Decision Tree Pruning: Biased or Optimal?

Sholom M. Weiss[†] **and Nitin Indurkhya**[‡]
† Department of Computer Science, Rutgers University
New Brunswick, New Jersey 08903, USA
‡ Department of Computer Science, University of Sydney
Sydney, NSW 2006, AUSTRALIA

Abstract

We evaluate the performance of weakest-link pruning of decision trees using cross-validation. This technique maps tree pruning into a problem of tree selection: Find the best (i.e. the right-sized) tree, from a set of trees ranging in size from the unpruned tree to a null tree. For samples with at least 200 cases, extensive empirical evidence supports the following conclusions relative to tree selection: (a) 10-fold cross-validation is nearly unbiased; (b) not pruning a covering tree is highly biased; (c) 10-fold cross-validation is consistent with optimal tree selection for large sample sizes and (d) the accuracy of tree selection by 10-fold cross-validation is largely dependent on sample size, irrespective of the population distribution.

Introduction

Decision trees methods have evolved from straightforward recursive partitioning algorithms that cover sample data to more complex techniques that also prune the covering tree and estimate future performance (Breiman *et al.* 1984; Quinlan 1993). The motivation for pruning a tree is to maximize predictive performance, which is often described as "overfitting avoidance." However, too much pruning can readily lead to "underfitting," and a more appropriate objective would be to find "the right size" tree.

Many techniques have evolved over the years for pruning trees to the right size. Practical experience has also led to the adaptation of these techniques to alternative learning models such as rule induction (Cohen 1993; Weiss & Indurkhya 1993) or neural nets (Hassibi & Stork 1993). With a very large set of independent test data, there is little difficulty in describing the efficacy of pruning. The estimated error rate and the standard error of the estimate have a precise formal description.

When very large numbers of independent test cases are not available, relatively complex techniques involving resampling can be employed to estimate performance and to select the pruned tree. Resampling with decision trees is more complex than for other classifiers. In addition to generating multiple trees, these trees must be pruned such that the complexities of the subtrees are matched for each subsample. In a series of papers studying the effects of pruning on decision tree performance (Schaffer 1992b; 1992a; Wolpert 1992; Schaffer 1993), it was demonstrated that pruning does not always lead to improved results. Moreover, in some instances it may even degrade performance. Generalizing from these experimental results, often with small samples, the authors of these studies concluded that pruning using cross-validation is inevitably biased and is often ineffective without knowledge of the sampled population.

In this paper, we reconsider the efficacy of decision tree pruning. Tree pruning is mapped into a problem of tree selection among competing subtrees. Previous experiment results are reevaluated and additional experiments are performed. We address several major issues such as the bias of tree pruning by cross-validation, the effect of sample size, the divergence from optimal selection, and the extent to which knowledge about overall population characteristics is essential for accurate results.

We reserve our discussion in this paper to the most general case: samples of moderate to large size, samples with at least 200 cases. Small samples, with their attendant high variability, require special attention (Efron 1983; Crawford 1989; Weiss 1991) and are discussed in a separate paper (Weiss & Indurkhya 1994).

Tree Pruning and Selection

Tree induction methods generate a covering tree to discriminate the training data. For generalization to new cases, a subtree of the covering tree may actually make fewer errors on new cases. Hence the use of pruning techniques that excise portions of the covering tree. Pruning can be described in the following general terms:

- Generate a set of "interesting" trees;
- Estimate the true performance of each of these trees;
- Select the best tree.

Although there are a number of pruning techniques (Quinlan 1987; Cestnik & Bratko 1991), a prime example of a form of pruning that matches these steps is weakest link (cost-complexity) pruning (Breiman *et al.* 1984). A covering tree is recursively pruned into a series of subtrees, based on eliminating the weak points

T_i i=	Nodes	Err_{app}	Err_{test}	$Test_{SE}$
0	18	.0000	.1074	.0282
1	15	.0083	.0909	.0261
2	13	.0165	.0909	.0261
3*	7	.0661	.0744	.0239
4	6	.0826	.1322	.0308
5	4	.1322	.1322	.0308
6	3	.2975	.2975	.0416
7	2	.5372	.5620	.0451
8	1	.6529	.6529	.0433

Table 1: Example of Summary Table for Tree Pruning

of the current tree. These weak points are determined strictly from the training data.

Having obtained a set of decision trees, $(T_0, \ldots, T_n)$ one is now faced with the *tree selection problem*: given a set of trees, select the best one. The usual definition of *best* is that of the minimum true error rate, which must be estimated. It is useful to order the set of trees by some complexity measure such as tree size. If the set of trees is obtained by pruning, then T_0 is the unpruned covering tree, and T_n is a tree that consists only of the root node. Figure 1 gives an example of a pruning summary table, such as found in CART, with the covering tree T_0 having 18 terminal nodes and T_8 representing the fully pruned tree with a single terminal node. Err_{test} is the estimate of the true error rate for each tree, and $Test_{SE}$ is an estimate of the standard error of the error rate. In this example, T_3 is selected because it has the minimum estimated true error rate.

Thus, tree pruning is mapped into a problem of tree selection: Find the best tree, i.e the right-sized tree, from a set of trees ranging in size from the unpruned tree to a null tree. Tree selection does not depend on the techniques for generating the trees. Error estimation is the sole basis of tree selection; the tree with the lowest error-estimate is selected. The quality of the results depends on the accuracy of these estimates. Several error-estimation procedures might be hypothesized:

Ideal: The ideal situation occurs when an oracle is available that can tell us the future performance of each decision tree. Then we will be able to make the optimal tree selection. Such an oracle is usually approximated accurately by testing each tree, T_i, on a very large, independent test set.

NP: While we would like to use an oracle-based method, this may not be possible if insufficient cases are available. One strategy might be to base decisions on the *apparent* error for the training cases. Because the apparent error rate is minimum for the covering tree, this strategy reduces to not pruning the initial covering tree.

Cross-Validation: When large numbers of independent test cases are not available, resampling methods are the principal technique for error rate estimation. Cross-validation is generally the procedure of choice, and 10-fold cross-validation (the test results of 10 runs using 90-percent training and 10-percent testing cases, with 10 mutually exclusive test partitions) has been widely used for many different learning models.

Our objective in the remainder of this paper is to compare the performance of tree pruning for these three alternative methods of estimating error rates.

Basic Principles

Fundamental Statistical Model of Evaluation

The standard model of evaluation of a learning system is by testing on an independent, randomly drawn sample from the general population. If performance is measured in terms of a proportion of failure, i.e. an error rate, then the situation corresponds to the binomial sampling model. This testing situation is the standard statistical coin tossing problem, where here we "toss" the classifier on each of the test cases. If we have n test cases, then there are n success or failure outcomes, each outcome representing a correct or incorrect classification of a test case. The standard error of this proportion is given in Equation 1, where n is the test-set size and p is the *true error rate*. For a given sample size, the standard error roughly tells us the average amount that the error rate will diverge from the truth.

$$Variance = \frac{p(1-p)}{n}; SE = \sqrt{Variance} \quad (1)$$

We have a statistical model of how far off the error estimate for a single test sample is from the truth. With unlimited test samples, the efficacy of pruning would be obvious. The pruned tree with the minimum test error is the best to a very high degree of confidence. Just based on the variation among random samples, the error rate on test cases will vary from the truth according to Equation 1. This variance is based solely on two terms, the true error rate, p, and the test size n. Considering the range of p, the worst case (i.e. the highest variance) is for p=.5. However, the true error rate, p, has a relatively minor effect on the variance, and the key factor is n, the test-set size. The accuracy of the evaluation on the test cases is mostly determined by test-sample size. When n is large enough the standard error becomes quite small.

Given only a single sample, without large numbers of test cases, the task is to estimate the true error rate. Resampling techniques such as cross-validation attempt to approach the performance of testing on the same number of independent cases, while still using the full sample for training purposes. Resampled estimators are still subject to the random variation of the sample. At best, the resampled estimates reflect the error-rate for treating the sample as an independent test set. Their variance from the true error-rate would approximately follow Equation 1, their accuracy mostly dependent on the sample size n, and independent of the original population distribution.

Estimation and Tree Selection

Bias and Consistency of Estimators An estimator, x, of a metric (such as an error-rate) is *unbiased* if

its expected value (i.e. the average of its values over all samples) is equal to the true value of the metric. If sufficient number of independent random samples, N, are used, then for an unbiased estimator, Equation 2, summarizes this relationship, where x is the estimator, X_i is its mean value for the i-th sample, $T(x)$ is the true value of the metric being estimated by x and N is the number of samples.

$$T(x) = \frac{\sum_{i=1}^{N} X_i}{N} \quad (2)$$

The key concept of an unbiased estimator is that over a large enough set of independent samples it averages to the true answer. It may vary from sample to sample, but over all samples the average is correct. An example of an unbiased estimator is the error-rate estimate on an independent test set. While the estimate from a particular test set may differ from the true value, the average value of the estimate over all possible (independently sampled) test sets is the same as the true value. There is some empirical evidence that suggests that cross-validated estimates are relatively unbiased under quite general conditions (Efron 1983). While an unbiased estimator averages to the the true value, it is also desirable that the estimate tend to be close to the true value. Equation 1 shows how close a typical estimate will be for a given sample size.

Another desirable statistical property of an estimator is *consistency*: results improve with increasing sample size. For example, error-rate estimation from an independent test set is consistent. As the test-sample size increases, the error-rate estimate varies less and less from the true error-rate.

Optimality and Unbiased Tree Selection Pruning can be posed as a problem of tree selection with the objective of minimizing the true error rate. An optimal procedure always selects the best tree from the set of pruned trees generated for a sample. Such a procedure would be obtained if ideal error-rates were available. In their absence, we must rely on estimates.

While we may use estimates of error rates for tree selection and pruning, the absolute magnitude of these estimates is not critical. Instead, the relative ranking is critical. As long as the relative ranking (in terms of error-rates) of the pruned trees is correct, then the right-size tree can be selected. If estimators are used for tree selection, the tree selection bias should be measured. An appropriate measure of bias is the average size of trees that are selected. An optimal tree selection procedure will always select the right-sized tree for each sample. However, an unbiased procedure is not necessarily optimal. An unbiased procedure may select the wrong-sized tree for any given sample. Although these trees may range from undersized to oversized, the procedure can be considered unbiased if the average size over many samples is correct.

Bias is one of two principal components of error in estimation. The other is variance. As indicated by Equation 1, samples randomly drawn for a large population will vary. They are not a perfect reflection of the general population. The variance decreases with increasing sample size. Thus, it is not unusual when we flip an honest coin ten times, that we will see seven heads. But if we flip it a thousand times, we are far less likely to see seven hundred heads. When the sample size grows large, the variance decreases greatly. An unbiased strategy with zero variance is an optimal strategy. As the sample size increases, the variance should move closer to zero and an unbiased strategy should also approach an optimal strategy.

Data	Cases		Feature type	Classes
	Train/Test	Features		
Mush	8124	122	Boolean	2
Hypo	3772	22	Mixed	2
Hyper	3772/3428	22	Mixed	2
Pb	1494	2	Numer.	10
Wave	pgm/5000	21	Numer.	3
Letter	20000	16	Numer.	26
Heart	282	13	Numer.	2
German	6479	80	Numer.	2
LED	pgm/10000	7	Boolean	10
Noise	5000	10	Numer.	2

Table 2: Dataset Characteristics

The classical formal definition of statistical bias may differ from the descriptions given in the machine learning literature (Schaffer 1993; Mitchell 1990; Utgoff 1986), where a reader might conclude that unbiased estimators are optimal. The fundamental statistical concept of bias recognizes that predictive error is not attributable solely to the bias of a decision model. Instead, the problem may be with the sample! Inaccuracy of estimation can be a byproduct of random sampling variance, particularly for small samples that diverge greatly from the general population characteristics.

With a large enough sample, an unbiased tree selection strategy should approach an optimal solution, but an unbiased strategy will not always beat a biased strategy. If the bias fits the characteristics of the population, then for samples drawn from that population, the biased strategy will be closer to the truth. For example, if someone always calls heads, then with a coin slightly biased for heads, that strategy should be superior. With a large enough sample of coin flips one would discover this, but for smaller samples inferior performances for unbiased guesses are unavoidable.

Sources of Error in Pruning

Even with unbiased estimation techniques, all induction and pruning algorithms are at the mercy of the random variance of a sample. There is also another inherent source of error. When estimating error rates, cross-validation will train on less than the full sample. During each train and test cycle, some of the data must be reserved for testing. The usual variation is 10-fold: 90% training and 10% percent testing. For error rate estimation, this means that the estimates are those for 90% trees, not 100%. Thus these estimates should be somewhat pessimistic. For tree selection and pruning, the situation may be somewhat better. The relative ranking is critical, not the absolute magnitude. Still the basis of the rankings is 90% trees, implying some weakness when the true answer is near the unpruned

Dataset	n	Ideal		10-cv		NP	
		Err	Size	Err	Size	Err	Size
Mush	200	.013	6.2	.014	6.3	.013	6.5
	500	.005	8.3	.005	8.3	.005	8.4
	1000	.002	9.8	.002	9.8	.002	10.3
Hypo	200	.018	3.5	.020	3.4	.020	4.4
	500	.010	5.5	.012	5.0	.011	6.5
	1000	.006	6.8	.007	6.7	.006	8.5
Heart	200	.211	11.4	.242	9.7	.265	34.7
Pb	200	.327	30.1	.341	31.5	.354	67.5
	500	.282	38.3	.292*	37.5	.328	158.8
	1000	.253	43.3	.262*	34.8	.313	298.3
Wave	200	.292	13.0	.303	13.9	.306	28.0
	500	.264	22.3	.272	23.5	.281	63.9
	1000	.247	35.5	.254*	34.2	.267	119.5
Letter	200	.574	70.5	.581	67.9	.573*	88.1
	500	.436	156.4	.442	145.5	.438	177.9
	1000	.357	282.3	.361	261.4	.358	299.8
German	200	.278	12.5	.292*	15.2	.307	40.6
	500	.245	15.8	.254*	19.5	.291	97.6
	1000	.230	20.3	.235*	22.0	.282	192.3
LED	200	.554	21.4	.569*	26.8	.584	50.1
	500	.520	28.9	.528	29.4	.536	67.2
	1000	.503	37.1	.509	38.9	.515	73.9
Noise	200	.251	1.0	.255*	1.3	.385	38.9
	500	.251	1.0	.252*	1.1	.387	98.3
	1000	.251	1.0	.252*	1.0	.386	194.9

Table 3: Comparison of Ideal, 10-cv and NP

tree.

Another potential source of error is more specific to trees. Error estimation by 10-fold cross-validation involves the somewhat complicated matching of tree complexity. As the sample size increases, this is a relatively accurate process. With smaller samples, the matching process is imperfect and some interpolation is required (Breiman *et al.* 1984).

We have noted the potential sources of error in tree pruning using cross-validation. We now examine how strongly these factors affect its performance, and we compare its performance to the hypothetically ideal solution and to a strategy of not pruning at all.

Methods

For purposes of comparison, the same datasets reported in (Schaffer 1992b; 1992a; 1993) were used in the simulations. Unlike previous experiments, we postulate a strong connection of sample size to performance. Thus, for each dataset, random samples of size 200, 500, and 1000 were drawn from the overall population. The true answer was determined by results on either the remainder of the dataset or where available a second independently drawn test set. In addition to the original datasets, four others were also considered. These include the following:

- Random noise for two classes with a prevalence of approximately 75% for one class.
- A two class problem with features representing word frequency counts in German Reuters news stories (Apté, Damerau, & Weiss 1994).
- The Peterson/Barney Vowel Formant Dataset in which two features (the first two formant values) are used to discriminate among ten vowel classes (Watrous 1991).
- The Waveform data discussed in (Breiman *et al.* 1984) with three classes and twenty one features all of which have added noise. The Bayes error-rate for this problem is 14%.

These added datasets allow us to examine a wider spectrum of true answers, with some falling near the unpruned tree and others far away. With the exception of the heart dataset, which only allowed for a size 200 sample, all datasets were large enough for both training and testing on relatively large numbers of cases. The characteristics of the datasets are described in Table 2. For some datasets, such as the hyperthyroid application, independent test data were available. For others, such as the letter recognition application, a random subset was drawn for training and the remaining case were used for testing.[1] For some applications, such as LED, the training data were generated dynamically by a program. In addition to the fixed sample size experiments, we also ran some experiments with even larger samples. These sizes were selected based on the number of available cases in the dataset. Each simulation

[1]For the german text data, a second set of 1888 independent test cases were used for the large training sample experiment.

encompassed at least 100 train and test trials.

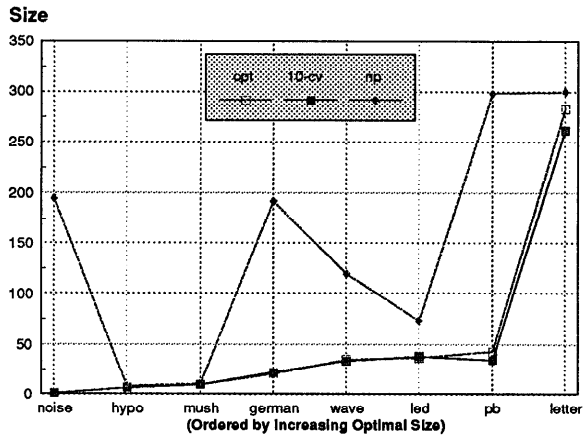

Figure 1: Bias: Tree Sizes for Size 1000 Samples

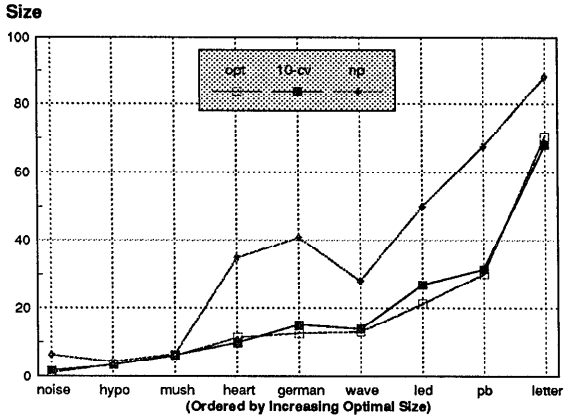

Figure 2: Bias: Tree Sizes for Size 200 Samples

The CART tree induction program was used in all experiments. The minimum error tree was selected by cross-validation. Ten-fold cross-validation was used in all experiments. The following slight modifications were made to the program:

- Each trial was initiated with a new random seed.
- Ties were broken in favor of the larger tree.

In the interest of experimental replication, many induction programs use the same random seed. In a laboratory setting, it may be beneficial to maximize randomness by reseeding after each trial. While it is tempting to break ties with the simpler tree, the 90% tree is actually being estimated, and therefore the larger tree is somewhat more likely for the full sample.

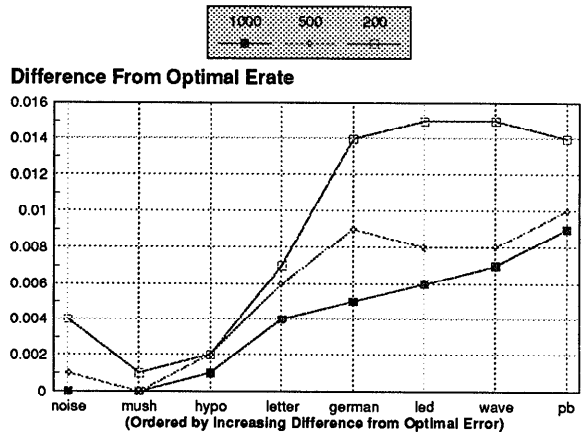

Figure 3: Consistency of 10-cv Performance for Varying Size Samples

As reported in (Schaffer 1992b), experiments were performed for cross-validation (10-cv), and not pruning (NP). The average error rates and sizes for the 10-cv and NP trees were recorded. Missing from the original analysis was crucial information about the average error rates and sizes for the hypothetically optimal tree-selection strategy (opt). In our experiments, this was determined by evaluating each of the ordered pruned trees directly on the independent test data.

Results

The results of the experiments for the fixed-sized samples are listed in Table 3. Table 4 lists the results for even larger sample sizes. Differences between NP and 10-cv of more than 2 standard errors (>95% confidence) are noted by a "*". Figure 1, plots the tree sizes for NP, 10-cv and opt for size 1000 samples; Figure 2 plots them for size 200 samples. Figure 3, compares the difference of 10-cv from the optimal error rate for sample sizes 200, 500, and 1000. Figure 4 plots the difference from the optimal error rate for NP and 10-cv for size 1000 samples; Figure 5 plots this difference for size 200 samples.

Significance Testing

For binomial trials, such as estimating error rates, the variance can be directly computed from Equation 1, and 2 standard errors is a reasonable significance test. In those instances where the dataset is randomly partitioned into train and test partitions, the standard error for a single trial is computed with n equal to the size of the test set. For many multiple trials, n approaches the full sample size, which is usually used to estimate the variance (Breiman *et al.* 1984). No matter how many multiple trials are performed, the results are bounded by the size of the full sample and its variance from the true population. For these applications, NP demonstrates a significantly better result only for the sample size of 200 letter recognition application (with its 26

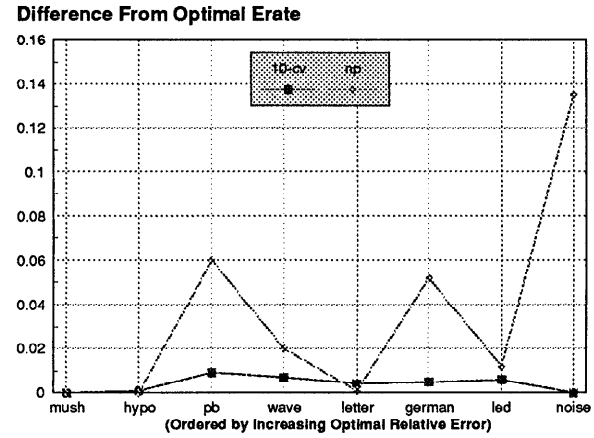

Figure 4: Tree Selection Performance for Size 1000 Samples

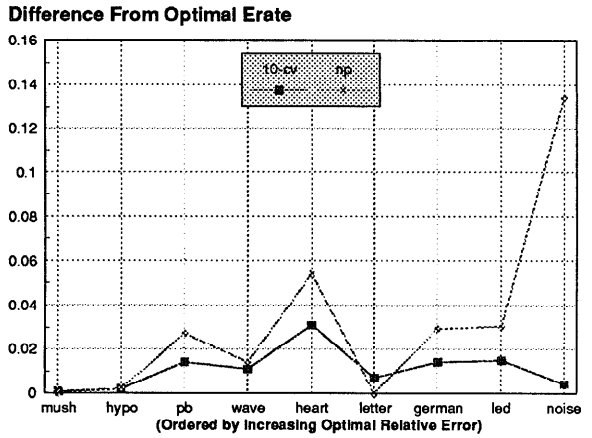

Figure 5: Tree Selection Performance for Size 200 Samples

classes and small samples for each class).

With any significance test, two statistical problems remain:

- Significance testing does not directly measure the magnitude of the difference.
- Even with a comparative result below two standard errors, there may still be a competitive edge. The difference in performance of competing solutions is usually more accurately determined than the individual estimates (Breiman *et al.* 1984; Shibata 1981).

These factors should lead one to consider the overall pattern of performance, and the relative advantages of competing solutions on large numbers of independent test cases. Figures 4 and 5 illustrate this overall pattern.

Discussion

The results listed in Tables 3 and 4, which are plotted in Figures 1 and 2, strongly suggest that pruning by 10-cv is nearly unbiased. Figure 3 shows that 10-cv pruning is consistent: as the sample size increases, the results get better and the difference from the optimal answer decreases. Not pruning is clearly a highly biased (optimistic) strategy.

When the bias of NP is close to the true answer, such as in the letter application, NP performs well, even better than the nearly unbiased 10-cv strategy. For size 200, the sampling variance is still moderate so that NP is sometimes competitive with 10-cv. By size 1000, the case for 10-cv is overwhelming, and we see 10-cv approaching an optimal selection strategy. Even for size 200 samples, 10-cv is competitive across the board, with typically slight losses to NP. But an NP strategy, with hugely optimistic predictions, can lead to disaster for noisy applications. Unfortunately, many real-world applications turn out to be collections of noisy features.

The fundamental unifying theme in an analysis of tree pruning performance must be the binomial model with the variance of Equation 1. This model demonstrates the difficulties in smaller sample estimation and the increasingly better performance for larger samples. It explains the sometimes weak behavior of unbiased tree selection for smaller samples. It also explains the near optimal results for larger samples due to the reduced variance.

Considering the variety of datasets used in this study, including many found in previous studies, one can reasonably conclude that these data are representative of typical real-world applications. By computing average tree sizes and comparing results to ideal trees, we have provided an objective basis to compare bias and accuracy of selecting the right-sized tree. Most importantly, the results of this study are consistent with an underlying theory of tree pruning using cross-validation. Pruning is mapped into a a form of binomial testing (coin tossing) to determine a proportion (the error rate). Direct testing on independent test cases is known to be unbiased with the standard binomial variance for sample estimators. The accuracy of independent testing is mostly dependent on test sample size and independent of solution complexity. This study shows that cross-validation estimators are good approximators to estimates based on independent test cases.

Overall, these results demonstrate that NP is usually inferior to 10-cv, sometimes by very large margins, for samples of at least 200 cases. If one were aware of the characteristics of the true answer, such as likely solution complexity, one might achieve slightly better results by biasing the solution in that direction. For a size 1000 sample, such knowledge would be of marginal value. The results are entirely consistent with sample size variation. With sample size of at least 200, good results for tree pruning and selection should generally be achievable without any knowledge of the population.

Data	n	Ideal		10-cv		NP	
		Err	Size	Err	Size	Err	Size
Mush	4800	.000	13.8	.000	13.8	.000	13.8
Hypo	2000	.003	7.6	.004	7.7	.004	11.9
Hyper	3772	.011	6.0	.011	7.2	.014	29.0
Letter	10000	.155	1332.2	.156	1311.7	.156	1463.8
Wave	5000	.215	101.9	.219	78.2	.240	542.0
German	6479	.197	74.0	.200	83.0	.266	1146.0
Noise	2500	.251	1.4	.252	1.1	.389	491.1
LED	10000	.488	49.7	.489	49.6	.490	77.7

Table 4: Results for very large training samples

The same binomial model should also be used to compare significance of results (Breiman *et al.* 1984). Standard significance tests, such as t-tests or nonparametric ranked sign tests on the results of each trial or the pooled data of all cases and trials, will overweight the significance of results for increasing numbers of non-independent trials.

One might wonder whether the experimental results suggest that the standard tree induction estimation techniques should be modified. Unlike our single-minded search for minimum error pruning, in the real world there is a strong tendency to simplify results. One sometimes chooses a simpler tree that is close to the best solution (Breiman *et al.* 1984). The usual rationale is in terms of explanatory capabilities. However, the real world may not be the perfect laboratory setting that was presented for the experiments of this paper. While the ideal model is a random sample from an infinite population, future samples may actually be drawn from a slightly changing population, where the simpler solution actually performs better. Whichever variation is used, we believe that these experimental results strongly confirm the efficacy of resampling estimators for tree pruning and selection. Although we have not examined the effects of pruning on other learning models, the known generality of resampling techniques should produce similar results.

References

Apté, C.; Damerau, F.; and Weiss, S. 1994. Automated Learning of Decison Rules for Text Categorization. Technical Report RC 18879, IBM T.J. Watson Research Center. To appear in ACM Transactions on Office Information Systems.

Breiman, L.; Friedman, J.; Olshen, R.; and Stone, C. 1984. *Classification and Regression Tress.* Monterrey, Ca.: Wadsworth.

Cestnik, B., and Bratko, I. 1991. On estimating probabilities in tree pruning. In *Machine Learning, EWSL-91.* Berlin: Springer Verlag.

Cohen, W. 1993. Efficient pruning methods for separate-and-conquer rule learning systems. In *Proceedings of IJCAI-93*, 988–994.

Crawford, S. 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. *Journal of the American Statistical Association* 78:316–333.

Hassibi, B., and Stork, D. 1993. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5.* San Mateo, CA: Morgan Kaufmann. 164–171.

Mitchell, T. 1990. The need for biases in learning generalizations. In *Readings in Machine Learning.* San Mateo, CA: Morgan Kaufmann. 184–191.

Quinlan, J. 1987. Simplifying decision trees. *International Journal of Man-Machine Studies* 27:221–234.

Quinlan, J. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Schaffer, C. 1992a. Deconstructing the digit recognition problem. In *Proceedings of the Ninth International Conference on Machine Learning*, 394–399. San Mateo, CA: Morgan Kaufmann.

Schaffer, C. 1992b. Sparse data and the effect of overfitting avoidance in decision tree induction. In *Proceedings of AAAI-92*, 147–152. Cambridge, MA: MIT Press.

Schaffer, C. 1993. Overfitting avoidance as bias. *Machine Learning* 10:153–178.

Shibata, R. 1981. An optimal selection of regression variables. *Biometrika* 68:45–54.

Utgoff, P. 1986. Shift of bias for inductive concept learning. In *Machine Learning: An Artificial Intelligence Approach. Volume 2.* San Mateo, CA: Morgan Kaufmann. 107–148.

Watrous, R. 1991. Current status of peterson-barney vowel formant data. *Journal of the Acoustical Society of America* 89(3).

Weiss, S., and Indurkhya, N. 1993. Optimized Rule Induction. *IEEE EXPERT* 8(6):61–69.

Weiss, S., and Indurkhya, N. 1994. Small sample decision tree pruning. In *Proceedings of the Eleventh International Conference on Machine Learning.*

Weiss, S. 1991. Small sample error rate estimation for k-nearest neighbor classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(3):285–289.

Wolpert, D. 1992. On overfitting avoidance as bias. Technical Report SFI TR 92-03-5001, The Sante Fe Institute.