Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 548906, 12 pages
doi:10.1155/2010/548906

Research Article

Small-Sample Error Estimation for Bagged Classification Rules

T. T. Vu and U. M. Braga-Neto

Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA

Correspondence should be addressed to U. M. Braga-Neto, ulisses@ece.tamu.edu

Received 2 April 2010; Accepted 16 July 2010

Academic Editor: Harri Lahdesmaki

Application of ensemble classification rules in genomics and proteomics has become increasingly common. However, the problem of error estimation for these classification rules, particularly for bagging under the small-sample settings prevalent in genomics and proteomics, is not well understood. Breiman proposed the "out-of-bag" method for estimating statistics of bagged classifiers, which was subsequently applied by other authors to estimate the classification error. In this paper, we give an explicit definition of the out-of-bag estimator that is intended to remove estimator bias, by formulating carefully how the error count is normalized. We also report the results of an extensive simulation study of bagging of common classification rules, including LDA, 3NN, and CART, applied on both synthetic and real patient data, corresponding to the use of common error estimators such as resubstitution, leave-one-out, cross-validation, basic bootstrap, bootstrap 632, bootstrap 632 plus, bolstering, semi-bolstering, in addition to the out-of-bag estimator. The results from the numerical experiments indicated that the performance of the out-of-bag estimator is very similar to that of leave-one-out; in particular, the out-of-bag estimator is slightly pessimistically biased. The performance of the other estimators is consistent with their performance with the corresponding single classifiers, as reported in other studies.

1. Introduction

Ensemble classification methods combine the decision of multiple classifiers designed on randomly perturbed versions of the available data [1–5]. The most popular version of this scheme is known as bootstrap aggregating, or "bagging" [4, 5] where the ensemble classifier corresponds to a majority vote among classifiers designed on bootstrap samples [6] from the available training data.

There has been considerable interest recently in the application of bagging in the classification of both gene expression data [7–10] and protein-abundance mass spectrometry data [11–16]. The popularity of bagging is based on the expectation that combining the decision of several classifiers will regularize and improve the performance of unstable, overfitting classification rules (the so-called "weak learners"). In a related study [17], the authors have investigated this claim, in the context of small-sample genomics and proteomics data. On the other hand, a different issue is the performance of error estimators for bagged classifiers. Accurate error estimation is a critical issue in Genomics, as it decisively impacts the scientific validity of hypotheses derived from application of pattern recognition methods to biomedical data [18–20]. On the topic of error estimation, Breiman proposed a general method, which he called "out-of-bag", for estimating statistics of bagged classifiers [21], and, subsequently, other authors applied it to the estimation of the classification error [22, 23]. In this paper, we give an explicit definition of the out-of-bag estimator that is intended to remove estimator bias, which is done by formulating carefully how the error count is normalized. The performance of out-of-bag estimators with general bagged classification rules is not in fact well understood, especially in connection with bagging ensemble classifiers derived from classification rules other than decision trees (which was Breiman's primary interest). In addition, to our knowledge, no studies have attempted to assess the performance of error estimators for bagged classifiers in the context of Genomics data, particularly in the prevalent small-sample setting usually found in these applications.

To investigate these issues, we conducted an extensive simulation study of bagging of common classification rules, including LDA, 3NN, and CART, applied on both synthetic and real patient data, corresponding

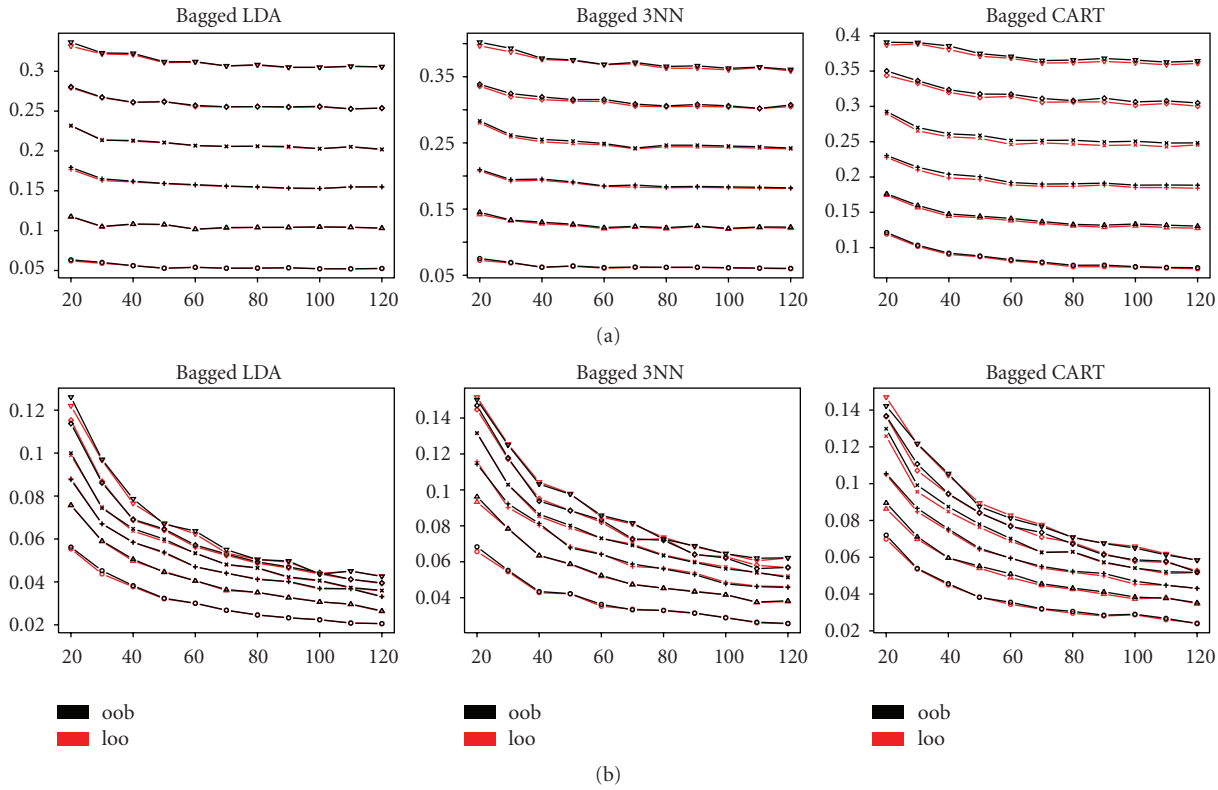

Figure 1: Comparison of out-of-bag and leave-one-out for different Gaussian models over the number of samples, for dimensionality $p = 2$. (a) Sample mean. (b) Sample standard deviation.

to the use of common error estimators such as resubstitution, leave-one-out, cross-validation, basic bootstrap, bootstrap 632, bootstrap 632 plus, bolstering, semibolstering, in addition to the out-of-bag estimator itself. We present here selected representative results; the full set of results can be found on the companion website, at http://gsp.tamu.edu/Publications/supplementary/oob. The results from the numerical experiments indicated that the performance of the out-of-bag error estimator is very similar to that of leave-one-out; in particular, the out-of-bag estimator is slightly pessimistically biased. The performance of the other estimators is for the most part consistent with their performance with the corresponding single classification rules assessed in other studies, with the best performance being provided by the bolstered error estimators, in terms of root mean square error.

This paper is organized as follows. In Section 2, we review briefly the definition of bagged ensemble classification rules. In Section 3, we describe the error estimators considered in this study. In Section 4, we present the results of a large simulation study on the performance of error estimators with bagged classification rules. Finally, Section 5 provides concluding remarks.

2. Bagged Classification Rules

In pattern recognition, classification is the process of assigning a group *label* to an object, based on information available about it in the form of a data vector called a *feature vector*. Suppose we have a binary classification problem with feature vector X in a feature space V and label $Y \in \{0, 1\}$. A classifier is a function $\psi : V \to \{0, 1\}$. The stochastic properties of the classification problem are completely determined by the *joint feature-label distribution* F of the pair (X, Y). F is, in practice, rarely known. Classification is implemented empirically, by means of the design of a *classifier* based on a finite set of i.i.d. samples drawn from F:

$$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}. \tag{1}$$

For a fixed n, a classification rule would be a function that maps the sample data to a classifier:

$$\Psi_n : [V \times \{0, 1\}]^n \longrightarrow \{0, 1\}^V. \tag{2}$$

For a given training set S_n, we have a designed classifier $\psi_n = \Psi_n(S_n)$. The classification error is the chance of incorrectly classifying a future sample (X, Y) given the training sample set S_n:

$$\varepsilon_n = P(\psi_n(X) \neq Y \mid S_n). \tag{3}$$

It is clear that ϵ_n is random as it depends on S_n. The expected error taken over the randomness of S_n is called expected classification error $E[\epsilon_n]$ and this is a deterministic quantity which is a function of classification rule and the joint feature-label distribution.

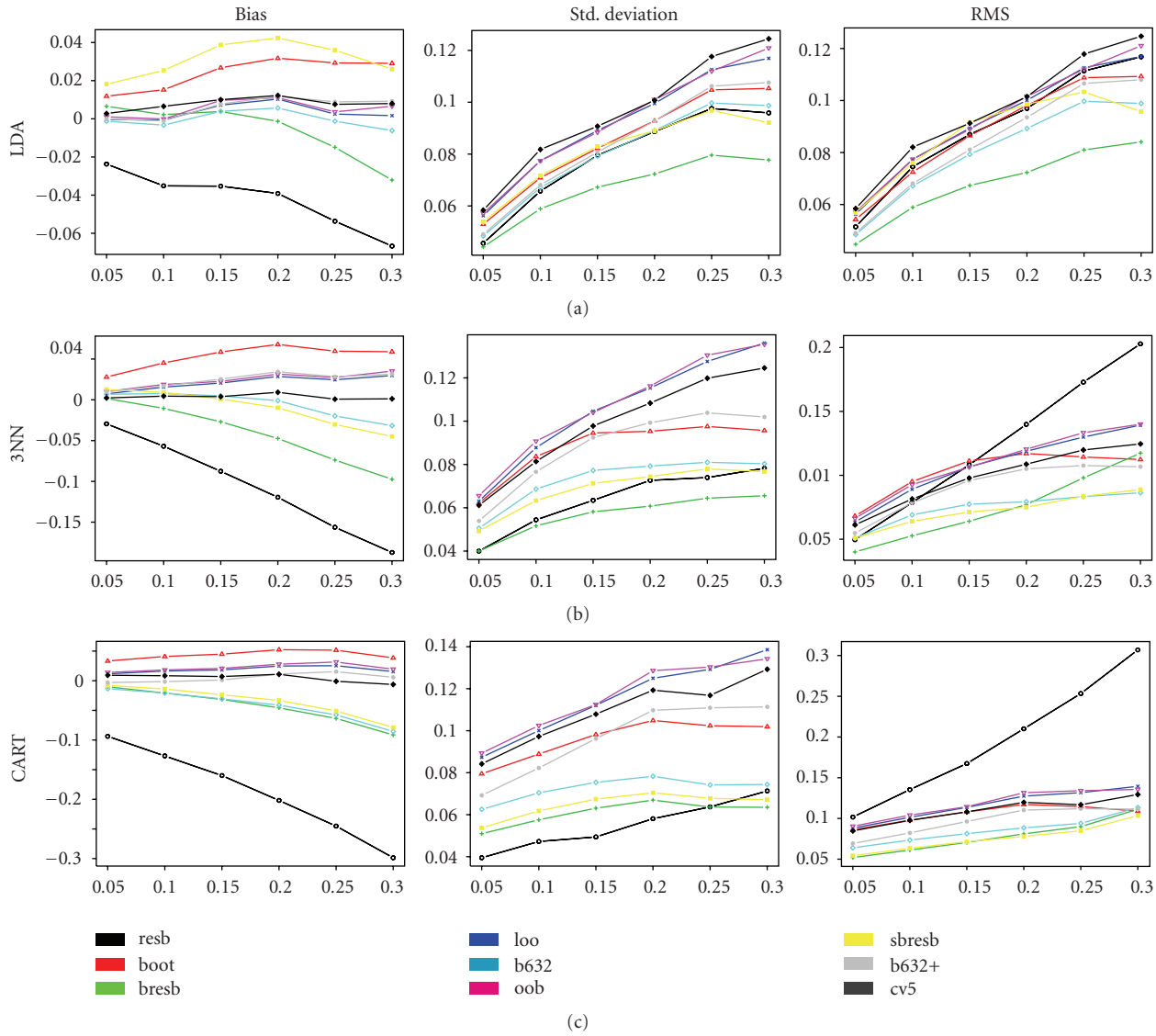

Figure 2: Bias, variance (standard deviation), and RMS of as a function of the Bayes error, for the synthetic data, sample size $n = 20$, and dimensionality $p = 2$, with different base classification rules.

The number of training samples n is, in practice, always limited. Much effort is spent on exploiting and reusing the samples as much as possible. Randomization is one resampling technique in which multiple bootstrap sets S_k^* are created by randomly drawing points from S_n, either with or without replacement, corresponding to a resampling distribution F^* on the training data. The cardinality k of S_k^* can be smaller, equal to or larger than n, depending on the application of interest. In a bootstrap set, a sample point can appear multiple times or not at all. In bagging, different choices of resampling distribution and k lead to variants, but the most common one is uniform resampling with $k = n$.

An ensemble classifier is acquired based on majority voting among component classifiers. Each component of the ensemble is built up on a bootstrap set using the original classification rule Ψ_n. The bagged classification is defined as

$$\psi_n^R(x) = \Psi_n^R(S_n)(x) = \begin{cases} 1, & E\left[\Psi_n\left(S_k^*\right)(x) \mid S_n\right] \geq \dfrac{1}{2}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where the expectation is taken with respect to the random mechanism F^*, fixed at the observed value of S_n. Bagging is a version of ensemble classifier, in which the expectation in (4) is approximated by Monte-Carlo sampling:

$$\psi_{n,m}^B(x) = \Psi_n^R(S_n)(x) = \begin{cases} 1, & \dfrac{1}{m}\displaystyle\sum_{j=1}^{m} \psi_n^{*(j)}(x) \geq \dfrac{1}{2}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where the classifiers $\psi_n^{*(j)}$ are designed by the original classification rule Ψ_n on bootstrap samples $S_n^{*(j)}$, for $j = 1, \ldots, m$, for large enough m. How large m should be is an important topic of bagging so that it is computationally

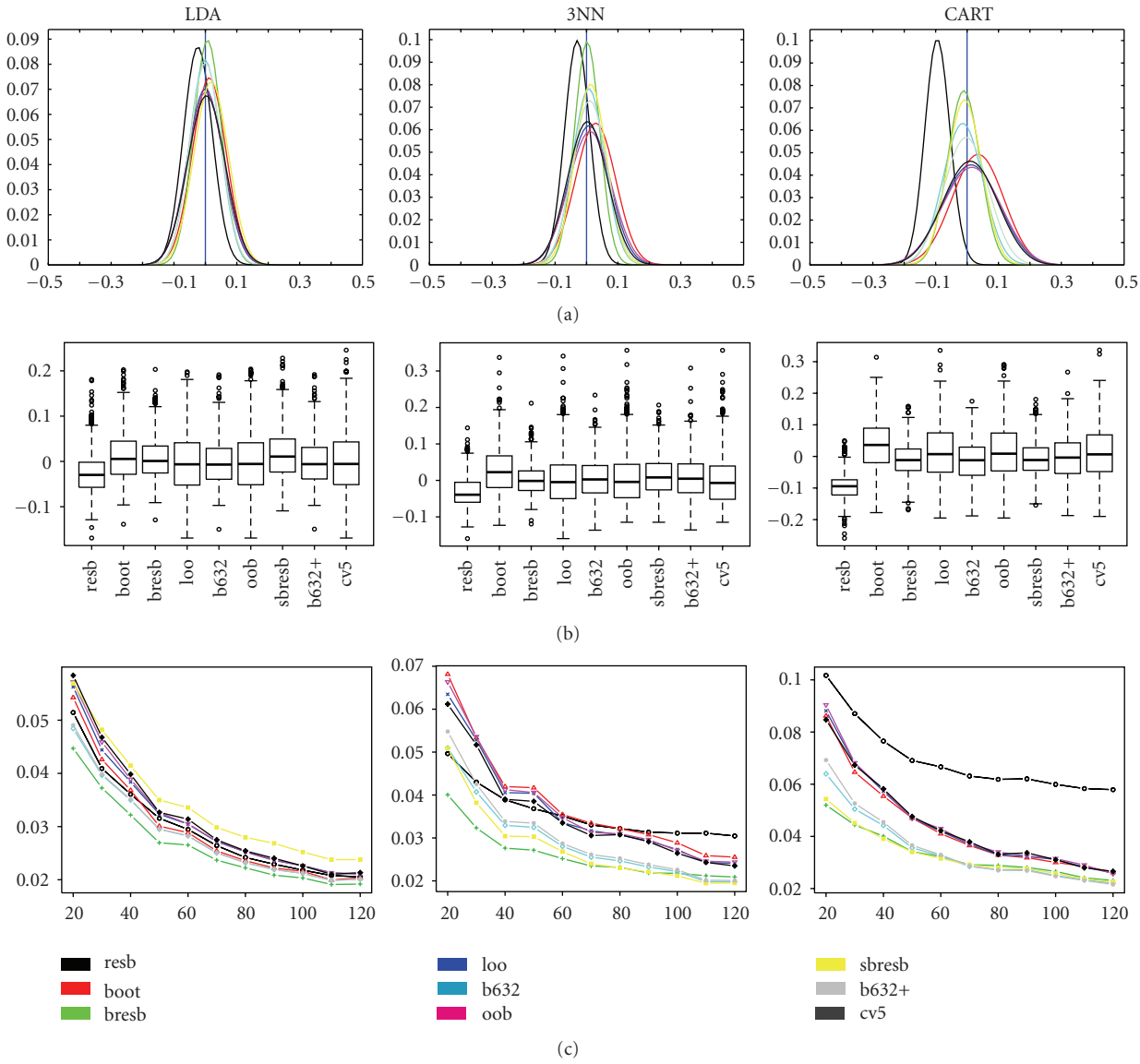

Figure 3: Empirical deviation distribution (a), box plots (b), and RMS as a function of sample size (c), for synthetic Gaussian model with Bayes error = 0.05, sample size $n = 20$, and dimensionality $p = 2$, with different base classification rules.

efficient and the Monte Carlo approximation is accurate enough. In this paper, we chose $m = 51$ according to the recommendation from Breiman [21] and from our observations on the convergence of mean error of bagged classifiers in our previous study [17]. It is important to select an odd m to avoid the issue of tie breaking in the majority vote. Experimental results in our previous study [17] showed that increasing m beyond $m = 51$ leads to negligible differences in performance.

3. Error Estimation

3.1. Classical Methods. Data in practice are often limited, and the training sample S_n has to be used for both designing the classifier ψ_n and estimating the true error ϵ_n. An obvious method to estimate ϵ_n is to use S_n itself as the test set, which leads often, but not always, to optimistic bias. This is called the *resubstitution* estimator:

$$\hat{\epsilon}_{\text{resub}} = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - \Psi_n(S_n)(x_i) \right|. \tag{6}$$

In *k-fold cross-validation*, S_n is partitioned into k *folds* $S_{(i)}$, for $i = 1, \ldots, k$ (for simplicity, we assume that k divides n), each fold is left out of the design process and used as a testing set, and the estimate is the overall proportion of error committed on all folds [24]:

$$\hat{\epsilon}_{\text{cvk}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n/k} \left| y_j^{(i)} - \Psi_n\left(S_n \setminus S_{(i)}\right)\left(x_j^{(i)}\right) \right|, \tag{7}$$

where $(x_j^{(i)}, y_j^{(i)})$ is a sample in the ith fold. The process may be repeated, where several cross-validated estimates are

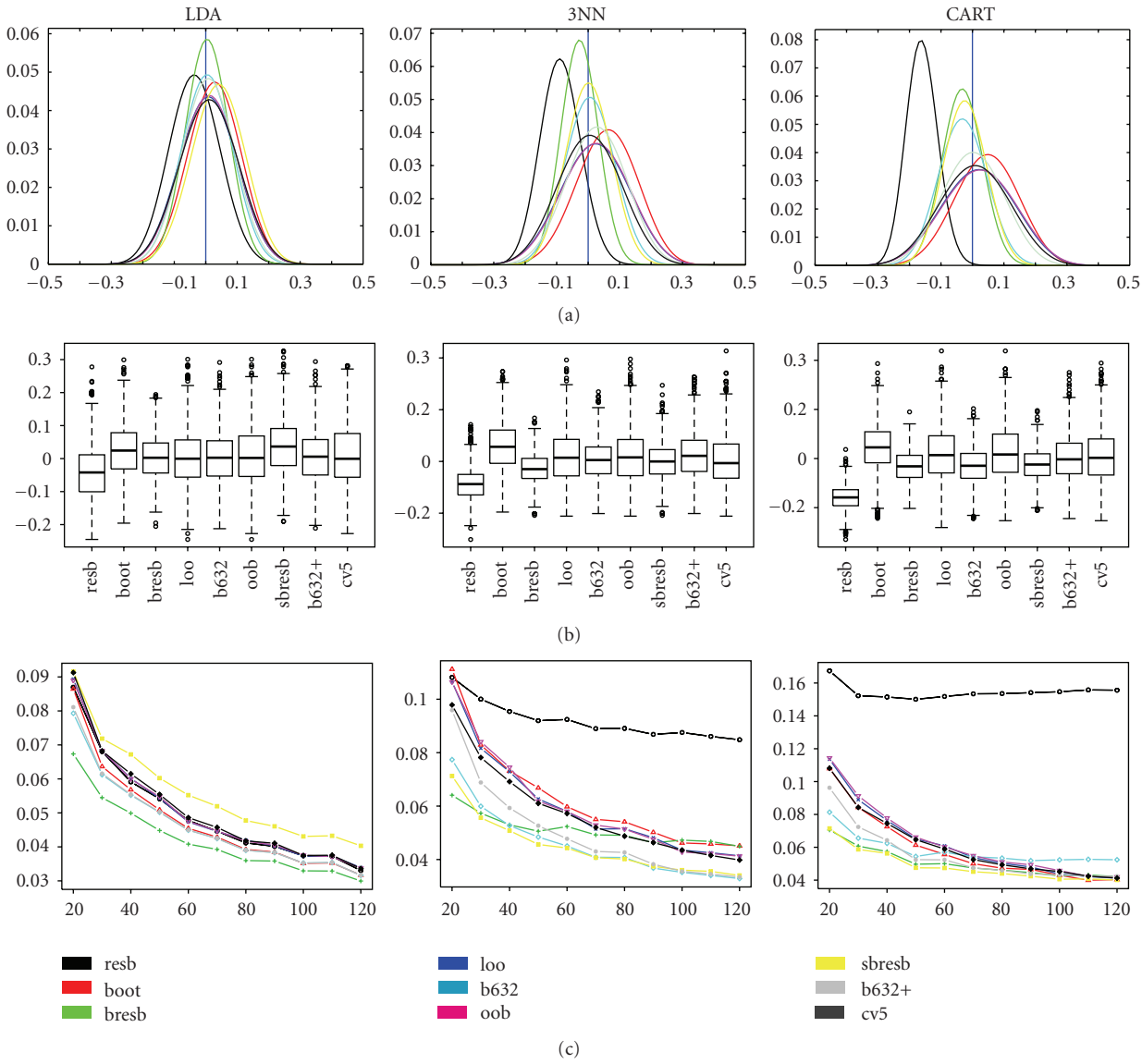

FIGURE 4: Empirical deviation distribution (a), box plots (b), and RMS as a function of sample size (c), for synthetic Gaussian model with Bayes error = 0.15, sample size $n = 20$, and dimensionality $p = 2$, with different base classification rules.

computed, using different partitions of the data into folds, and the results averaged. In *leave-one-out estimation*, a single observation is left out each time, which corresponds to n-fold cross-validation. The leave-one-out estimator is nearly unbiased as an estimator of $E[\epsilon_n]$.

3.2. Bootstrap Error Estimators. Resampling methodology, as mentioned above in generating ensemble classifiers, can be used for estimating errors. In fact, bootstrap error estimation was proposed by Efron [25], before its use in bagging. The actual proportion of times a data point (x_i, y_i) appears in a bootstrap sample S_n^* can be written as $P_i^* = (1/n)\Sigma_{j=1}^n I_{(x_j^*, y_j^*)=(x_i,y_i)}$, where $I_S = 1$ if the statement S is true, zero otherwise. The basic bootstrap (or "zero bootstrap") is given by

$$\widehat{\epsilon}_0 = \frac{\sum_{b=1}^{B}\sum_{i=1}^{n}\left|y_i - \Psi_n\left(S_n^{*b}\right)(x_i)\right| I_{P_i^{*b}=0}}{\sum_{b=1}^{B}\sum_{i=1}^{n} I_{P_i^{*b}=0}}, \tag{8}$$

With the number of bootstrap sample B being between 25 and 200, as recommended in [25]. Bootstrap 632 is a variant, which tries to correct the bias of the basic bootstrap estimator by performing an average with the resubstitution estimator [25]:

$$\widehat{\varepsilon}_{b632} = (1 - 0.632)\widehat{\varepsilon}_{\text{resub}} + 0.632\widehat{\varepsilon}_0. \tag{9}$$

Bootstrap 632 plus is another modified version of bootstrap, proposed in [26], which is intended for highly-overfitting classification rules. Bootstrap 632 plus attempts to adaptively find the weights in (9) that offset the effects of overfitting.

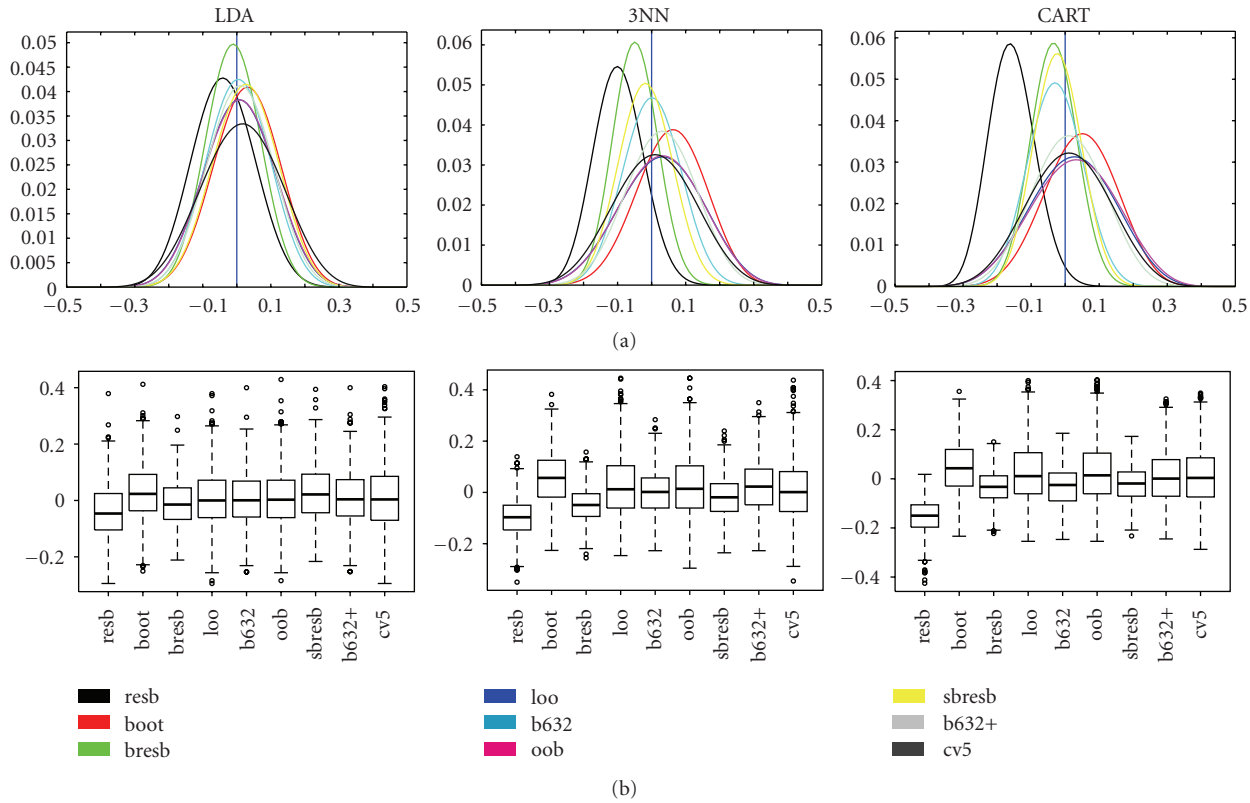

Figure 5: Empirical deviation distribution (a) and box plots (b), for breast cancer gene expression data, sample size $n = 20$, and dimensionality $p = 2$, with different base classification rules.

The weights depend on the *relative overfitting rate* R and *no-information error rate* γ. In dichotomous classification, R and γ are estimated from $\hat{p}_1$, the proportion of observed samples belonging to class 1 and $\hat{q}_1$, the proportion of classifier outputs belonging to class 1. The relations are as follows:

$$\begin{aligned}
\hat{\gamma} &= \hat{p}_1(1-\hat{q}_1) + \hat{q}_1(1-\hat{p}_1), \\
\hat{R} &= \frac{\hat{\varepsilon}_0 - \hat{\varepsilon}_{\text{resub}}}{\hat{\gamma} - \hat{\varepsilon}_{\text{resub}}}, \\
\hat{w} &= \frac{.632}{1 - .368\hat{R}}, \\
\hat{\varepsilon}_{b632+} &= (1-\hat{w})\hat{\varepsilon}_{\text{resub}} + \hat{w}\hat{\varepsilon}_0.
\end{aligned} \tag{10}$$

3.3. Bolstered Error Estimators. Bolstered estimation was proposed in [27]. It has shown promising performance for small sample sizes in terms of root mean square error. While it is comparable to bootstrap methods in many cases, bolstered estimators are typically much more computationally efficient than the bootstrap. The main idea of bolstering is to put a kernel at each of the sample point, called "bolstering kernel" to smooth the variance of counting-based estimation methods (in this paper, we adopt Gaussian bolstering kernels). When the classifiers are overfitted, and hence, resubstitution estimates are optimistically biased, then bolstering at a misclassified point will increase this bias. Semibolstering is suggested for correcting this, by conducting no bolstering at misclassified points. We refer the reader to [27] for the full details (in this paper, we employ the bolstered and semibolstered resubstitution estimators of [27]).

3.4. Out-of-Bag Error Estimators. Breiman [21] originally proposed the out-of-bag method to estimate the generalization error of bagged predictors of CART and the node priority probabilities. Bylander [22] later did a simulation study comparing out-of-bag and cross-validation for tree classification C4.5 and concluded that both are biased. Banfield et al. [23] used out-of-bag in a large simulation of investigating performances of a variety of ensemble methods. Martínez-Muñoz and Suárez [28], in an attempt to find the optimal number of components of ensembles, employed out-of-bag as the optimization criterion. Despite that, the properties of the out-of-bag estimator remain largely unclear, in particular, the issue of bias. We propose in the sequel a modification to the standard out-of-bag estimator that removes nearly all of its bias (as evidenced by the numerical experiments in Section 4).

In bagging, component classifiers are designed based on bootstrap sets, each of which contain on average 63% of the original sample set. Hence, there are approximately 37% of the data which are not used to build the classifier and are therefore uncorrelated with it. Out-of-bag estimates are obtained by testing the majority voting classifier via those individual classifiers in the ensemble that are uncorrelated

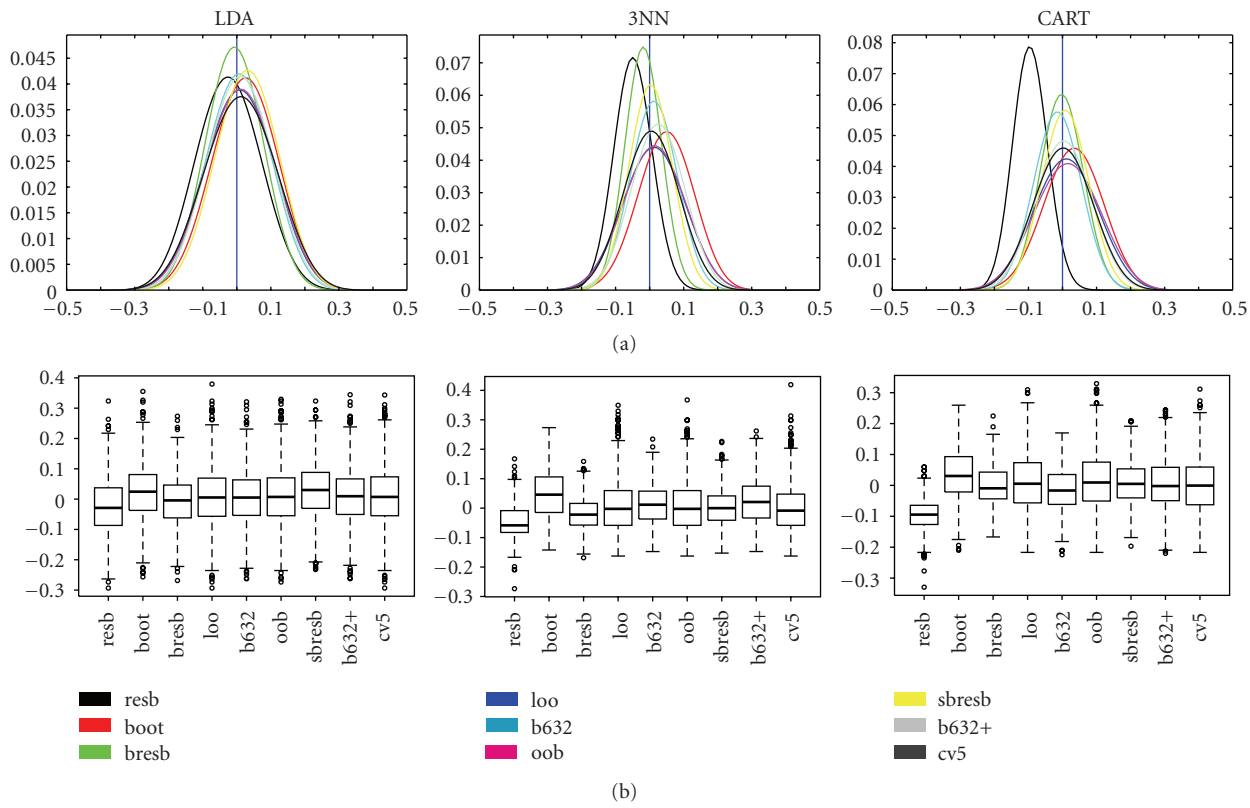

FIGURE 6: Empirical deviation distribution (a) and box plots (b), for lung cancer gene expression data, sample size $n = 20$, and dimensionality $p = 2$, with different base classification rules.

with the testing point, that is, those classifiers whose training sets do not contain the testing points. Suppose we resample the original sample set k times, leading to k bootstrap sample sets S^{*j}. Let $P_i^j = 1$ if sample i appears in the bootstrap sample S^{*j}, and $P_i^j = 0$, otherwise, for $i = 1, \ldots, n$. Denote

$$\begin{aligned} A_0(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{Y_i=0\}}, \\ B_0(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{\Psi_n(S^{*j})(X_i)=1\}} I_{\{Y_i=0\}}, \\ A_1(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{Y_i=1\}}, \\ B_1(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{\Psi_n(S^{*j})(X_i)=0\}} I_{\{Y_i=1\}}, \end{aligned} \tag{11}$$

for $i = 1, \ldots, n$. Notice that $A_m(i)$ is equal to the number of times that sample i in class m appears across all bootstrap sample sets, while $B_m(i)$ is equal to the number of times that sample i in class m appears and is *misclassified* across all bootstrap sample sets. Then the out-of-bag error estimator, as proposed by Breiman in [4], can be written as

$$\hat{\varepsilon}_{\text{oob}} = \frac{1}{n} \sum_{i=1}^{n} \left[I_{\{B_0(i) \geq A_0(i)/2\}} I_{\{A_0(i)>0\}} + I_{\{B_1(i) \geq A_1(i)/2\}} I_{\{A_1(i)>0\}} \right]. \tag{12}$$

The estimator, as formulated above, will be optimistically biased, in general, according to the following rationale. Clearly, when $Y_i = j$ and $A_j(i) = 0$, then the ith sample point belongs to all of the bootstrap samples, so there are no individual classifiers to test on the ith point. In other words, the "out-of-bag ensemble" of classifiers for that point is empty in this case. That means that, with training sample size of n, we often have fewer than n samples to perform the out-of-bag estimation. In computing the proportion of incorrect classification by the ensemble, one should therefore divide not by n as in (12), but rather by n minus the number of times when the out-of-bag ensembles are empty, which leads to the following modified out-of-bag estimator:

$$\hat{\varepsilon}_{\text{oob}}^{m} = \frac{1}{n - \sum_{i=1}^{n} \left[I_{\{A_0(i)=0\}} + I_{\{A_1(i)=0\}} \right]} \times \sum_{i=1}^{n} \left[I_{\{B_0(i) \geq A_0(i)/2\}} I_{\{A_0(i)>0\}} + I_{\{B_1(i) \geq A_1(i)/2\}} I_{\{A_1(i)>0\}} \right]. \tag{13}$$

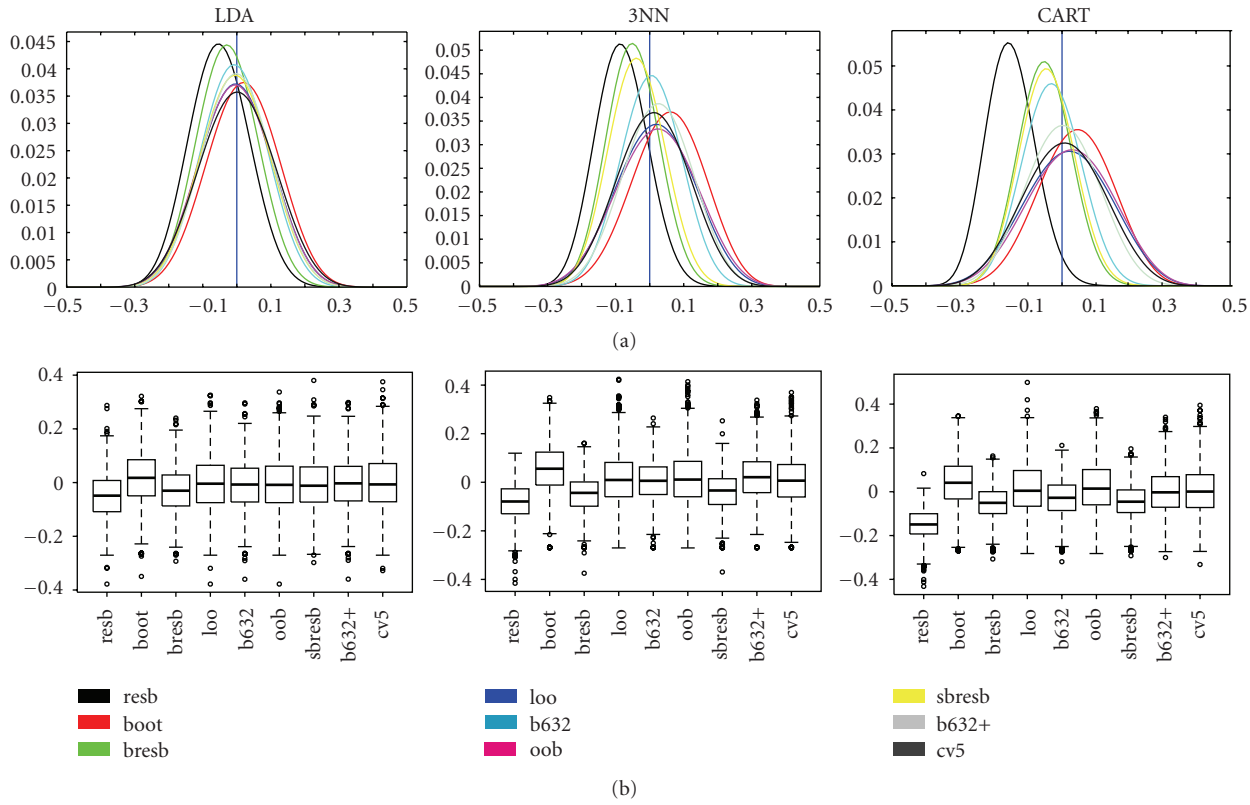

Figure 7: Empirical deviation distribution (a) and box plots (b), for prostate cancer mass-spectrometry data, sample size $n = 20$, and dimensionality $p = 2$, with different base classification rules.

As shown by the numerical results in Section 4, this estimator has approximately the bias of leave-one-out; that is, it is only slightly pessimistically biased. As far as we know, this formulation of the out-of-bag estimator has not been explicitly given in the literature.

4. Simulation Study

This section reports the results of an extensive simulation study, which were conducted on both synthetic and publicly available microarray data and protein abundance mass spectrometry data. We present here selected representative results; the full set of results can be found on the companion website, at http://gsp.tamu.edu/Publications/supplementary/oob. We simulated bagged ensembles of linear discriminant analysis (LDA), 3-nearest neighbors (3NN), and decision trees (CART) [24], and computed actual and estimated errors, according to the different estimation methods. These estimators were evaluated based on the distribution of their deviation from the true error, and in terms of bias, variance, and root mean square (RMS) errors.

4.1. Methods. We compared the performances of estimators for varying number of training samples with different dimensions of the feature space. The dimensionality and number of samples are selected to be compatible with a small-sample scenario (in this paper, the dimensionality is kept fixed at $p = 2$). For patient data, a small number of features (once again, $p = 2$ in this paper) are first selected by the t-test. We afterwards randomly draw a number of samples to be used as the training set and employed the rest as a testing set. The number of training points are chosen to be small to keep the small sample setting, and to have a large enough testing set. This was repeated 1000 times to get the empirical deviation distribution [18], that is, the distribution of estimated minus actual errors, for the different error estimators. The results are presented in forms of beta-fit curves, box-plots, and bias, variance, and RMS curves in order to provide as detailed as possible a picture of the empirical deviation distributions of the error estimators.

4.2. Simulation Based on Synthetic Data. We employ here the spherical Gaussian model, where the covariance matrix is identity and the two mean vectors are symmetric over the origin. With that assumption, we varied the Bayes error of the model by changing the distance between the two means. Models with different Bayes errors and dimension are compared over varying number of samples. The feature-label distribution is known and this allows us to exactly compute the true error of the designed classifier, which is then used to derive the empirical deviation distribution for the different estimators.

Table 1: Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for breast cancer gene expression data, and dimensionality $p = 2$.

Rule	n	stat	resb	boot	bresb	loo	b632	oob	sbresb	b632plus	cv5
lda	20	bias	−0.0388	0.0287	−0.0104	0.0063	0.0039	0.0076	0.0244	0.0092	0.0143
		sd	0.0908	0.0944	0.0789	0.1004	0.0912	0.1003	0.0933	0.0938	0.1140
		rms	0.0988	0.0986	0.0795	0.1006	0.0913	0.1006	0.0964	0.0942	0.1149
lda	40	bias	−0.0198	0.0082	−0.0084	−0.0012	−0.0021	0.0002	0.0168	−0.0011	−0.0044
		sd	0.0657	0.0642	0.0614	0.0671	0.0638	0.0673	0.0676	0.0641	0.0714
		rms	0.0686	0.0647	0.0620	0.0671	0.0639	0.0673	0.0696	0.0641	0.0716
lda	60	bias	−0.0157	−0.0000	−0.0097	−0.0045	−0.0058	−0.0036	0.0104	−0.0054	−0.0011
		sd	0.0577	0.0559	0.0544	0.0580	0.0560	0.0581	0.0586	0.0560	0.0586
		rms	0.0598	0.0559	0.0553	0.0582	0.0563	0.0582	0.0595	0.0563	0.0587
cart	20	bias	−0.1554	0.0456	−0.0330	0.0226	−0.0284	0.0267	−0.0225	0.0096	0.0094
		sd	0.0653	0.1047	0.0671	0.1210	0.0798	0.1229	0.0700	0.1059	0.1187
		rms	0.1686	0.1142	0.0747	0.1231	0.0847	0.1258	0.0735	0.1063	0.1190
cart	40	bias	−0.1583	0.0323	−0.0358	0.0095	−0.0378	0.0143	−0.0284	−0.0094	0.0058
		sd	0.0484	0.0697	0.0502	0.0774	0.0533	0.0799	0.0516	0.0671	0.0810
		rms	0.1655	0.0769	0.0616	0.0780	0.0653	0.0812	0.0589	0.0677	0.0812
cart	60	bias	−0.1722	0.0211	−0.0377	0.0001	−0.0501	0.0043	−0.0317	−0.0232	−0.0050
		sd	0.0400	0.0624	0.0473	0.0705	0.0473	0.0701	0.0472	0.0590	0.0695
		rms	0.1768	0.0658	0.0605	0.0705	0.0689	0.0703	0.0569	0.0634	0.0697
3nn	20	bias	−0.0964	0.0575	−0.0478	0.0270	0.0009	0.0269	−0.0176	0.0273	0.0076
		sd	0.0716	0.0996	0.0649	0.1174	0.0835	0.1167	0.0778	0.1005	0.1156
		rms	0.1201	0.1150	0.0806	0.1204	0.0835	0.1197	0.0798	0.1041	0.1159
3nn	40	bias	−0.0952	0.0406	−0.0481	0.0109	−0.0094	0.0139	−0.0214	0.0075	0.0036
		sd	0.0529	0.0687	0.0493	0.0787	0.0590	0.0785	0.0577	0.0669	0.0801
		rms	0.1089	0.0798	0.0689	0.0794	0.0598	0.0797	0.0615	0.0673	0.0802
3nn	60	bias	−0.0962	0.0316	−0.0504	0.0034	−0.0154	0.0054	−0.0261	−0.0012	−0.0008
		sd	0.0432	0.0625	0.0452	0.0693	0.0526	0.0693	0.0514	0.0595	0.0680
		rms	0.1054	0.0701	0.0677	0.0694	0.0548	0.0695	0.0576	0.0595	0.0680

4.3. Simulation Based on Patient Data. We utilized the following publicly available data sets from published studies in order to study the performance of bagging in the context of genomics and proteomics applications.

4.3.1. Breast Cancer Gene Expression Data. These data come from the breast cancer classification study in [29], which analyzed $N = 295$ gene-expression microarrays containing a total of 25760 transcripts each. Filter-based feature selection was performed on a 70-gene prognosis profile, previously published by the same authors in [30]. Classification is between the good-prognosis class (115 samples), and the poor-prognosis class (180 samples), where prognosis is determined retrospectively in terms of survivability [29].

4.3.2. Lung Cancer Gene Expression Data. We employed here the data set "A" from the study in [31] on nonsmall cell lung carcinomas (NSCLC) that analyzed $N = 186$ gene expression microarrays containing a total of 12600 transcripts each. NSCLC is subclassified as adenocarcinomas, squamous cell carcinomas, and large-cell carcinomas, of which adenocarcinomas are the most common subtypes and of interest to classify from other subtypes of NSCLC. Classification is thus between adenocarcinomas (139 samples) and nonadenocarcinomas (47 samples).

4.3.3. Prostate Cancer Mass Spectrometry Data. Given the recent keen interest on deriving serum-based proteomic biomarkers for the diagnosis of cancer [32], we also included in this study data from a proteomic study of prostate cancer reported in [33]. It consists of SELDI-TOF mass spectrometry of $N = 326$ samples, which yield mass spectra for 45000 n/z (mass over charge) values. Filter-based feature selection is employed to find the top discriminatory n/z values to be used in the experiment. Classification is between prostate cancer patients (167 samples) and nonprostate patients, including benign prostatic hyperplasia and healthy patients (159 samples). We use the raw spectra values, without baseline subtraction, as we found that this leads to better classification rates.

Table 2: Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for lung cancer gene expression data, and dimensionality $p = 2$.

Rule	n	stat	resb	boot	bresb	loo	b632	oob	sbresb	b632plus	cv5
lda	20	bias	−0.0243	0.0238	−0.0070	0.0075	0.0061	0.0103	0.0294	0.0094	0.0106
		sd	0.0938	0.0938	0.0827	0.0989	0.0923	0.0988	0.0910	0.0932	0.1025
		rms	0.0969	0.0967	0.0830	0.0992	0.0925	0.0993	0.0956	0.0937	0.1031
lda	40	bias	−0.0118	0.0109	0.0012	0.0017	0.0025	0.0044	0.0273	0.0033	0.0045
		sd	0.0675	0.0655	0.0628	0.0684	0.0656	0.0685	0.0652	0.0656	0.0694
		rms	0.0685	0.0664	0.0628	0.0684	0.0657	0.0686	0.0707	0.0657	0.0695
lda	60	bias	−0.0092	0.0067	0.0023	−0.0004	0.0009	0.0015	0.0235	0.0012	0.0020
		sd	0.0606	0.0587	0.0570	0.0608	0.0590	0.0608	0.0586	0.0590	0.0610
		rms	0.0613	0.0591	0.0570	0.0608	0.0591	0.0609	0.0632	0.0590	0.0610
cart	20	bias	−0.0945	0.0321	−0.0025	0.0100	−0.0145	0.0139	0.0076	0.0031	0.0017
		sd	0.0502	0.0852	0.0623	0.0916	0.0683	0.0945	0.0676	0.0811	0.0849
		rms	0.1069	0.0911	0.0623	0.0921	0.0699	0.0955	0.0681	0.0812	0.0849
cart	40	bias	−0.0926	0.0226	−0.0230	0.0071	−0.0198	0.0088	−0.0141	−0.0071	0.0022
		sd	0.0384	0.0630	0.0439	0.0694	0.0504	0.0705	0.0472	0.0577	0.0654
		rms	0.1003	0.0670	0.0496	0.0698	0.0542	0.0710	0.0493	0.0581	0.0655
cart	60	bias	−0.0938	0.0202	−0.0277	0.0043	−0.0218	0.0068	−0.0210	−0.0103	0.0012
		sd	0.0335	0.0544	0.0397	0.0590	0.0438	0.0597	0.0414	0.0496	0.0571
		rms	0.0996	0.0580	0.0484	0.0592	0.0490	0.0601	0.0464	0.0507	0.0571
3nn	20	bias	−0.0483	0.0474	−0.0185	0.0114	0.0122	0.0132	0.0027	0.0238	0.0040
		sd	0.0552	0.0803	0.0529	0.0876	0.0677	0.0870	0.0623	0.0765	0.0787
		rms	0.0734	0.0932	0.0561	0.0884	0.0688	0.0880	0.0624	0.0802	0.0788
3nn	40	bias	−0.0489	0.0236	−0.0270	0.0043	−0.0031	0.0055	−0.0094	0.0027	−0.0004
		sd	0.0435	0.0602	0.0411	0.0626	0.0519	0.0624	0.0484	0.0555	0.0593
		rms	0.0655	0.0646	0.0492	0.0627	0.0520	0.0626	0.0493	0.0555	0.0593
3nn	60	bias	−0.0500	0.0198	−0.0317	0.0031	−0.0059	0.0036	−0.0147	−0.0009	−0.0028
		sd	0.0381	0.0526	0.0383	0.0555	0.0459	0.0553	0.0439	0.0486	0.0514
		rms	0.0629	0.0562	0.0497	0.0556	0.0462	0.0555	0.0463	0.0486	0.0514

4.4. Results and Discussion

4.4.1. Synthetic Data. The various error estimators can be grouped into four groups according to performance. The first group corresponds to resubstitution, which showed to be optimistically biased for the bagged LDA, 3NN, and CART classifiers, with a root mean square error that increases substantially with increasing Bayes error; resubstitution had been previously known to behave as such for single LDA, 3NN, and CART classifiers. The second group contains leave-one-out, fivefold cross-validation and out-of-bag. As we can see from Figure 1, the out-of-bag estimator, with the formulation given in (13), is almost identical to leave-one-out. This second group shows very small bias but considerably high variance. The resemblance of out-of-bag to cross-validation, which had been pointed out already in [22], is explained by the similar way of partitioning the sample set. This group shows much smaller bias than resubstitution, and this is consistent as the Bayes error increases. However, this group displayed larger variability than resubstitution and the bootstrap group, as we already knew from [19] on single classification rules. The third group includes the basic bootstrap, bootstrap 632, and bootstrap 632 plus; this group displays very competitive performance in terms of root mean square error. Even though they often perform better than the two previous groups, the estimators in this group took the longest time to compute across all experiments. The last group consists of the bolstered and semibolstered error estimators, which exhibit superior performance to the other groups, in terms of RMS error, despite being far less computationally expensive than cross-validation and bootstrap estimators.

Generally, for a fixed model, almost all the estimates work better when the sample size increases and this holds for all three bagged classifiers. In Figure 2, we see that there is a consistent trend; as the Bayes error increases or, equivalently, the classification problem becomes harder, error estimation performance decreases steadily, in term of root mean square error; this is true for all error estimation methods. Bolstered error estimators showed consistent superior performance to the others, in terms of accuracy (RMS) and computational cost. These conclusions are also supported by Figures 3 and 4.

Table 3: Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for prostate cancer mass-spectrometry data, and dimensionality $p = 2$.

Rule	n	stat	resb	boot	bresb	loo	b632	oob	sbresb	b632plus	cv5
lda	20	bias	−0.0506	0.0181	−0.0277	−0.0033	−0.0072	−0.0044	−0.0050	−0.0019	0.0006
		sd	0.0871	0.1025	0.0879	0.1031	0.0949	0.1037	0.0993	0.0985	0.1071
		rms	0.1007	0.1041	0.0921	0.1031	0.0951	0.1038	0.0994	0.0985	0.1071
lda	40	bias	−0.0283	0.0079	−0.0189	−0.0051	−0.0054	−0.0042	−0.0029	−0.0039	−0.0031
		sd	0.0609	0.0688	0.0626	0.0673	0.0647	0.0683	0.0674	0.0655	0.0693
		rms	0.0672	0.0693	0.0654	0.0675	0.0649	0.0684	0.0675	0.0656	0.0694
lda	60	bias	−0.0192	0.0045	−0.0141	−0.0042	−0.0042	−0.0044	−0.0008	−0.0035	−0.0017
		sd	0.0514	0.0572	0.0524	0.0542	0.0542	0.0549	0.0559	0.0546	0.0577
		rms	0.0549	0.0573	0.0542	0.0544	0.0544	0.0550	0.0560	0.0547	0.0577
cart	20	bias	−0.1504	0.0409	−0.0500	0.0164	−0.0295	0.0248	−0.0441	0.0014	0.0059
		sd	0.0693	0.1082	0.0765	0.1198	0.0847	0.1223	0.0791	0.1053	0.1169
		rms	0.1655	0.1157	0.0914	0.1209	0.0897	0.1247	0.0905	0.1054	0.1170
cart	40	bias	−0.1412	0.0320	−0.0436	0.0047	−0.0317	0.0096	−0.0418	−0.0108	0.0044
		sd	0.0461	0.0701	0.0497	0.0753	0.0539	0.0773	0.0503	0.0646	0.0787
		rms	0.1485	0.0771	0.0661	0.0755	0.0625	0.0779	0.0654	0.0655	0.0788
cart	60	bias	−0.1397	0.0284	−0.0404	0.0021	−0.0334	0.0088	−0.0393	−0.0155	0.0049
		sd	0.0347	0.0580	0.0418	0.0626	0.0441	0.0648	0.0424	0.0521	0.0636
		rms	0.1439	0.0646	0.0581	0.0627	0.0554	0.0654	0.0578	0.0544	0.0637
3nn	20	bias	−0.0820	0.0554	−0.0488	0.0165	0.0048	0.0200	−0.0371	0.0233	0.0104
		sd	0.0748	0.1041	0.0757	0.1100	0.0871	0.1129	0.0805	0.0993	0.1037
		rms	0.1110	0.1179	0.0901	0.1112	0.0872	0.1147	0.0886	0.1020	0.1043
3nn	40	bias	−0.0673	0.0405	−0.0377	0.0029	0.0008	0.0067	−0.0271	0.0099	0.0040
		sd	0.0458	0.0643	0.0460	0.0679	0.0536	0.0695	0.0504	0.0585	0.0644
		rms	0.0814	0.0760	0.0595	0.0680	0.0536	0.0698	0.0572	0.0593	0.0645
3nn	60	bias	−0.0660	0.0304	−0.0375	0.0015	−0.0051	0.0040	−0.0269	0.0016	0.0006
		sd	0.0389	0.0534	0.0393	0.0560	0.0451	0.0563	0.0435	0.0482	0.0557
		rms	0.0766	0.0614	0.0543	0.0560	0.0454	0.0564	0.0511	0.0482	0.0557

We observed that the performance of error estimators other than out-of-bag (which can only be applied to ensemble rules) were consistent with their performance with the corresponding single classifier, as reported in other studies [18, 27].

4.4.2. Patient Data. The results for the real patient data sets were entirely consistent with those for the synthetic data, as can be seen in Figures 5, 6, and 7 and Tables 1, 2, and 3. We again observed the division of the error estimators in the same four groups according to performance. We also observed that the bolstered error estimator group displayed the best performance, as measured by RMS.

5. Conclusion

We presented an extensive study of several error estimation methods for bagged ensembles of typical classifiers. We provided here an explicit formulation for the out-of-bag error estimator, which is intended to remove estimator bias. We observed that this out-of-bag error estimator was almost identical to leave-one-out, under spherical Gaussian models, and conjectured a very close relationship between the two. The results of our simulation study were consistent between synthetic and real patient data, and the performance of error estimators that can be applied to single classifiers (i.e., all of them save for the out-of-bag estimator) with the bagged classifiers was comparable to their performance with the corresponding single classifier, as reported elsewhere. The bolstered error estimators exhibited the best performance, in terms of RMS error, in our simulation study, despite being far less computationally expensive than cross-validation and bootstrap estimators. We hope this work will provide useful guidance to practitioners working with bagged ensemble classifiers designed on small-sample data.

Acknowledgment

This work was supported by the National Science Foundation, through NSF Award CCF-0845407.

References

[1] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[2] Y. Freund, "Boosting a weak learning algorithm by majority," in *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pp. 202–216, 1990.

[3] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[4] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] B. Efron, "Bootstrap methods: another look at the jackknife," *Annals of Statistics*, vol. 7, pp. 1–26, 1979.

[7] S. Alvarez, R. Diaz-Uriarte, A. Osorio et al., "A predictor based on the somatic genomic changes of the BRCA1/BRCA2 breast cancer tumors identifies the non-BRCA1/BRCA2 tumors with BRCA1 promoter hypermethylation," *Clinical Cancer Research*, vol. 11, no. 3, pp. 1146–1153, 2005.

[8] E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento, and M. P. Heyes, "Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9608–9613, 2003.

[9] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article 3, 2006.

[10] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinformatics*, vol. 9, article 319, 2008.

[11] G. Izmirlian, "Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial," *Annals of the New York Academy of Sciences*, vol. 1020, pp. 154–174, 2004.

[12] B. Wu, T. Abbott, D. Fishman et al., "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[13] P. Geurts, M. Fillet, D. de Seny et al., "Proteomic mass spectra classification using decision tree based ensemble methods," *Bioinformatics*, vol. 21, no. 14, pp. 3138–3145, 2005.

[14] B. Zhang, T. D. Pham, and Y. Zhang, "Bagging support vector machine for classification of SELDI-ToF mass spectra of ovarian cancer serum samples," in *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence (AI '07)*, vol. 4830 of *Lecture Notes in Computer Science*, pp. 820–826, Gold Coast, Australia, December 2007.

[15] A. Assareh, M. H. Moradi, and V. Esmaeili, "A novel ensemble strategy for classification of prostate cancer protein mass spectra," in *Proceedings of the 29th Annual International Conference of IEEE-EMBS, Engineering in Medicine and Biology Society (EMBC '07)*, pp. 5987–5990, August 2007.

[16] W. Tong, Q. Xie, H. Hong et al., "Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence," *Environmental Health Perspectives*, vol. 112, no. 16, pp. 1622–1627, 2004.

[17] T. T. Vu and U. M. Braga-Neto, "Is bagging effective in the classification of small-sample genomic and proteomic data?" *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2009, Article ID 158368, 2009.

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

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

[20] U. Braga-Neto and E. Dougherty, "Exact performance of error estimators for discrete classifiers," *Pattern Recognition*, vol. 38, no. 11, pp. 1799–1814, 2005.

[21] L. Breiman, *Out-of-bag estimation*, Department of Statistics, University of California, ftp://ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.Z.

[22] T. Bylander, "Estimating generalization error on two-class datasets using out-of-bag estimates," *Machine Learning*, vol. 48, no. 1–3, pp. 287–297, 2002.

[23] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A comparison of decision tree ensemble creation techniques," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 173–180, 2007.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

[25] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.

[26] B. Efron and R. Tibshirani, "Improvements on cross-validation: the .632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, 1997.

[27] U. Braga-Neto and E. Dougherty, "Bolstered error estimation," *Pattern Recognition*, vol. 37, no. 6, pp. 1267–1281, 2004.

[28] G. Martínez-Muñoz and A. Suárez, "Out-of-bag estimation of the optimal sample size in bagging," *Pattern Recognition*, vol. 43, no. 1, pp. 143–152, 2010.

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

[30] L. J. Van't Veer, H. Dai, M. J. van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

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

[32] H. J. Issaq, T. D. Veenstra, T. P. Conrads, and D. Felschow, "The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification," *Biochemical and Biophysical Research Communications*, vol. 292, no. 3, pp. 587–592, 2002.

[33] B.-L. Adam, Y. Qu, J. W. Davis et al., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Research*, vol. 62, no. 13, pp. 3609–3614, 2002.