Computational Statistics and Data Analysis 53 (2009) 3735–3745

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap

Ji-Hyun Kim*

Department of Statistics and Actuarial Science, Soongsil University, Dongjak-Ku Sangdo-Dong, Seoul 156-743, Republic of Korea

ARTICLE INFO

Article history:
Received 26 September 2006
Received in revised form 17 April 2008
Accepted 21 April 2009
Available online 3 May 2009

ABSTRACT

We consider the accuracy estimation of a classifier constructed on a given training sample. The naive resubstitution estimate is known to have a downward bias problem. The traditional approach to tackling this bias problem is cross-validation. The bootstrap is another way to bring down the high variability of cross-validation. But a direct comparison of the two estimators, cross-validation and bootstrap, is not fair because the latter estimator requires much heavier computation. We performed an empirical study to compare the .632+ bootstrap estimator with the repeated 10-fold cross-validation and the repeated one-third holdout estimator. All the estimators were set to require about the same amount of computation. In the simulation study, the repeated 10-fold cross-validation estimator was found to have better performance than the .632+ bootstrap estimator when the classifier is highly adaptive to the training sample. We have also found that the .632+ bootstrap estimator suffers from a bias problem for large samples as well as for small samples.

1. Introduction

For the classification problem, it is important to estimate the true error rate of a given classifier when an independent set of testing samples is not available. Cross-validation (CV) is a traditional approach, which gives a nearly unbiased but highly variable estimator. The bootstrap is another approach, which gives a less variable estimator for small samples (Efron, 1983; Efron and Tibshirani, 1997). The estimators obtained by two methods were compared in the vast literature (Efron, 1983; Crawford, 1989; Fitzmaurice et al., 1991; Efron and Tibshirani, 1997; Braga-Neto and Dougherty, 2004; Wehberg and Schumacher, 2004; Molinaro et al., 2005; Schiavo and Hand, 2000). The bootstrap estimator is, on the whole, known to have better performance for small samples because of its small variance. A straightforward comparison, however, is not fair since the bootstrap estimator requires much more computation. To make it even and to reduce the variance of the CV estimator, we consider a repeated CV estimator, in which we repeat CV several times and then take the average.

Most comparison studies of CV versus bootstrap were done with no repetition of the CV. The better performance of bootstrap estimators for small samples may have been attained by its extra cost of computation. There has not been much study on the repeated CV. Braga-Neto and Dougherty (2004) studied the repeated CV and the bootstrap for a small-sample microarray classification, and concluded that the .632 bootstrap estimator is the best in their simulation study with the caution that it might be susceptible to a downward bias. But they did not consider the .632+ bootstrap estimator, which is a bias-corrected version of the .632 estimator proposed by Efron and Tibshirani (1997). And they considered only small samples varying from 20 to 120, and did not apply a highly adaptive classification rule such as boosting. The classifier by boosting algorithm, called the *discrete adaboost* (Freund and Schapire, 1997), is highly adaptive in the sense that the

* Tel.: +82 2 820 0445; fax: +82 2 823 1746.
E-mail address: jxk61@ssu.ac.kr.

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.04.009

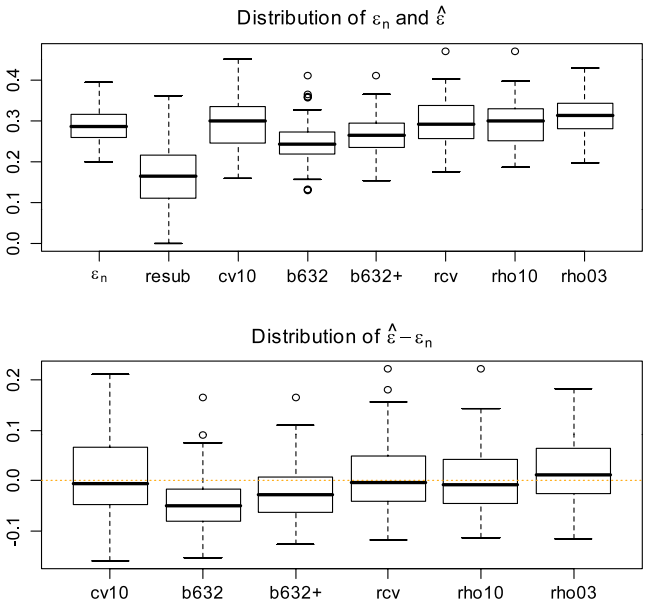

Fig. 1. Distribution of estimators for pruned tree model: Distribution from 100 experiments of Simulation 1; sample size $n = 100$; ε_n is the true error rate; resub is the resubstitution estimator of ε_n; the bootstrap estimators, b632 and b632+, underestimate ε_n, and the repeated CV estimator rcv is less variable than the ordinary 10-fold CV estimator cv10.

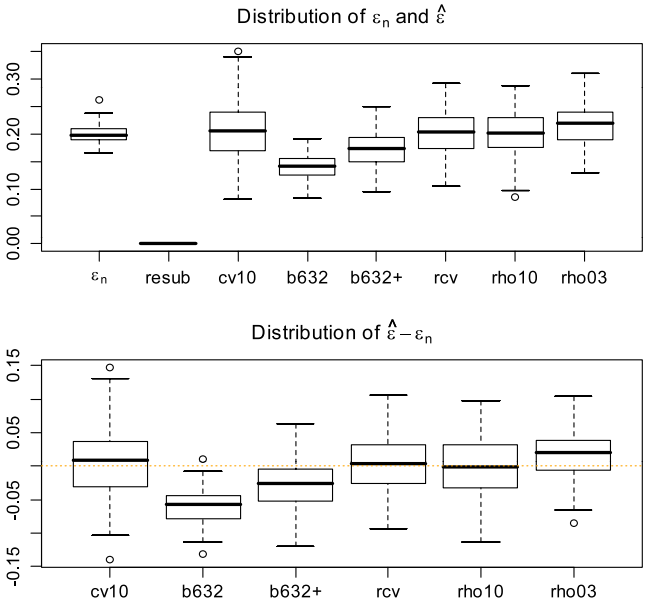

Fig. 2. Distribution of estimators for boosting: Distribution from 100 experiments of Simulation 1; The classifier by boosting is more accurate since the true error rates ε_n for boosting are smaller than those for the pruned tree model; the values of resubstitution estimator of ε_n, resub, are very small; the .632 bootstrap estimator b632 severely underestimates ε_n, and the repeated CV estimator rcv is less variable than the ordinary 10-fold CV estimator cv10.

resubstitution error estimate of the classifier induced by the algorithm easily becomes close to zero. Since the bootstrap estimate directly depends on the resubstitution estimate, the choice of the best estimator can be different when an adaptive algorithm is used for classification rule induction. Actually, Kim and Cha (2006) report that the bias of bootstrap estimators tends to be serious even for large samples when the boosting algorithm is used for rule induction.

Burman (1989) examined the repeated CV, but he did it in the context of regression, not of classification. Its derivation of large sample property of the repeated CV estimator is only valid for a continuous response variable as in the regression. The large sample property of the bootstrap estimator is not available either. The theoretical basis of the bootstrap estimator is weak as mentioned in Efron (1983). Theoretical comparison between the repeated CV and bootstrap estimator in the classification problem does not seem to be feasible.

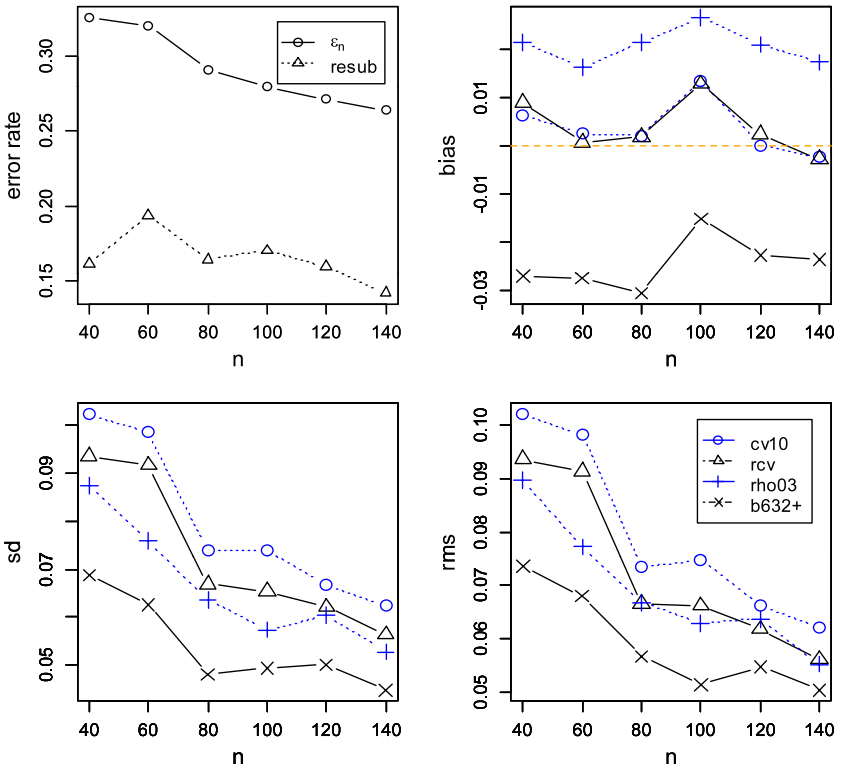

Fig. 3. Learning curve and three performance measures for the pruned tree model: Small-to-moderate sample sizes in simulation1; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. The bootstrap estimator b632+ shows the best overall performance.

The main purpose of this study is to compare empirically the repeated CV estimator with the .632+ bootstrap estimator in the binary classification problem. Both a highly adaptive classifier by boosting and a relatively smooth classifier by the pruned tree model (Breiman et al., 1984) are used for the simulation study, where the sample sizes vary from 40 to 1000.

This paper is organized as follows: the problem and all the error rate estimators of concern are described in the next section; simulation studies on two artificial data structures are done in Section 3; and finally conclusions and comments are made in the last section.

2. Estimation of classification error rate

A *training sample* $\mathcal{D}_n$ consists of n observations $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ is a predictor or a feature vector, and y_i a response or a label. For simplicity we assume the two-class problem, $y_i \in \{0, 1\}$. A *classifier* $r_n(\mathbf{x}) = r(\mathcal{D}_n, \mathbf{x})$ is constructed based on the training sample $\mathcal{D}_n$ by a classification rule. A *classification rule* maps the training sample $\mathcal{D}_n$ into the classifier $r_n(\cdot)$.

We want to estimate the *true conditional error rate* of a classifier given the training sample $\mathcal{D}_n$ defined as

$$\epsilon_n = E[I(Y \neq r_n(\mathbf{X}))|\mathcal{D}_n]$$

where the expectation is taken over the distribution of a future observation $(\mathbf{X}, Y)$. The true conditional error rate is the conditional probability of misclassification given the training sample. In order to select the best classification rule rather than to estimate the accuracy of a given classifier, one might want to estimate the unconditional error rate $E(\epsilon_n)$ where the expectation is taken over the distribution of all possible training samples $\mathcal{D}_n$.

The resubstitution estimator of the error rate is obtained by using the same sample to construct a classifier and also to assess its performance. Hence the resubstitution estimator underestimates the true error rate. Cross-validation is the traditional method of choice to tackle the downward bias problem of the resubstitution estimator.

In k-fold CV, the training sample is partitioned into k-folds as equally as possible, and each fold is left out of the learning process and used as a testing set. The estimate of error rate is the overall proportion of error incurred on all folds of testing set. Leave-one-out CV is the special case of k-fold CV where $k = n$, and its computational cost becomes higher in proportion to the size of training sample. The choice of $k = 10$ is the most popular since the number of model fitting to get the estimate now becomes independent of the size of training sample, and since the bias of the estimator often matters with smaller k (Kohavi, 1995).

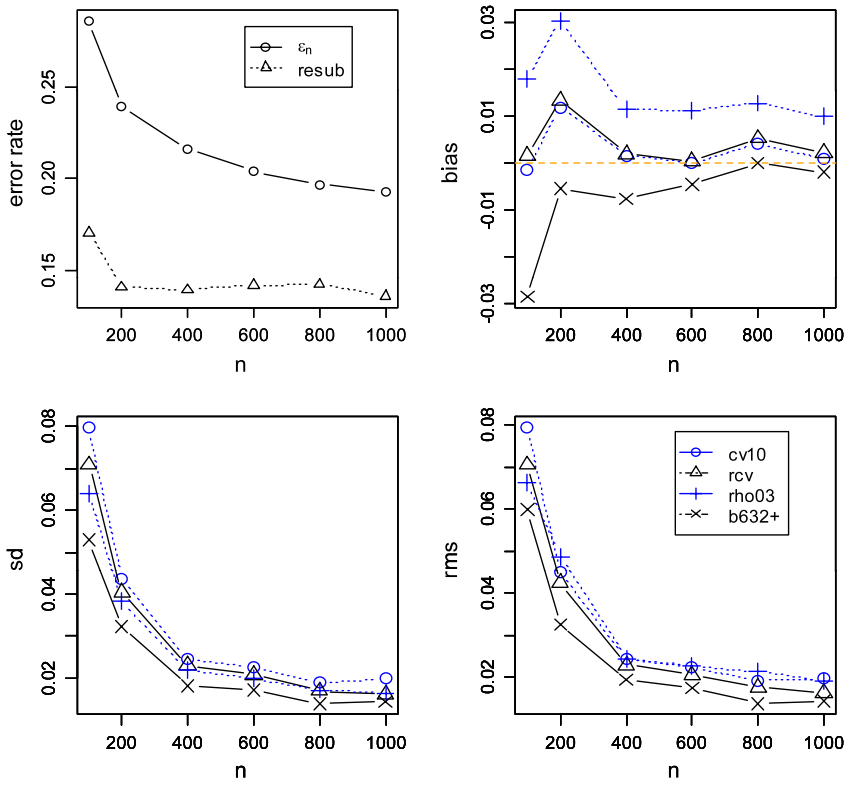

Fig. 4. Learning curve and three performance measures for the pruned tree model: Moderate-to-large sample sizes in simulation 1; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. The bootstrap estimator b632+ shows the best overall performance.

The CV estimator is known to be nearly unbiased but can be highly variable for small samples. To reduce its variability, bootstrap estimators were proposed by Efron (1983) and Efron and Tibshirani (1997), which are described briefly here.

If we denote the resubstitution estimator as $\hat{\epsilon}_{\text{resub}}$, the true (conditional) error rate ϵ_n can be rewritten as

$$\epsilon_n = \hat{\epsilon}_{\text{resub}} + (\epsilon_n - \hat{\epsilon}_{\text{resub}}).$$

Now the estimation of the term $\epsilon_n - \hat{\epsilon}_{\text{resub}}$ is of concern. Efron (1983) proposed the .632 bootstrap estimator of ϵ_n defined as

$$\hat{\epsilon}_{\text{b632}} = \hat{\epsilon}_{\text{resub}} + .632(\hat{\epsilon}_{\text{b0}} - \hat{\epsilon}_{\text{resub}}).$$

The naive bootstrap estimator $\hat{\epsilon}_{\text{b0}}$ is based on B bootstrap samples (The exact formula of $\hat{\epsilon}_{\text{b0}}$ is referred to in equation (6.11) in Efron (1983). The value of B in this study is set to 50 as recommended there). Since each bootstrap sample of size n has only $.632n$ different observations on average, $\hat{\epsilon}_{\text{b0}}$ tends to overestimate ϵ_n. The weight .632 mitigates this overestimation. Now, the .632 bootstrap estimator can be rewritten as

$$\hat{\epsilon}_{\text{b632}} = .368\hat{\epsilon}_{\text{resub}} + .632\hat{\epsilon}_{\text{b0}}.$$

Efron and Tibshirani (1997) pointed out that $\hat{\epsilon}_{\text{b632}}$ gets downward-biased when a highly overfitting classification rule like 1-nearest-neighbor is employed to construct the classifier. They proposed to put more weight on $\hat{\epsilon}_{\text{b0}}$ when the amount of overfitting, as measured by $\hat{\epsilon}_{\text{b0}} - \hat{\epsilon}_{\text{resub}}$, gets larger. Their new estimator, called the .632+ bootstrap estimator, is defined as

$$\hat{\epsilon}_{\text{b632+}} = (1 - \hat{w})\hat{\epsilon}_{\text{resub}} + \hat{w}\hat{\epsilon}_{\text{b0}}$$

where $\hat{w} = .632/(1 - .368\hat{R})$ and $\hat{R} = (\hat{\epsilon}_{\text{b0}} - \hat{\epsilon}_{\text{resub}})/(\hat{\gamma} - \hat{\epsilon}_{\text{resub}})$. Here, $\hat{\gamma}$ is the estimate of the error rate when there is no information, i.e. when Y is independent of $\mathbf{X}$. For the binary classification problem it can be computed as $\hat{\gamma} = \hat{p}_1(1 - \hat{q}_1) + (1 - \hat{p}_1)\hat{q}_1$, where $\hat{p}_1$ is the observed proportion of y_i equalling 1 and $\hat{q}_1$ is the observed proportion of $\hat{y}_i = r_n(\mathbf{x}_i)$ equalling 1. The weight $\hat{w}$ varies from .632 to 1, and $\hat{\epsilon}_{\text{b632+}}$ gets closer to $\hat{\epsilon}_{\text{b632}}$ as the measure of overfitting $\hat{\epsilon}_{\text{b0}} - \hat{\epsilon}_{\text{resub}}$ becomes negligible as compared to $\hat{\gamma} - \hat{\epsilon}_{\text{resub}}$.

With the training sample fixed, the k-fold CV estimator has a variation due to the randomness of partitioning the sample into k-folds. Some literature (Efron and Tibshirani, 1997; Braga-Neto and Dougherty, 2004) called this variation the *internal variance*. One way to reduce this internal variance of the k-fold CV estimator is to repeat the whole process of partitioning and estimating. In this study we obtain the 10-fold CV estimates 5 times, and take the average as the final estimate. We call

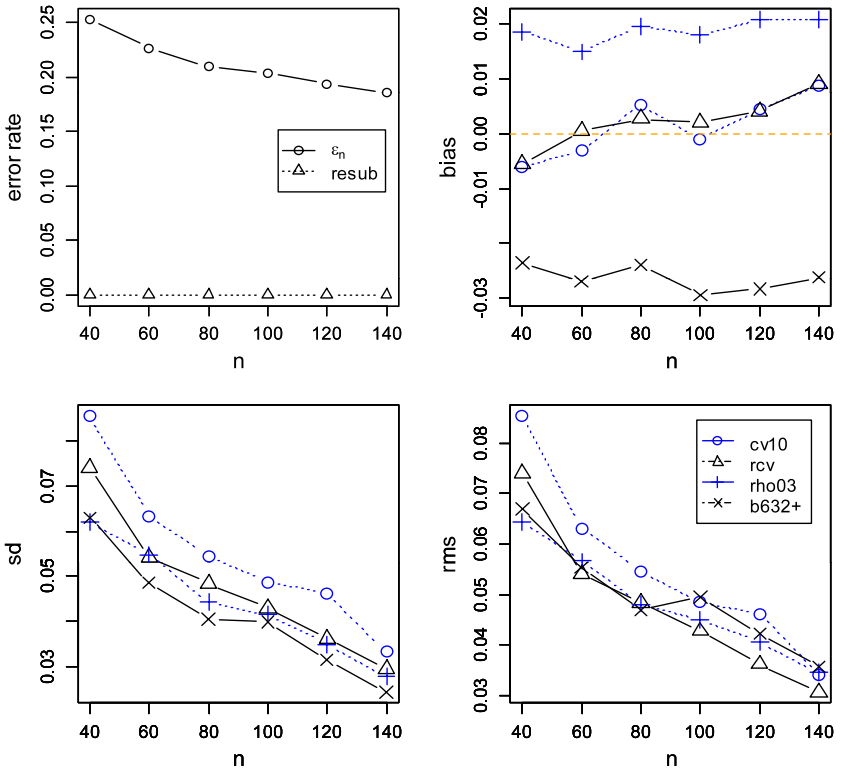

Fig. 5. Learning curve and three performance measures for boosting: Small-to-moderate sample sizes in simulation 1; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. The repeated CV estimator rcv shows the best overall performance for the adaptive classifier by the boosting algorithm except a very small sample case.

it the *repeated CV estimator*, and denote it by $\hat{\epsilon}_{\text{rcv}}$. The number of repetitions of the 10-fold CV is set to 5 because the number of model fitting (or the number of constructing the classifiers) for the two estimates, $\hat{\epsilon}_{\text{rcv}}$ and $\hat{\epsilon}_{\text{b632+}}$ with $B = 50$, then gets equal to 50.

Instead of k-fold cross-validation, a repeated holdout method is often used in the field of application. When given no testing sample independent of the training sample, one randomly selects and holds out a portion of the training sample for testing, and constructs a classifier with only the remaining sample. The true error rate of the constructed classifier is estimated with the held-out testing sample, and this whole process is repeated many times, and the average of repeatedly obtained estimates of error rate is called the *repeated holdout estimate*. Often one third of the training sample is set aside for testing. We also tried a one-tenth holdout. The holdout procedure is repeated 50 times in this study for fair comparison. We denote the resulting estimators by $\hat{\epsilon}_{\text{rho03}}$ and $\hat{\epsilon}_{\text{rho10}}$, respectively.

To study the performance of an error estimator $\hat{\epsilon}$, it has been pointed out that the distribution of $\hat{\epsilon} - \epsilon_n$ should be examined instead of the distribution of $\hat{\epsilon}$ by Krzanowski and Hand (1997), and the distribution of $\hat{\epsilon} - \epsilon_n$ was called the *deviation distribution* by Braga-Neto and Dougherty (2004). Taking $\hat{\epsilon} - \epsilon_n$ as a random variable, the relation

$$E[(\hat{\epsilon} - \epsilon_n)^2] = Var(\hat{\epsilon} - \epsilon_n) + [E(\hat{\epsilon} - \epsilon_n)]^2 \tag{2.1}$$

holds, where the total performance measure of $\hat{\epsilon}$, $E[(\hat{\epsilon} - \epsilon_n)^2]$, is decomposed into a variance component and a bias component. Note that the variance component is $Var(\hat{\epsilon} - \epsilon_n)$, not $Var(\hat{\epsilon})$. The three terms will be estimated numerically in the simulation.

3. Simulation study

Two kinds of classification rules are applied. The first one is based on the boosting algorithm which is *adaptive* in the sense that it usually gives classifiers fitting the training sample well, so that the resubstitution error estimate is very small. To avoid overfitting we adopted boosting with maximum-depth-2 tree as weak learners, and with the number of iterations of boosting as 50, which is determined by looking at two error-rate curves for the training sample and testing sample respectively. The other classification rule is the pruned tree model which is relatively smooth or *less-adaptive* in the sense that it produces relatively simpler classifiers, so that the resubstitution error rate is not too small. We adopted *one-SE rule* to select the best pruned tree (Breiman et al., 1984).

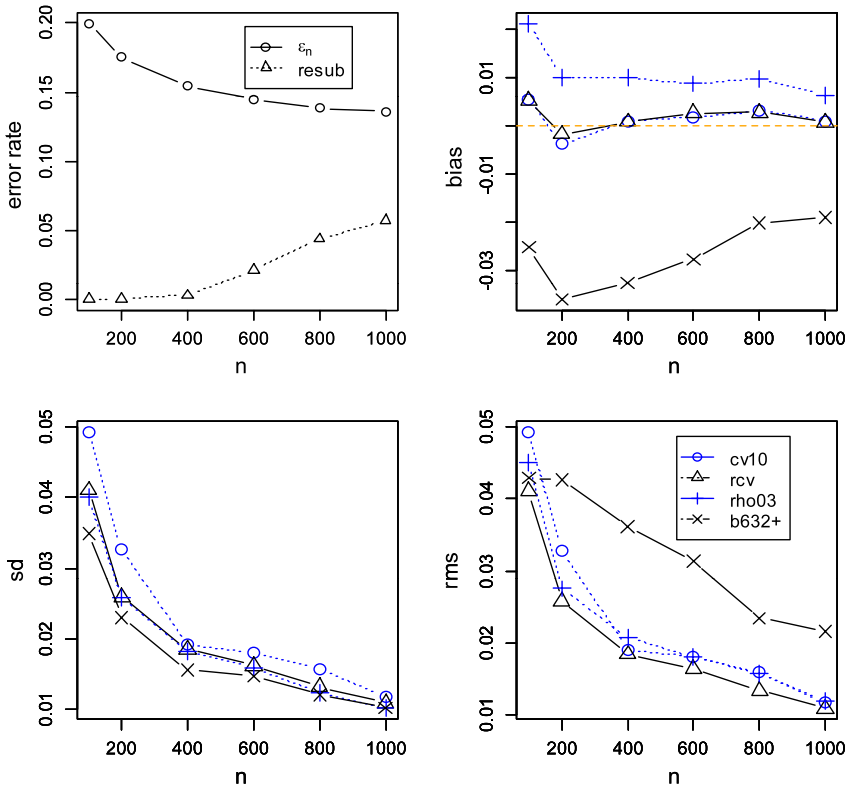

Fig. 6. Learning curve and three performance measures for boosting: Moderate-to-large sample sizes in simulation 1; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. The repeated CV estimator rcv shows the best overall performance.

The statistical environment R (R Development Core Team, 2004) was used for simulation. For the construction of the tree model we used the RPART package (Therneau and Atkinson, 2004), and for the boosting we used the algorithm in Freund and Schapire (1997), called the discrete adaboost. For the bootstrap estimators we used the BOOTSTRAP package (Tibshirani, 2004) with a little modification of the input and output.

We wanted to investigate the following things through the simulation study:

- How much can we reduce the variance of the CV estimator by repeating the random k-fold partitions in CV?
- Is the commonly-used one-third holdout procedure valid?
- Does the bootstrap estimator have an advantage over the repeated CV? Or which is the winner?
- What is the effect of the size of training sample? How does the performance of estimators change as the size of the training sample varies?
- What is the effect of the types of classification rules (adaptive and less-adaptive)?

3.1. Simulation 1

A Bernoulli random variable Y with the probability of success $\mu = 1/[1 + \exp(-F(\mathbf{x}))]$ is generated, where

$$F(\mathbf{x}) = -10 + 10\sin(\pi x_1 x_2) + 5(x_3 - 1/2)^2 + 5x_4 + 2x_5.$$

The first four predictors have uniform distribution on the interval (0, 1), and the last predictor x_5 has a discrete uniform distribution on {1, 2, 3}. To make the classification harder we introduce another five random variables $(x_6, \ldots, x_{10})$ which have the same distribution as $(x_1, \ldots, x_5)$ but are not related to Y. All ten predictors are independently generated.

We first draw boxplots to examine the distribution of various estimators of the true conditional error rate for the training sample size $n = 100$. To get the distribution of estimators and to estimate the three terms in the Eq. (2.1), we iterate the experiments 100 times. For the estimation of ϵ_n in each experiment, we draw a testing sample of size 2000 independently of the training sample. Fig. 1 shows the result for the pruned tree model, and Fig. 2 for the boosting. The upper graph in each figure shows the distribution of $\hat{\epsilon}$, and lower one shows the distribution of $\hat{\epsilon} - \epsilon_n$. Note that ϵ_n also has a variation along the iterations.

The upper graphs in Figs. 1 and 2 tell us that the naive resubstitution estimate $\hat{\epsilon}_{\text{resub}}$ severely underestimates ϵ_n. For the adaptive classifier they are all zeroes, in which case the bootstrap estimator $\hat{\epsilon}_{\text{b632}}$ underestimates too much as shown

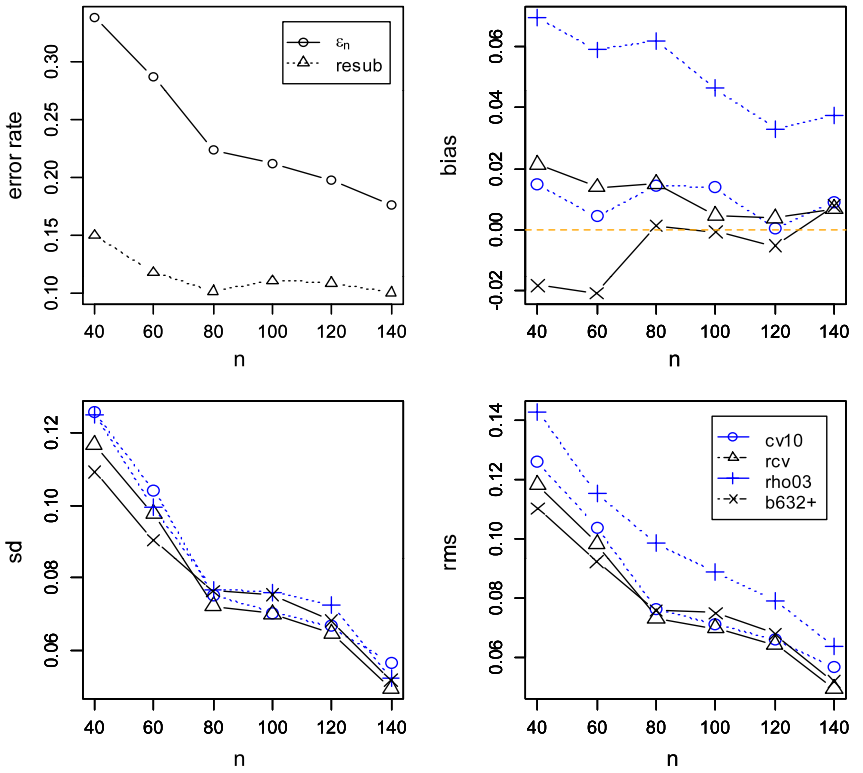

Fig. 7. Learning curve and three performance measures for the pruned tree model: Small-to-moderate sample sizes in simulation 2 with $p = 0.9$; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. The one-third holdout estimator rho03 is biased upward.

by the lower graph in Fig. 2 since it directly depends on $\hat{\epsilon}_{\text{resub}}$. Kim and Cha (2006) gave empirical evidence that the bias problem of $\hat{\epsilon}_{\text{b632}}$ is so serious in the adaptive classifier by boosting that in the total performance measure of root-mean-square (RMS), $\sqrt{E[(\hat{\epsilon} - \epsilon_n)^2]}$, $\hat{\epsilon}_{\text{b632}}$ is far inferior to $\hat{\epsilon}_{\text{b632+}}$ or even to the 10-fold CV estimator $\hat{\epsilon}_{\text{cv10}}$. From now on we only report on $\hat{\epsilon}_{\text{b632+}}$ which was devised to correct the bias of $\hat{\epsilon}_{\text{b632}}$.

The repeated 10-fold CV estimator $\hat{\epsilon}_{\text{rcv}}$ has less variability than the non-repeated one $\hat{\epsilon}_{\text{cv10}}$. The ratios of standard deviations of the two estimates, $\frac{\sqrt{Var(\hat{\epsilon}_{\text{cv10}} - \epsilon_n)}}{\sqrt{Var(\hat{\epsilon}_{\text{rcv}} - \epsilon_n)}}$ are 1.08 and 1.21 for the pruned tree model and boosting, respectively. (The statistical significance of these two values is not easy to quantify. Assuming two samples of estimates, $\{\hat{\epsilon}_{\text{cv10},t}, t = 1, \ldots, 100\}$ and $\{\hat{\epsilon}_{\text{rcv},t}, t = 1, \ldots, 100\}$, are independent, p-values of F-test for the equal-variance hypothesis with one-sided alternative are 0.222, 0.020, respectively. But these values are not valid since the two estimators are paired as $\{(\hat{\epsilon}_{\text{cv10},t}, \hat{\epsilon}_{\text{rcv},t}), t = 1, \ldots, 100\}$. However, we are fairly sure the difference in variances of the two estimates is not by chance, since it is observed consistently for other experiments with sample sizes as reported later.)

Figs. 1 and 2 also show that estimators based on CV, $\hat{\epsilon}_{\text{cv10}}$ and $\hat{\epsilon}_{\text{rcv}}$, are nearly unbiased, whereas $\hat{\epsilon}_{\text{rho03}}$ is biased upward since it uses only two thirds of the available training sample. In the simulation study the two estimators, $\hat{\epsilon}_{\text{rcv}}$ and $\hat{\epsilon}_{\text{rho10}}$, have similar distributions because the only difference between the two is that for $\hat{\epsilon}_{\text{rcv}}$ each observation of the training sample is used for testing exactly once in each repetition of CV, whereas for $\hat{\epsilon}_{\text{rho10}}$ some observations of the training sample may not be used at all or may be used more than once for testing.

To examine the distribution of $\hat{\epsilon} - \epsilon_n$ for the varying sample sizes, and also to decompose the variation in Figs. 1 and 2 into the variance component and the bias component, we draw Figs. 3 through 6. The upper left graph in Fig. 3 shows the learning curve and the resubstitution estimate for the pruned tree model. The point in the solid line denotes the mean of 100 conditional true errors ϵ_n. The point in the dotted line is the mean of 100 resubstitution estimates. Comparing this graph for the pruned tree model with the corresponding one in Fig. 4 for the boosting, we confirm that the pruned tree model is not as adaptive as the boosting one. The upper right graph in Fig. 3 shows the (estimated) bias component $E(\hat{\epsilon} - \epsilon_n)$ for the four different estimates of ϵ_n. Note that $\hat{\epsilon}_{\text{cv10}}$ and $\hat{\epsilon}_{\text{rcv}}$ are nearly unbiased, whereas $\hat{\epsilon}_{\text{rho03}}$ is biased upward and $\hat{\epsilon}_{\text{b632+}}$ is biased downward. But $\hat{\epsilon}_{\text{b632+}}$ has a smaller standard deviation $\sqrt{Var(\hat{\epsilon}_{\text{b632+}} - \epsilon_n)}$, having smaller RMS as shown in the two lower graphs in Fig. 3. A similar pattern is observed for moderate-to-large samples as in Fig. 4, but here $\hat{\epsilon}_{\text{b632+}}$ has less bias than for the small sample case. Overall for the pruned tree model, $\hat{\epsilon}_{\text{b632+}}$ is the winner by the total performance measure RMS.

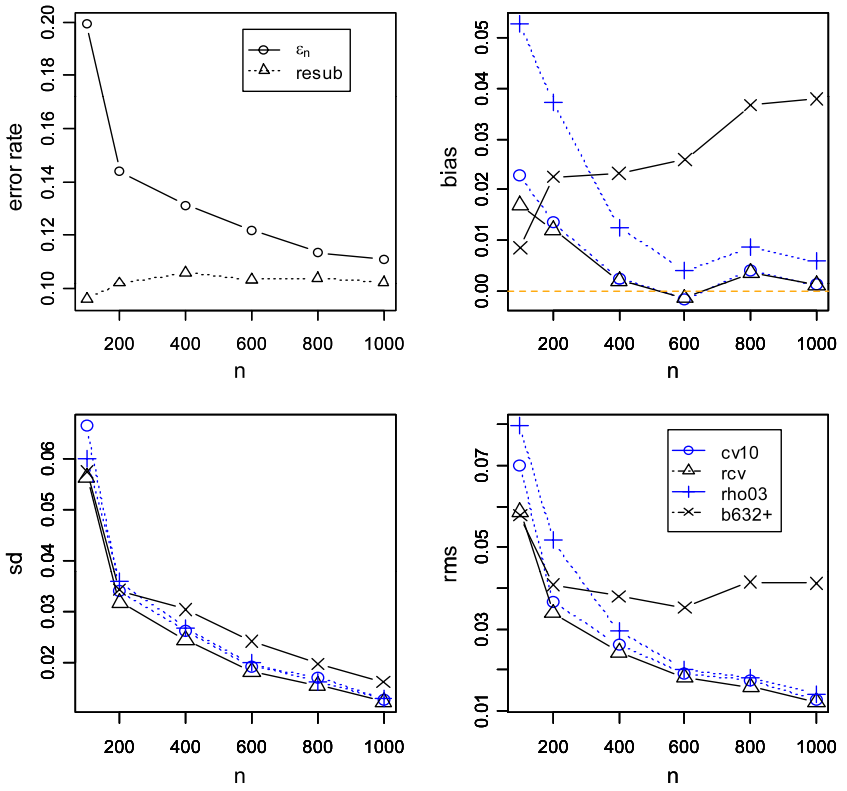

Fig. 8. Learning curve and three performance measures for the pruned tree model: Moderate-to-large sample sizes in simulation 2 with $p = 0.9$; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. The bootstrap estimator b632+ shows a deviant behavior when there is little difference between the true error rate and the resubstitution error estimate.

The boosting, an adaptive classification rule, tells a different story. For small samples as in Fig. 5, the naive estimates $\hat{\epsilon}_{\text{resub}}$ are zeroes, causing the bootstrap estimate $\hat{\epsilon}_{\text{b632+}}$ as well as $\hat{\epsilon}_{\text{b632}}$ to be biased downward. Hence despite the small standard deviation, bootstrap estimators are poor in RMS except at the smallest sample case with $n = 40$. For moderate-to-large samples, this tendency of bootstrap estimators persists as in Fig. 6. For large samples, the difference in standard deviation between $\hat{\epsilon}_{\text{b632+}}$ and $\hat{\epsilon}_{\text{rcv}}$ becomes small but the bias of $\hat{\epsilon}_{\text{b632+}}$ is still noticeable. Overall for the boosting, $\hat{\epsilon}_{\text{rcv}}$ is the winner. ($\hat{\epsilon}_{\text{rho10}}$ has very similar distributions as $\hat{\epsilon}_{\text{rcv}}$ throughout the simulation study. We did not include it in the graph for the sake of readability.)

The one-third holdout estimate $\hat{\epsilon}_{\text{rho03}}$ is biased upward in both classifiers since it uses only two-thirds of the available sample to construct the classifier. But the larger size of the held-out testing sample gives more stability in estimating the error rate, so that it has smaller variance than $\hat{\epsilon}_{\text{cv10}}$ or $\hat{\epsilon}_{\text{rcv}}$. In the total performance measure, however, it is not the best one.

3.2. Simulation 2

In this experiment, there are only two explanatory variables X_1 and X_2, which are independent and uniformly distributed on $(-1, 1)$. When $X_1X_2 > 0$ we generate a Bernoulli response variable Y with the probability of success $p = P(Y = 1)$, and when $X_1X_2 < 0$, we generate a binary response variable Y with $p = P(Y = 0) = 1 - P(Y = 1)$. In the experiment we vary p from 0.5 to 1.0 incremented by 0.1. The value 0.5 means no association between (X_1, X_2) and Y, whereas the value 1.0 means a deterministic association. The sample size n is also varied from 40 to 140 to see the small sample behavior, and from 100 to 1000 to see the moderate-to-large sample behavior. By varying both p and n we can compare the estimators in more diverse conditions.

For the pruned tree model with small samples, Simulation 2 is similar in results to Simulation 1. The bootstrap estimator $\hat{\epsilon}_{\text{b632+}}$ is biased downward but has smaller variance, hence comes to have a smaller RMS as in Fig. 7. The repeated CV estimator $\hat{\epsilon}_{\text{rcv}}$ has smaller variance than $\hat{\epsilon}_{\text{cv10}}$. The repeated one-third holdout estimator $\hat{\epsilon}_{\text{rho03}}$ is biased upwards when the learning curve of the true error rate has a steep gradient as in $p = 0.9$.

For the pruned tree model with moderate-to-large samples, there is one new finding. The bootstrap estimate $\hat{\epsilon}_{\text{b632+}}$ is biased *upward* and more variable than $\hat{\epsilon}_{\text{rcv}}$ when the difference between the true error rate and the resubstitution estimate becomes negligible as in Fig. 8. In this case, $\hat{\epsilon}_{\text{rcv}}$ outperforms $\hat{\epsilon}_{\text{b632+}}$ in RMS.

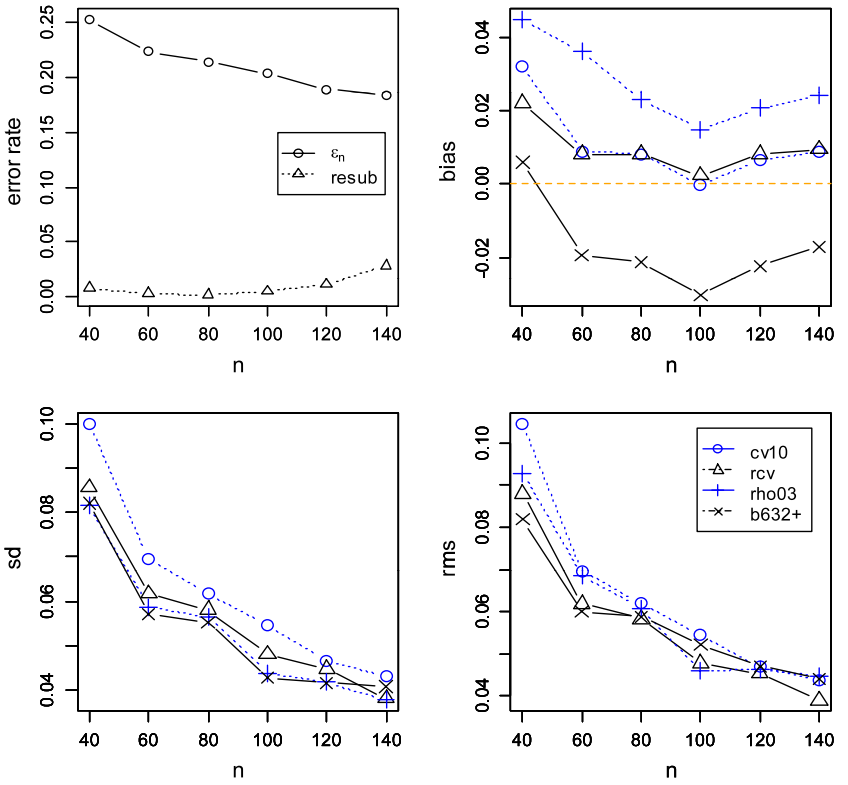

Fig. 9. Learning curve and three performance measures for boosting: Small-to-moderate sample sizes in simulation 2 with $p = 0.9$; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. The bootstrap estimate b632+ is biased downward when the resubstitution estimate is very small.

Fig. 9 shows that the downward bias of $\hat{\epsilon}_{\text{b632+}}$ matters for boosting with small samples. And the standard deviation of $\hat{\epsilon}_{\text{b632+}}$ is larger than $\hat{\epsilon}_{\text{rcv}}$ for $p = 0.7$ or 0.8. (The graphs for p other than 0.9 are not shown here.) The upward bias of $\hat{\epsilon}_{\text{rho03}}$ is noticeable for $p = 0.9$ or 1.0. Because of the small standard deviation, however, the holdout estimator $\hat{\epsilon}_{\text{rho03}}$ is competitive in RMS for cases with $p \leq 0.7$. The repeated CV estimator $\hat{\epsilon}_{\text{rcv}}$ shows stable performance overall.

The learning curve in Fig. 10 for the boosting with moderate-to-large samples shows that it is hard to adapt the classifier to the training set for this experiment with large samples. The bootstrap estimator and the holdout estimator have a bias problem. Although there is no clear winner in RMS, the repeated CV estimator seems to be a reasonable choice.

To summarize Simulation 2, $\hat{\epsilon}_{\text{b632+}}$ is the winner when the relatively smooth pruned tree model is applied to small samples. Other than that case, there is no clear winner. The bias-corrected bootstrap estimator $\hat{\epsilon}_{\text{b632+}}$ still has a bias problem, and shows a severely deviant behavior when the pruned tree model is applied to large samples. The repeated holdout estimator $\hat{\epsilon}_{\text{rho03}}$ has a serious bias problem for small samples. The repeated CV estimator $\hat{\epsilon}_{\text{rcv}}$ shows a stable performance overall, although it is not always the best.

Simulation 2 has a somewhat different result from Simulation 1. Note that the learning curves in the two simulations are different. The curves of the true error rate become flat for large samples in Simulation 2, and the two curves for the true error rate and resubstitution error rate move close together in many cases in Simulation 2. In addition, the resubstitution error rates of the classifiers by boosting are not small enough in Simulation 2, which means that the adaptive classifier does not get adaptive enough for the data structure in Simulation 2. Somewhat confusing results of Simulation 2 seem to come from these facts. But Simulation 2 has its value since it shows that there are many factors to be considered in the choice of the best error estimator.

4. Conclusion

We intended to compare empirically the estimators of the true conditional error rate in the binary classification problem. We performed a comparison study for the adaptive (but not necessarily overfitting) classifier by boosting as well as for the less-adaptive classifier by the pruned tree model, and for moderate-to-large samples as well as for small samples. In the limited cases considered here, we have drawn some conclusions. More extensive study is necessary in order to establish these as general facts:

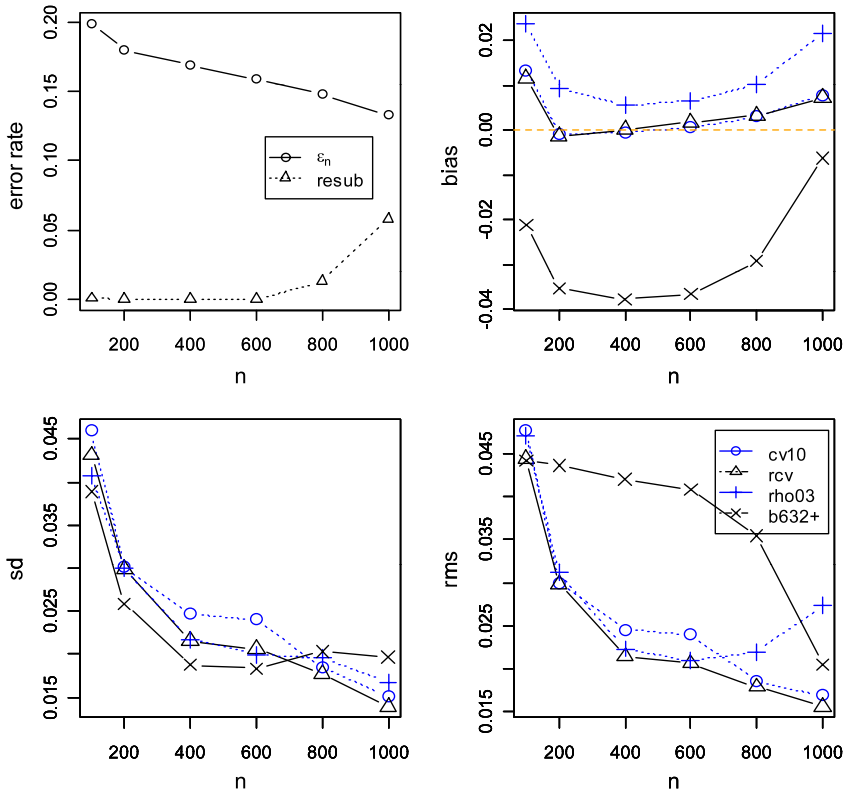

Fig. 10. Learning curve and three performance measures for boosting: Moderate-to-large sample sizes in simulation 2 with $p = 0.9$; Upper left graph shows the learning curve and the resubstitution estimate. The other three graphs show the three performance measures (i.e. bias, standard deviation and RMS) of four estimators of error rate. Each point is the average of 100 experiments. To increase the accuracy of the classifier, the maximum depth of weak learner, the classification tree, was set to 20 instead of 2. Note the severe downward bias of the .632+ estimate.

- The repeated CV estimator outperforms the non-repeated one by reducing the variability of the estimator. The heavier computation required for repetition is worth carrying out.
- The variance of the repeated one-third holdout estimator is usually smaller than the repeated CV estimator, and larger than the .632+ bootstrap estimator. But it is severely biased upward for the small sample case.
- The .632+ bootstrap estimator is usually biased downward with an exception in Fig. 8. Because of its small variance, it has the smallest RMS in a small sample less-adaptive classifier case. But due to its bias, it may show poor performance in a large sample case as shown in Figs. 6 and 8.
- The performance of estimators depends on the sample size. The winner in the small sample case is not necessarily the winner in the large sample case. And the bias of the .632+ bootstrap estimator matters even for large sample cases.
- The adaptive classifier by boosting shows somewhat different results from the less-adaptive one by the pruned tree model. When a classifier is highly adaptive, the .632+ bootstrap estimator has a downward bias problem even for large samples, so that it shows a poorer performance in terms of RMS than the repeated CV estimator.
- Except for the small sample less-adaptive classifier case where the .632+ bootstrap estimator is the best choice, the repeated CV estimator is recommended for general use. It shows a stable performance, whereas the repeated one-third holdout estimator and .632+ bootstrap estimator sometimes show deviant behavior as in Figs. 6, 8 and 10.

Acknowledgement

This work was supported by the Soongsil University Research Fund.

References

Braga-Neto, U.M., Dougherty, E.R., 2004. Is cross-validation valid for small-sample microarray classification? Bioinformatics 20, 374–380.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Pacific Grove, CA, Wadsworth.
Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76, 503–514.
Crawford, S.L., 1989. Extensions to the CART algorithm. International Journal of Man–Machine Studies 31, 197–217.
Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association 78, 316–331.
Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: The.632+ bootstrap method. Journal of the American Statistical Association 92, 548–560.

Fitzmaurice, G.M., Krzanowski, W.J., Hand, D.J., 1991. A monte carlo study of the 632 bootstrap estimator of error rate. Journal of Classification 8, 239–250.
Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.
Kim, J., Cha, E., 2006. Estimating prediction errors in binary classification problem: Cross-validation versus bootstrap. Korean Communications in Statistics 13, 151–165.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of Fourteenth International Joint Conference on Artificial Intelligence, IJCAI, Montreal, CA, pp. 1137–1143.
Krzanowski, W.J., Hand, D.J., 1997. Assessing error rate estimators: The leave-one-out method reconsidered. The Australian Journal of Statistics 39, 35–46.
Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: A comparison of resampling methods. Bioinformatics 21, 3301–3307.
R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available from http://www.R-project.org.
S original by Tibshirani, R. R port by Leisch, F., 2004. Bootstrap: functions for the book "An Introduction to the Bootstrap". R package version 1.0-15.
Schiavo, R., Hand, D.J., 2000. Ten more years of error rate research. International Statistical Review 68, 295–310.
Therneau, T., Atkinson, B., R port by Ripley, B., 2004. Rpart: recursive partitioning. R package version 3.1-20. S-PLUS 6.x original at http://www.mayo.edu/hsr/Sfunc.html.
Wehberg, S., Schumacher, M., 2004. A comparison of nonparametric error rate estimation methods in classification problems. Biometrical Journal 46, 35–47.