially similar to those on synthetic data except that in real data experiments we use *t*-test feature selection to reduce the dimensionality to $d = 3$. In real data experiments, we consider only $n = 80$, which allows sufficient data for holdout error estimation.

All experiments are performed for a range of $r = \frac{n_0}{n} \in [0.15, 0.85]$. We fix n and determine n_0 according to $n_0 = \lceil nr \rceil$. At each iteration, S_0 and S_1 are randomly picked from either a synthetic model or real data to train the classifier and compute the two cross-validation estimates. Finding the bias requires knowing the true error, which is estimated on 5000 independent sample points from the synthetic distributions, or held out points in the case of real data; however, owing to separate sampling the ordinary holdout method cannot be applied, and we use separate-sampling holdout as explained by Esfahani and Dougherty (2014). We consider $c = 0.001, 0.1, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 0.999$. For each classification rule, we repeat the process of obtaining the true error and its estimates 4000 times for each value of r and c to obtain a distribution of estimates and true errors from which to compute the bias.

Table 1. Microarray studies used in this study

Dataset	Description	Features	n_0/n_1
(Desmedt *et al.*, 2007)	Breast cancer	22 215	98/77
(Yeoh *et al.*, 2002)	Pediatric ALL	5077	149/99
(Valk *et al.*, 2004)	AML	22 215	116/157
(Zhan *et al.*, 2006)	Multiple myeloma	54 613	156/78

In Figure 2, we provide the results for the synthetic data with unequal covariance matrices $[(\rho_0, \rho_1) = (0.8, 0.4)]$ for $n = 80$, 1000 and for two of the real datasets (Desmedt *et al.*, 2007; Valk *et al.*, 2004). The complete set of results is given in the Supplementary Material. In the figure, from left to right, the columns correspond to LDA, QDA, L-SVM and RBF-SVM, respectively. The top two rows of the figure correspond to the real data from Desmedt *et al.* (2007) and Valk *et al.* (2004), and the third and fourth rows correspond to the synthetic data with $n = 80$ and $n = 1000$. The x-axis corresponds to the sampling ratio r, the y-axis gives the bias, the solid lines are for the proposed separate-sampling cross-validation, the dashed lines are for classical cross-validation, and the colors code the value of c.

The trends are consistent across all experiments (including those in the Supplementary Material): (i) for classical cross-validation with c near 0.5, there is significant optimistic bias for large $|r - c|$; (ii) for classical cross-validation with small or large c, there is optimistic bias for large $|r - c|$ and pessimistic bias for small $|r - c|$ as long as $|r - c|$ is not very close to 0; (iii) for separate-sampling cross-validation, estimation is slightly optimistic and almost unbiased across the range of $|r - c|$. Combined with the results of Esfahani and Dougherty (2014), the bias behavior of classical cross-validation is especially harmful for large $|r - c|$ because it masks the increase in classifier error that occurs for large $|r - c|$, as shown in Figure 1. Furthermore, although the deviation variance of classical cross-validation can be mitigated by large samples, the bias issue generally remains just as bad for large samples.

3.2 Two case studies

To further illustrate the effects of separate sampling on classical cross-validation bias, we consider two published studies. The first (Ambroise and McLachlan, 2002) uses a colon microarray dataset containing gene-expression measurements taken from 2000 genes for 62 tissue specimens, 40 tumorous tissues (class 0) and 22 normal tissues (class 1). Using the SVM-RFE classification rule (Guyon *et al.*, 2002), the authors split the data into a training and a test set, each including 31 specimens, by sampling without replacement, such that the training data contain 20 tumorous and 11 normal specimens. They compare the 10-fold cross-validation error using (8) to the standard holdout estimate obtained by counting the errors on the test set. But the standard holdout estimate is unbiased under random sampling, not separate sampling. For the latter, holdout estimation must take into account the value of c to be unbiased (Esfahani and Dougherty, 2014). Assuming the classifier is applied to the US population, based on the incidence rate of colorectal cancer among the US population, which is 40/100 000 (Haggar and Boushey, 2009), $c = 40/100\,000$. The black solid and dotted curves in Figure 3 resemble the curves plotted in Figure 1 of Ambroise and McLachlan (2002). The gray solid and dashed curves are obtained by considering the cross-validation scheme (14) and computing the true error from (3). The error bars refer to the 95% confidence interval. All curves show the averaged error and estimated error obtained on 200 random splits of the data as mentioned above. These curves show that regardless of the number of genes considered in the classifier, using the classical cross-validation (8) induces ~13% optimistic bias with respect to the true

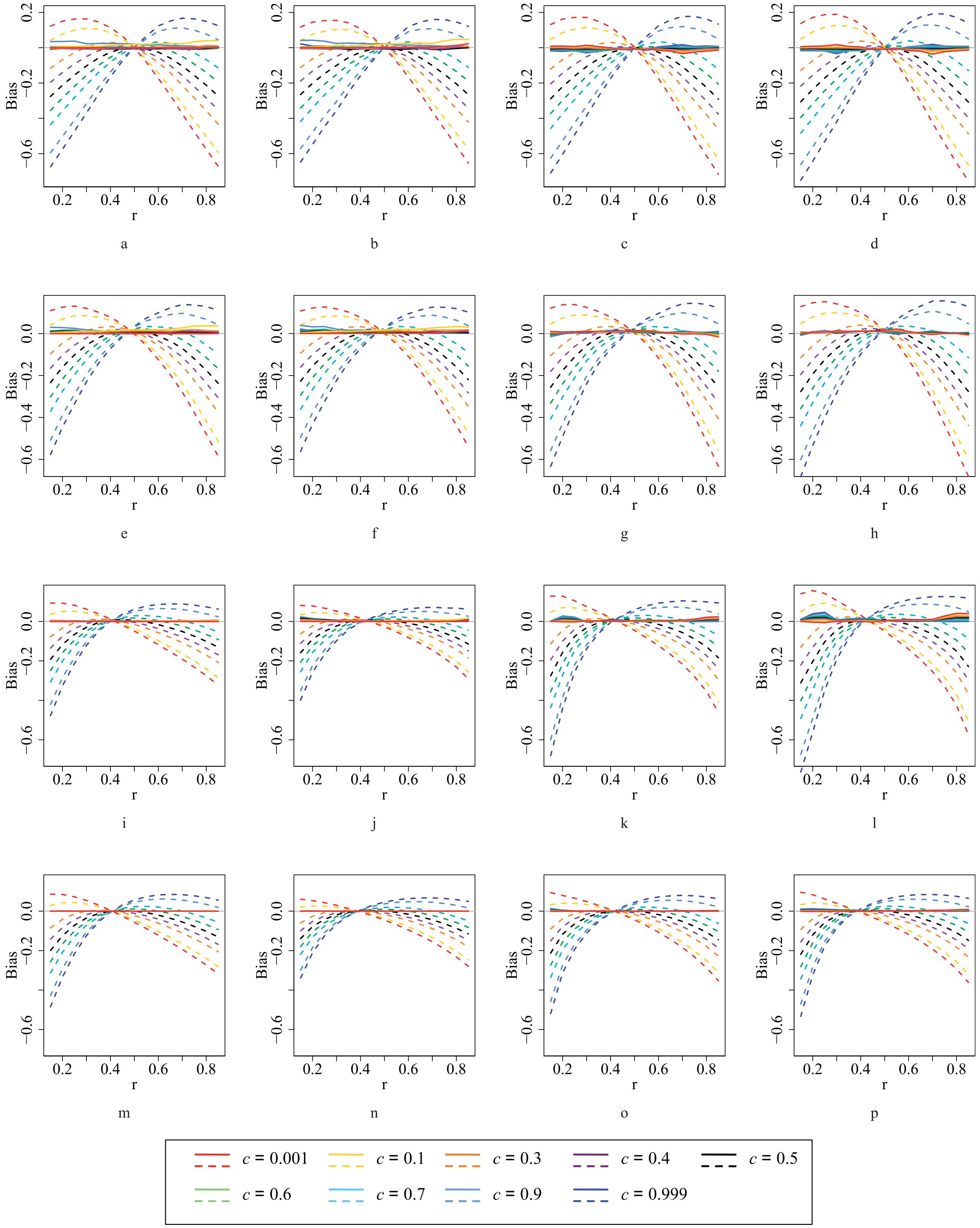

Fig. 2. Bias of 5-fold CV as a function of r for LDA, QDA, L-SVM and RBF-SVM classifiers using real and synthetic data. Rows from top to bottom: data taken from Desmedt *et al.* (2007); data taken from Valk *et al.* (2004); synthetic data $n=n_0+n_1=80$ and $\Sigma_0 \neq \Sigma_1$; synthetic data $n=n_0+n_1=1000$ and $\Sigma_0 \neq \Sigma_1$; Columns from left to right: LDA; QDA; L-SVM; RBF-SVM. Dashed curves: regular cross-validation scheme. Solid curve: new scheme of cross-validation. For each r and n, n_0 is determined as $n_0=\lceil nr\rceil$

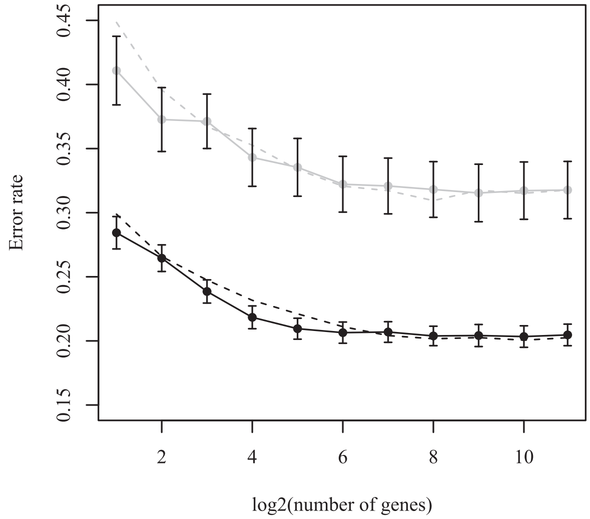

Fig. 3. The black solid and dotted curves resemble the expected true error and cross-validation error rates reported in Figure 1 of Ambroise and McLachlan (2002). The gray solid and dotted curves are the expected true error and estimated error by using cross-validation scheme (14) when the prior probability of colon cancer is set to be the incidence rate across the USA. All curves show the averaged error and estimated error obtained on 200 random splits of the data

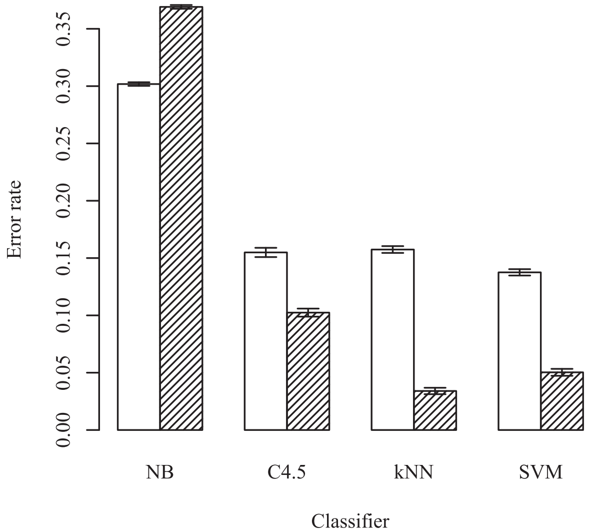

Fig. 4. The white bars are the expected classical cross-validation error rates on the Parkinson's dataset used by Kaya *et al.*, (2011) for four classifiers. The shaded bars are the estimated error rates by using cross-validation scheme (14) when the prior probability of Parkinson's disease is set to be the incidence rate across the USA. The bars show the averaged estimated error obtained on 200 samplings of the data

error, while the proposed cross-validation scheme is almost unbiased.

In the second case study, we use the Parkinson's dataset used by Kaya *et al.* (2011). This dataset contains 22 biomedical voice features and 195 measurements in which 48 belong to individuals with Parkinson (class 0) and 147 measurements are taken from healthy individuals (class 1), so that $r = 0.246$. The authors use this dataset to construct classifiers for diagnosis of Parkinson's disease based on distorted voice features. Four classifiers are constructed: naive Bayes (NB; Friedman *et al.*, 1997), C4.5 (Dietterich, 2000), kNN (k = 5) (Devroye *et al.*, 1996) and RBF-SVM. Although Kaya *et al.* (2011) have reported the estimated classical cross-validation error on a single sample of the data, we repeat the sampling procedure 200 times to get an estimate of the expected cross-validation error using both the classical (8) and the corrected cross-validation scheme (14). We assume the prior probability c of Parkinson's disease is determined by the incidence rate of Parkinson's disease across the USA, which is 13.4/100 000 (Van Den Eeden *et al.*, 2003). In Figure 4, the white bars are the expected classical cross-validation error rates; the shaded bars are the estimated error rates using the separate-sampling cross-validation scheme. The bars show the averaged estimated error obtained on 200 samplings of the data. The behavior observed in Figure 2 makes it plausible that the error estimates for classical cross-validation will exceed those of separate-sampling cross-validation, which is nearly unbiased. This is true in all cases except for NB. However, if we look carefully at Figure 2, we see that the point at which the bias becomes optimistic (for increasing r) can be well left of 0.5. This point is affected by the covariance structure and the classification rule. In this case, for NB, it is to the left of 0.246.

4 CONCLUDING REMARKS

We show in this article that classical cross-validation may display substantial bias when it is applied in the separate sampling scenario, which is common in biomedical studies. If one wishes to use cross-validation with separate sampling, then one should use the separate-sampling version of cross-validation, which is proposed here, or else, significant bias may result. This means that one must know the prior probability c (at least a good approximation of it). A similar requirement was made by Esfahani and Dougherty (2014) to ensure proper performance of the classification rule. Using a sampling ratio significantly different from c will result in poor classifier design and, often, optimistic bias to obscure the poor design. As concluded by Esfahani and Dougherty (2014), given the ubiquity of separate sampling in biomedicine, although it would incur some cost, it would behoove the medical community to gather population statistics so that accurate estimates of prior class probabilities would be available. In the absence of such statistics, separate sampling should not be used.

Funding: This work was supported by NSF award CCF-0845407 (Braga-Neto) and NIH grant 2R25CA090301 (Nutrition, Biostatistics and Bioinformatics) from the National Cancer Institute.

Conflict of interest: none declared.

REFERENCES

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Anderson,T. (1951) Classification by multivariate analysis. *Psychometrika*, **16**, 31–50.

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

Desmedt,C. *et al.* (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.

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

Dietterich,T.G. (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach. Learn.*, **40**, 139–157.

Duda,R.O. *et al.* (2000) *Pattern Classification*. Wiley, New York.

Esfahani,M.S. and Dougherty,E.R. (2014) Effect of separate sampling on classification accuracy. *Bioinformatics*, **30**, 242–250.

Friedman,N. *et al.* (1997) Bayesian network classifiers. *Mach. Learn.*, **29**, 131–163.

Glick,N. (1973) Sample-based multinomial classification. *Biometrics*, **29**, 241–256.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines in machine learning. *Mach. Learn.*, **46**, 389–422.

Haggar,F.A. and Boushey,R.P. (2009) Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin. Colon. Rectal. Surg.*, **22**, 191–197.

Kaya,E. *et al.* (2011) Effect of discretization method on the diagnosis of parkinsons disease. *Int. J. Innov. Comput. Inf.*, **7**, 4669–4678.

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

Valk,P.J. *et al.* (2004) Prognostically useful gene-expression profiles in acute myeloid leukemi. *N. Engl. J. Med.*, **350**, 1617–1628.

Van Den Eeden,S.K. *et al.* (2003) Incidence of parkinson's disease: variation by age, gender, and race/ethnicity. *Am. J. Epidemiol.*, **157**, 1015–1022.

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.

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