Estimating Generalization Error Using Out-of-Bag Estimates

Tom Bylander and Dennis Hanzlik
Division of Computer Science
University of Texas at San Antonio
San Antonio, Texas 78249-0667
bylander@cs.utsa.edu

Abstract

We provide a method for estimating the generalization error of a bag using out-of-bag estimates. In bagging, each predictor (single hypothesis) is learned from a bootstrap sample of the training examples; the output of a bag (a set of predictors) on an example is determined by voting. The out-of-bag estimate is based on recording the votes of each predictor on those training examples omitted from its bootstrap sample. Because no additional predictors are generated, the out-of-bag estimate requires considerably less time than 10-fold cross-validation. We address the question of how to use the out-of-bag estimate to estimate generalization error. Our experiments on several datasets show that the out-of-bag estimate and 10-fold cross-validation have very inaccurate (much too optimistic) confidence levels. We can improve the out-of-bag estimate by incorporating a correction.

Introduction

Supervised learning involves finding a hypothesis to correctly classify examples in a domain. If, for example, we wanted to classify mushrooms as edible or poisonous based on relevant characteristics such as color, smell, habitat, etc., we could learn a hypothesis by using mushrooms whose characteristics and classifications are known.

Much work has been done in supervised learning in developing learning algorithms for decision trees, neural networks, Bayesian networks, and other hypothesis spaces. As an improvement on these learning algorithms, work has recently been done using algorithms that combine several "single hypotheses" (called "predictors" from this point onward) into one "aggregate hypothesis." One such algorithm is bagging (bootstrap aggregating) (Breiman 1996a). Bagging involves repeated sampling with replacement to form several bootstrap training sets from the original dataset. Bagging should not be viewed as a competitor to other aggregation algorithms (such as boosting) because bagging can use any learning algorithm to generate predictors.

Over many types of predictor algorithms, bagging has been shown to improve on the accuracy of a single predictor (Breiman 1996a; Dietterich 1998b; Freund & Schapire 1996; Maclin & Opitz 1997; Quinlan 1996). An important issue is determining the generalization error of a bag (a bagging aggregate hypothesis). Usually, generalization error is estimated by k-fold cross-validation over the dataset (Michie, Spiegelhalter, & Taylor 1994; Weiss & Kulikowski 1991).

There are two potential problems with the cross-validation estimate (Wolpert & Macready 1996). One is the additional computation time. If there are B predictors in the bag, then $10B$ additional predictors must be generated for 10-fold cross-validation. This becomes a serious issue if significant time is needed to generate each predictor, e.g., as in neural networks.

The other is that the cross-validation estimate does not directly evaluate the aggregate hypothesis. None of the $10B$ predictors generated during 10-fold cross-validation become part of the bag (except by coincidence). It is an assumption that the performance of the hypotheses learned from the cross-validation folds will be similar to the performance of the hypothesis learned using the whole dataset (Kearns & Ron 1997).

One solution is to use the predictors in the bag to estimate generalization error (Breiman 1996b; Wolpert & Macready 1996; Tibshirani 1996). Each predictor is generated from a bootstrap sample, which typically omits about 37% of the examples. The *out-of-bag estimate* records the votes of each predictor over the examples omitted from its corresponding bootstrap sample. The aggregation of the votes followed by plurality voting for each example results in an estimate of generalization error.

We performed experiments on 10 two-class datasets. We used ID3 (Quinlan 1986) and C4.5 (Quinlan 1993) to generate predictors. Generalization error is represented by the empirical error of the bag on a separate test set.

In these experiments, the out-of-bag estimate slightly overestimates generalization error on average. 10-fold cross-validation has similar behavior. A two-sample t test (the two samples are the training examples and test examples), for both the out-of-bag estimate and 10-fold cross-validation, has very inaccurate (much too optimistic) confidence levels. In several cases, a supposedly 95% confidence interval corresponds to less than 90% empirically; in one case, less than 75%.

Previous research (Kohavi 1995) has shown that 10-fold cross-validation tends to have a pessimistic bias, i.e., the estimated error rate tends to have a higher expected value

than the true error rate. Besides duplicating this finding in the context of bagging, our methodology uses a better statistical test and also studies the confidence intervals of the estimates.

We can improve the out-of-bag estimate by incorporating a correction. If there are B predictors in the bag, then there are B votes for each test example compared to about $0.37B$ out-of-bag votes on average for each training example. We propose two corrections by taking this factor into account.

One correction is based on the voting patterns in the test examples, where a voting pattern is specified by the number of votes for each class, e.g., 29 votes for class A and 21 votes for class B. This correction is not practical for estimating generalization error because the test-example voting-patterns are unknown if all available examples are used for training. However, we gain an understanding of what is needed to calculate a correction.

For a given test example, we can simulate out-of-bag voting by drawing a subsample of the votes on the test examples, i.e., each vote is selected with probability $1/e$. For two-class datasets, we can directly compute two values: the expected value and the variance of the difference between the simulated out-of-bag voting and test error. Using these statistics and an appropriate t test, we obtain acceptable confidence intervals in our experiments.

Our second correction tries to reverse this process. It uses the out-of-bag voting patterns on the training examples to estimate the distribution of B-vote patterns. Based on this estimated distribution, we compute the expected value and variance of the difference between the out-of-bag estimate and B-vote voting. This second correction has a heuristic component because it (roughly) assumes that the B-vote distribution is selected from a a uniform distribution of B-vote distributions. Perhaps for this reason, the second estimate often leads to optimistic confidence levels, though they are better than the uncorrected out-of-bag estimate.

The remainder of this paper is organized as follows. First, we describe the experimental procedure. Next, we provide the results of the experiments and their implications. Finally, we conclude with a summary and future research issues.

Experimental Procedure

We selected a number of two-class datasets from the UCI repository and the C4.5 distribution (certain quantities are easier to compute precisely with two-class datasets). Several of these datasets were used extensively to develop the generalization error estimates. The other datasets (see the Appendix) were used for the experiments presented in this paper.

We used two learning algorithms. One algorithm was C4.5 using default parameters (Quinlan 1993). We also used the ID3 decision-tree learning algorithm with no pruning (Quinlan 1986). In our version of ID3, missing values are handled by creating an extra branch from each internal node to represent the case of a missing value. If there are no examples for a leaf node, it is given a classification equal to the most common class of its parent.

For this paper, the following procedure for experimenting with the bagging method was used:

1. The data set is randomly divided in half to create a training set S and a test set T.
2. A bootstrap sample S_1^* is selected from S and a predictor is created from S_1^* using a learning algorithm. This is repeated B times to create B predictors, $h_1, \ldots, h_B$, from the B bootstrap samples $S_1^*, \ldots, S_B^*$.
3. The out-of-bag estimate is determined from the training set S by allowing each predictor h_i to vote only on the examples $S - S_i^*$, i.e., the training examples omitted from the ith bootstrap sample. Then the predicted class of each example is determined by a plurality vote with ties broken in favor of the most common class in S. On average, about 37% of the examples are excluded from each bootstrap sample, so on average, about 37% of the predictors vote on each training example.
4. Test error is determined from the test set T by a plurality vote on each example over the B predictors. Ties are broken in favor of the most common class in S. Test error is considered to be an accurate estimate of generalization error.[1]
5. The above steps 1–4 are repeated 1000 times for each data set, learning algorithm, and value for B (we used $B = 50$). Averages and standard deviations for the out-of-bag estimate, test error, and the paired difference were computed. 1000 trials were used for two reasons. Any substantial difference in the averages ought to become statistically significant after 1000 trials. Also, we performed a two-sample t test on each trial with the anticipation of falling within the calculated 95% confidence interval at least 95% of the time, i.e., to determine if the confidence level of the test can be trusted.

Other Generalization Error Estimates

Besides the out-of-bag estimate, we also evaluated 10-fold cross-validation and two different corrections to the out-of-bag estimate.

10-Fold Cross-Validation For $B = 50$, we computed a 10-fold cross-validation estimate of generalization error. Cross validation has been widely accepted as a reliable method for calculating generalization accuracy (Michie, Spiegelhalter, & Taylor 1994; Weiss & Kulikowski 1991), and experiments have shown that cross validation is relatively unbiased (less biased than bootstrap sampling) (Efron & Tibshirani 1993). However, there is some evidence that 10-fold cross-validation has high type I error for comparing learning algorithms (Dietterich 1998a).

In order to compute the cross-validation estimate, a step is inserted between steps 4 and 5 in the procedure described

[1] This assumes that the examples in the dataset are independently drawn from some probability distribution, and that the probability mass of the training set S is near 0%. These assumptions are not true for at least the monks datasets. In this case, the generalization error estimates can be treated as estimates of error on examples outside of the training set.

above. In this new step, the training set S is partitioned into 10 cross-validation sets or folds of nearly equal size. Then for each cross-validation fold F_i, the examples $S - F_i$ are used with bagging to form B predictors. The resulting bag is used to classify the examples in F_i and produce an error measure. The average and variance are computed from the 10 iterations.

Test Error Correction In the out-of-bag estimate, there are about $0.37B$ out-of-bag votes on average for each training example. The test error is determined from B votes for each test example, so it might be expected that the out-of-bag voting would be inaccurate.

A "test error correction" is determined as follows. Given the voting patterns on the test examples, we can simulate out-of-bag voting by choosing each vote with probability $1/e$. This is expected to be a good simulation of the out-of-bag estimate because test examples and training examples should be interchangeable as far as out-of-bag voting is concerned. We did not perform the simulation, but instead directly computed the expected value of the out-of-bag simulation, using this value as the sample mean in statistical tests. For sample variance, we treated the mean as if it was the result of n Bernoulli trials, where n is the number of test examples. The second appendix describes the calculations in more detail.

Out-of-Bag Correction A serious disadvantage of the test error correction is its dependence on the voting on the test examples. We would like to estimate generalization error purely from the training examples so that all available examples can be used for training. So one might alternatively compute an "out-of-bag correction" based on the out-of-bag voting patterns by simulating B-vote patterns from the out-of-bag voting patterns. The difficulty here is that determining $P(E_B \mid E_O)$, where E_B and E_O are respectively events of some B-vote pattern and out-of-bag voting pattern, depends on the distribution of B-vote patterns.

We heuristically guess the distribution of B-vote patterns as follows. Two different distributions are calculated for the two classes. Consider one of the classes and the training examples with that class as their label. Let $\mathcal{U}$ be a probability distribution in which all B-vote patterns are equally likely, e.g., the pattern of 30 votes for class A and 20 votes for class B is as likely as 15 votes for class A and 35 votes for class B. We determine $P_{\mathcal{U}}(E_B \mid E_O)$ for each training example, sum up the probabilities for each B-vote pattern, and normalize them so they sum to 1; these are used to define a probability distribution $\mathcal{D}$ over the B-vote patterns. This distribution is adjusted to better account for the number of out-of-bag errors. Then we determine $P_{\mathcal{D}}(E_B \mid E_O)$ for each training example and compute the expected value of the simulated test error.

The two expected values from the two different distributions are combined and used as sample mean in statistical tests. For sample variance, we treated the mean as if it was the result of n Bernoulli trials, where n is the number of training examples. The second appendix describes the calculations in more detail.

Statistical Tests

The following statistical tests were employed to compare the results of different experiments over the 1000 trials. In our notation, μ_i, $\overline{X_i}$, and s_i^2 are respectively the expected value, the sample average, and sample variance over n_i samples.

A paired difference t test (paired comparison of means) was performed over 1000 pairs to compare the four different estimates of generalization error with test error. This test evaluates the hypothesis that the average estimate of generalization error has the same expected value as the average test error. To pass this test with a 5% significance level, the magnitude of the t value should be no more than 1.962. t is computed by:

$$t = \frac{\overline{X_1}}{\sqrt{s_1^2/n_1}} \tag{1}$$

where $n_1 = 1000$ in our experiments

Two different *t* tests (unpaired comparison of means) was performed on each trial to determine whether the generalization error estimate on that trial was significantly different (5% level) from the test error on that trial. That is, for each trial, we evaluate the hypothesis that the generalization error estimate has the same expected value as the test error. Because this hypothesis is tested for each of 1000 trials, we can count the number of trials that fail the test and see if the number of failures is about 5% (or less). This would imply that a 95% confidence interval according to the test appears to correspond to a 95% confidence interval in reality. For a binomial distribution with probability of success $p = .95$, the probability of 939 successes or more (61 failures or less) is about 95%.

The number of failures should be interpreted cautiously in our experiments. Simply splitting the file into a training set and test set generally results in a negative correlation between the error estimates and the test error. This appears to be because the "bad" examples might not be proportionally distributed between the two sets, which causes the error estimate and the test error to go in opposite directions.

For the out-of-bag estimate, the test correction, and the out-of-bag correction, we used the standard two-sample t test assuming equal variances (Cohen 1995; Snedecor & Cochran 1980). To test the hypothesis $\mu_1 = \mu_2$, we compute

$$s_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{2}$$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{s_{pooled}^2(1/n_1 + 1/n_2)}} \tag{3}$$

and compare t against the critical value for $n_1 + n_2 - 2$ degrees of freedom. Here, n_1 and n_2 are the number of examples in the training set and test set, respectively.

For 10-fold cross-validation, we used a two-sample t test allowing for unequal variances (Snedecor & Cochran 1980). This is because the variance over $n_1 = 10$ folds will

be much different from the variance over n_2 test examples. To test the hypothesis $\mu_1 = \mu_2$, we compute

$$a_i = \frac{s_i^2}{n_i}, \qquad v_i = n_i - 1 \tag{4}$$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{a_1 + a_2}} \tag{5}$$

$$v = \frac{(a_1 + a_2)^2}{a_1^2/v_1 + a_2^2/v_2} \tag{6}$$

and compare t against the critical value for $\lfloor v \rfloor$ degrees of freedom. Here, n_1 and n_2 are the number of examples in the training set and test set, respectively.

Justification for The Two-Sample t Tests

An informal survey of several statistics texts indicated a variety of ways to estimate the degrees of freedom v for a two-sample t test with unequal variances. A common, but cautious, approach is to set v to the minimum of v_1 and v_2. However, this would put the uncorrected out-of-bag estimate at a disadvantage because it uses the more stringent two-sample t test assuming equal variances. The calculation of v by Equation (6) appears to have some acceptance as a better approximation.

It is unclear what sample variance should be used for the test correction and out-of-bag correction. Our choice is in some sense "fair" because it ensures that the size of the confidence interval will be similar to that used to the out-of-bag estimate. However, one might derive a sample variance for the test correction from the probabilities derived for each test example (likewise for the out-of-bag correction and each training example). We tried this, but it fails badly empirically because it leads to an artificially small variance. We believe the reason is that incorporating the corrections corresponds to adding more variance rather than reducing variance. In particular, our results suggest that additional variance should be added for the out-of-bag correction, but this would give the correction an apparently "unfair" advantage.

Results

Bagging Estimate

Table 1 shows the statistics that were collected for the 10 data sets using $B = 50$ predictors. The first column gives the abbreviations for the datasets (see the Appendix), and the second column gives the test error percentage. The next three columns compares the out-of-bag estimate to test error: the percent difference between the averages, the t value from the paired-difference t test over 1000 pairs, and the number of failures for 1000 applications of the two-sample t test at 5% significance. The last three columns provide the same information comparing 10-fold cross-validation with the out-of-bag estimate.

The table shows that both the out-of-bag estimate and 10-fold cross-validation can be expected to slightly overestimate test error. On average, the out-of-bag estimate differed from test error by 0.52% on average, and 10-fold cross-validation differed by 0.25%. The paired-difference t test showed a significant difference (5% level, $|t| > 1.962$) between the generalization error estimates and test error for all the datasets except FL.

Data Set	Test Error	OOB − Test Error Diff.	t	Fails	CV − Test Error Diff.	t	Fails
BA	27.55	0.88	6.20	111	0.31	1.99	112
CR	16.24	0.43	4.33	72	0.27	2.76	77
FL	20.52	0.07	0.85	78	0.15	1.69	82
IO	8.18	0.46	4.59	61	0.26	2.60	68
M1	1.77	0.78	13.03	231	1.02	15.28	115
M2	51.93	−0.90	−5.40	79	−1.39	−8.42	93
M3	0.00	0.004	2.31	0	0.014	3.77	0
PI	24.74	0.71	6.96	57	0.22	2.15	72
PR	17.84	1.70	5.59	119	1.01	3.31	111
SO	22.81	1.05	5.17	82	0.64	3.21	63
Average		0.52	4.36	89	0.25	2.83	79

Table 1: Results for ID3, $B = 50$: Out-of-Bag Estimate and 10-Fold Cross-Validation

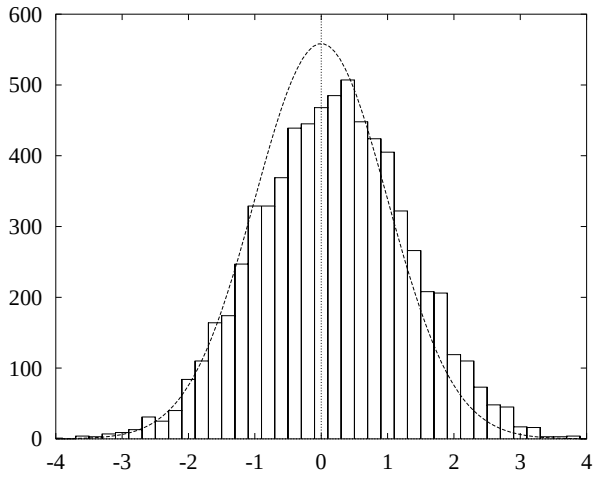

Figure 1: Histogram of t Values Comparing the Out-of-Bag Estimate to Test Error

The results of the two-sample t test also show that the out-of-bag estimate and 10-fold cross-validation yield similar performance. For most of the datasets, the two-sample t test has an overoptimistic confidence level (more than 61 failures). That is, a 95% confidence interval according to the test does not correspond to a 95% confidence interval for either the out-of-bag estimate or 10-fold cross-validation. On average, around 8% to 9% of the trials fail the test, with especially poor performance on the BA, M1, and PR datasets (over 100 fails each).

Figure 1 displays a histogram of the t values for 7 of the datasets (excluding M1, M2, and M3 because of extreme error rates). The normal density function is superimposed. It can be seen that the histogram has a normal-like shape, but flatter and skewed to the right (corresponding to where test error is overestimated).

Data Set	OOB − Test Error − Test Error Corr. Diff.	t	Fails	OOB − Test Error − OOB Correction Diff.	t	Fails
BA	−0.02	−0.14	57	−0.19	−1.34	97
CR	0.06	0.59	61	−0.05	−0.55	71
FL	0.02	0.19	63	−0.04	−0.46	85
IO	0.05	0.57	43	−0.01	−0.13	54
M1	0.07	1.21	52	0.05	0.80	115
M2	−0.05	−0.31	75	−0.24	−1.38	91
M3	0.003	1.81	0	0.003	2.39	0
PI	0.16	1.64	42	−0.06	−0.61	50
PR	0.06	0.23	61	0.10	0.32	113
SO	0.05	0.27	46	−0.32	−1.62	73
Avg.	0.04	0.61	50	−0.08	−0.26	75

Table 2: Results for ID3, $B = 50$: Corrections to the Out-of-Bag Estimate

Corrections to the Out-of-Bag Estimate

Table 2 shows the statistics that were collected for the 10 data sets using the corrections to the out-of-bag estimate. The first column gives the abbreviations for the datasets (see the Appendix). The next three columns compare the out-of-bag estimate to the test error and the test error correction: the percent difference between the averages, the t value from the paired-difference t test over 1000 pairs, and the number of passes for 1000 applications of a multiple-sample t test at 5% significance. The last three columns provide the same information comparing the out-of-bag estimate to the test error and the out-of-bag correction.

Table 2 shows that both corrections lead to better estimates of generalization error. The test error correction differed from by 0.04% on average, and so, comes very close to an unbiased estimate of generalization error. The paired-difference t test shows no significant difference (5% level, $|t| > 1.962$) on any of the datasets, though the test error is still overestimated on 8 of the 10 datasets. The out-of-bag correction is almost as good according to the paired-difference t test, with an average difference of -0.08% and with a significant difference on only one of the 10 datasets, but is better than the uncorrected out-of-bag estimate and 10-fold cross-validation (Table 1).

The results of the two-sample t test also show that the test error correction has excellent performance and is better than the out-of-bag correction, which in turn is slightly better than the uncorrected out-of-bag estimate and 10-fold cross-validation. This t test with the test error correction empirically provides a much more acceptable confidence level; there were more than 61 failures only 2 of the datasets. The t test with the out-of-bag correction is less acceptable with more than 61 failures on 7 of the 10 datasets. The slightly lower average compared to the out-of-bag and 10-fold cross-validation estimates is encouraging, but not statistically significant.

	OOB	10CV	TEC	OOBC
Rejections Paired t	7	6	0	2
Avg. Number of Fails	110	96	55	99

Table 3: Results for C4.5, $B = 50$: Generalization Error Estimates

Additional Results

For C4.5 and $B = 50$, we obtained results for all four generalization error estimates, summarized in Table 3. The columns correspond to the difference estimates (out-of-bag estimate, 10-fold cross-validation, test error correction, out-of-bag correction). The first row shows the number of datasets in which the paired difference t shows that the generalization error estimate is biased, i.e., rejects the hypothesis that the average estimate of generalization error has the same expected value as the average test error. The second row shows the average number of failures using the two-sample t test. Failure means rejecting the hypothesis that the generalization error estimate has the same expected value as test error on that trial.

It can be seen that the test correction again clearly outperforms the other estimates. Compared to the out-of-bag and 10-fold cross-validation estimates, the out-of-bag correction again has less bias based on the number of paired-t rejections, but has similar performance on the 2-sample t test. In the detailed results (not shown), by far the worse performance was by the out-of-bag estimate, 10-fold cross-validation, and out-of-bag correction on the M1 dataset (292, 273, and 202 fails, respectively). The out-of-bag correction also performed badly on the M1 and PR datasets (147 and 151 fails, respectively, but still better than OOB and 10CV). The test error correction had 61 or fewer failures on 8 of the datasets, with the worse performance on the PR dataset (93 fails).

Conclusion

With the use of any learning algorithm, it is important to use as many examples as possible for training the hypothesis (or hypotheses) from a dataset. It is also important to determine a good estimate of generalization error so that we can have confidence that a good hypothesis has been learned. Our methodology statistically compares an estimate of generalization error determined from a training set to the empirical error on a separate test set.

Cross-validation is one way of estimating generalization error, while using all of the examples for training, but our experiments have shown that it is biased and can provide inaccurate confidence interval estimates of generalization error. When bagging is used, the out-of-bag estimate can be to estimate generalization error, and it also uses all examples that are available. Unfortunately, the out-of-bag estimate is also biased and leads to similarly inaccurate confidence intervals.

We have developed corrections that improve the out-of-bag estimate and outperform 10-fold cross-validation. A test error correction, i.e., based on the voting patterns on the

test examples, empirically provides a nearly unbiased estimate of generalization error and leads to good confidence intervals. However, this correction is not practical because it cannot be applied until the bag is evaluated on examples outside of the training set.

We also developed an out-of-bag correction, i.e., based on the voting patterns on the training examples. This correction makes an assumption about the distribution of domains, and so, must be regarded as heuristic. Perhaps as a result, this correction is not as good as the previous correction. The out-of-bag correction is relatively unbiased compared to 10-fold cross-validation and an uncorrected out-of-bag estimate, but does not significantly improve the accuracy of the confidence intervals.

We conclude that 10-fold cross-validation and the uncorrected out-of-bag estimate should be cautiously used for generalization error estimates because they can result in confidence levels that are much too high. We recommend the out-of-bag estimate with a correction based on the voting patterns on the training examples. The corrected out-of-bag estimate uses all the data, is unbiased, and avoids the additional time needed for 10-fold cross-validation; however, it still often leads to inflated confidence levels. Further research is needed to develop generalization error estimates with confidence intervals that can be fully trusted.

Acknowledgments

This research was funded in part by Texas Higher Education Coordinating Board grant ARP-225. We thank the anonymous reviewers for their comments.

Appendix: Datasets

For each dataset, we list our abbreviation, the number of examples, the number of attributes, and a brief description. The datasets come from the Irvine dataset (Blake, Keogh, & Merz 1998) or the C4.5 distribution (Quinlan 1993). We did not consider larger datasets because of the time required to perform bagging and 10-fold classification multiple times.

BA, 550, 35. The UCI cylinder bands dataset.

CR, 690, 15. The C4.5 credit card applications dataset.

FL, 1066, 10. The UCI solar flare dataset. This was changed to a two-class dataset: any flare activity vs. no flare activity.

IO, 351, 34. The UCI ionosphere dataset.

M1, 432, 6. The C4.5 monk 1 dataset.

M2, 432, 6. The C4.5 monk 2 dataset.

M3, 432, 6. The C4.5 monk 3 dataset. The dataset in the C4.5 distribution has no classification noise.

PI, 768, 8. The UCI Pima Indians diabetes dataset.

PR, 106, 57. The UCI promoter gene sequence dataset.

SO, 208, 60. The UCI sonar signals dataset.

Appendix: Deriving the Corrections

For a given trial with a two-class dataset, designate one class to be the majority class, and let the other class be the minority class. The majority class is determined using the training set. Assume that B predictors are in the bag.

For a given example, let $E_B(x, y)$ be the event of x votes for the majority class and y votes for the minority class, where $x + y = B$. Let $E_O(u, v)$ be the event of u votes for the majority class and v votes for the minority class, where the votes are a subsample of the B votes, where each vote is independently selected to be in the subsample with probability $1/e$. That is, we treat out-of-bag voting as if we were taking a subsample of B votes on that example. We call this "out-of-bag sampling." A probability distribution is specified by assigned priors to $P(E_B(x, y))$.

We note that:

$$\begin{aligned} & P(E_O(u, v) \mid E_B(x, y)) \\ & \quad = \quad b(u, x, 1/e)b(v, y, 1/e) \end{aligned} \tag{7}$$

where $b(k, n, p)$ is the probability of k successes in n i.i.d. Bernouilli trials, each with probability of success $1/e$. That is, $E_O(u, v)$ means that u of the x votes for the majority class were chosen, and v of the y votes for the minority class were chosen.

Test Error Correction

For test example i, let x be the number of votes for the majority class, let y be the number of votes for the minority class, and let $l(i)$ be the example's label, with a label of 1 corresponding to the majority class. Here, $x + y = B$.

We can then determine the probability that out-of-bag sampling favors the majority class or the minority class:

$$p_1(i) = \sum_{u \geq v} P(E_O(u, v) \mid E_B(x, y)$$

$$p_0(i) = \sum_{u < v} P(E_O(u, v) \mid E_B(x, y)$$

and the expected error by out-of-bag sampling from the B votes.

$$\mu(i) = \begin{cases} p_1(i) & \text{if } x < y \wedge l(i) = 0 \\ 1 - p_1(i) & \text{if } x < y \wedge l(i) = 1 \\ p_0(i) & \text{if } x \geq y \wedge l(i) = 1 \\ 1 - p_0(i) & \text{if } x \geq y \wedge l(i) = 0 \end{cases}$$

Over n test examples, we obtain a sample mean and sample variance:

$$\mu = \frac{\sum_{i=1}^{n} \mu(i)}{n}$$

$$s^2 = \frac{n(\mu - \mu^2)}{n - 1}$$

μ and s^2 are the values used in our two-sample t test. The sample variance is based on treating μ as the sample mean of n Bernoulli trials.

Out-of-Bag Correction

For training example i, let u be the number of votes for the majority class, and let v be the number of votes for the minority class. Here, $u + v$ on average will be about B/e. The out-of-bag correction is based on estimating the distribution of $E_B(x, y)$ based on the out-of-bag votes. This is done by pretending that the out-of-bag voting was really out-of-bag

sampling from B votes and assuming that each $E_B(x,y)$ is equally likely. We actually estimate two different distributions: one for when the label is the majority class, and the other for the minority class. Consider, then, those n training examples that have a majority class label.

Define $P_{\mathcal{U}}(E_B(x,y)) = 1/(B+1)$ for $x \in \{0,1,\ldots,B\}$ and $y = B - x$. We can then derive:

$$\begin{aligned} & P_{\mathcal{U}}(E_B(x,y) \mid E_O(u,v)) \\ = \; & \frac{P_{\mathcal{U}}(E_O(u,v) \mid E_B(x,y)) P_{\mathcal{U}}(E_B(x,y))}{P_{\mathcal{U}}(E_O(u,v)} \\ \sim \; & P_{\mathcal{U}}(E_O(u,v) \mid E_B(x,y)) \end{aligned}$$

because $P_{\mathcal{U}}(E_B(x,y))$ is a constant and the denominator is a normalizing term. Equation (7) gives the calculations.

We define an intermediate probability distribution $\mathcal{I}$:

$$P_{\mathcal{I}}(E_B(x,y)) = \theta \sum_{i=1}^{n} P_{\mathcal{U}}(E_B(x,y) \mid E_O(u,v))$$

setting θ so that $1 = \sum_{j=0}^{B} P_{\mathcal{I}}(E_B(j, B-j))$.

This probability distribution implies the probability that out-of-bag voting will result in predicting the majority class and minority class.

$$c_1 = \sum_{u>=v} P_{\mathcal{I}}(E_O(u,v))$$

Let $c_0 = 1 - c_1$. Let d_1 be the percentage of training examples that favor the majority class in out-of-bag voting. Let $d_0 = 1 - d_1$.

We obtain the probability distribution $\mathcal{D}$ by:

$$P_{\mathcal{D}}(E_B(x,y)) \sim \begin{cases} d_1 P_{\mathcal{I}}(E_B(x,y))/c_1 & \text{if } x >= y \\ d_0 P_{\mathcal{I}}(E_B(x,y))/c_0 & \text{if } x < y \end{cases}$$

These probabilities are normalized so they sum to 1. This mostly, but not completely, adjusts the probabilities so that the distribution of majority/minority class predictions corresponds to the out-of-bag voting.

Assuming this probability distribution, we estimate the B-vote predictions based on the out-of-bag voting by calculating $P_{\mathcal{D}}(E_B(x,y) \mid E_O(u,v))$. For training example i, we can determine the probability that B-vote voting favors the majority class ($x \geq y$) and the minority class ($x < y$). By adding the probabilities over all the examples, we estimate the number of majority and minority class predictions. If we are considering the training examples with a majority (minority) class label, then the number of minority (majority) class predictions is the estimated number of test errors.

Let μ be the estimated percentage of test errors by combining the results from the two distributions. Let n now be the total number of training examples. Then we use

$$s^2 = \frac{n(\mu - \mu^2)}{n-1}$$

The sample variance is based on treating μ as the sample mean of n Bernoulli trials.

References

Blake, C.; Keogh, E.; and Merz, C. J. 1998. UCI repository of machine learning databases.

Breiman, L. 1996a. Bagging predictors. *Machine Learning* 24:123–140.

Breiman, L. 1996b. Out-of-bag estimation. Technical report, Dept. of Statistics, Univ. of Calif., Berkeley. ftp://ftp.stat.berkeley.edu/pub/users/breiman/ OOBestimation.ps.Z.

Cohen, P. R. 1995. *Empirical Methods for Artificial Intelligence*. Cambridge, MA: MIT Press.

Dietterich, T. G. 1998a. Approximate statistical tests for comparing supervised classification algorithms. *Neural Computation* 10:1895–1924.

Dietterich, T. G. 1998b. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Technical report, Dept. of Computer Science, Oregon State Univ. ftp://ftp.cs.orst.edu/pub/tgd/papers/tr-randomized-c4.ps.gz.

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Freund, Y., and Schapire, R. E. 1996. Experiments with a new boosting algorithm. In *ICML-96*, 148–156.

Kearns, M. J., and Ron, D. 1997. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proc. Tenth Annual Conf. on Computational Learning Theory*, 152–162.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI-95*, 1137–1143.

Maclin, R., and Opitz, D. 1997. An empirical evaluation of bagging and boosting. In *AAAI-97*, 546–551.

Michie, D.; Spiegelhalter, D. J.; and Taylor, C. C. 1994. *Machine Learning, Neural and Statistical Classification*. Englewood Cliffs, NJ: Prentice Hall.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1:81–106.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Quinlan, J. R. 1996. Bagging, boosting, and C4.5. In *AAAI-96*, 725–730.

Snedecor, G. W., and Cochran, W. G. 1980. *Statistical Methods*. Ames, IA: Iowa State Univ. Press.

Tibshirani, R. 1996. Bias, variance and prediction error for classication rules. Technical report, Department of Statistics, University of Toronto. http://www-stat.stanford.edu/∼tibs/ftp/biasvar.ps.

Weiss, S. M., and Kulikowski, C. A., eds. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Mateo, CA: Morgan Kaufmann.

Wolpert, D. H., and Macready, W. G. 1996. An efficient method to estimate bagging's generalization error. Technical report, Santa Fe Institute. http://www.santafe.edu/sfi/publications/Working-Papers/96-06-038.ps.