# Decision Tree Pruning: Biased or Optimal?

Sholom M. Weiss

Department of Computer Science, Rutgers University
New Brunswick, New Jersey 08903, USA

Nitin Indurkhya

Department of Computer Science, University of Sydney
Sydney, NSW 2006, AUSTRALIA

This paper primarily focuses on Decision Tree Pruning by using different Estimators.This paper is re-evaluating the efficacy of Decision tree pruning.The main reason for pruning is to avoid overfitting of data but one must also be aware of the fact that too much pruning may lead to underfitting as well.Hence our objective is to find a tree with appropriate size. Therefore, pruning task is further mapped to a tree selection problem among various subtrees ranging in size from unpruned tree to a null tree. We must also keep this in mind that pruning may sometimes degrade performance. For this paper we will limit our discussion to general case with samples of moderate size to large size, lowest sample-size being 200.
While selecting the best tree, we only look for Error-estimation.The tree with lowest error-estimate is selected. And accuracy of these error-estimates depends on type of procedure we used.This paper for tree pruning compares the performance of 3 ways of estimating error rates. They are:
1. Ideal- Approximated accurately by testing each Tree $T_i$,on a very large and independent test set. This helps us to make optimal tree selection.
2. NP- This procedure is based on the apparent error rate for training cases. Since the apparent error is minimum for Covering tree, this strategy becomes not pruning for the initial Covering tree.
3. Cross-Validation: When there are less independent test cases we use resampling methods and Cross-Validation is the obvious choice and the 10-fold cross-validation(90 % training and 10 % test cases with 10 mutually exclusive test partitions) is most widely used method.
Revisiting some basic principles like, For a given sample size, Standard error tells us the average value by which our error rate diverge from the truth. It is given by $SE = \sqrt{Variance}$, where Variance is given by,

$$Variance = \frac{p(1-p)}{n} \tag{1}$$

Here n is the size of test set and p is true error rate.From equation 1 we see that p has a little effect and n plays the major role here.So when n is large ,standard error is very small.
An estimator is called unbiased when its expected value is equal to true value of the metric. Now, the true value of metric is estimated by,

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

where x is the estimator, $X_i$ is its mean value for the i-th sample and N is the number of samples.An unbiased estimator averages out to true answer for a large set of independent samples but may vary from sample to sample. We also want our estimator to be consistent which implies that as the sample size increases, the results get better and the difference from the optimal answer decreases. If estimators are used for tree selection, the tree selection bias should be measured.Bias is measured by average of selected trees.Also one must keep in mind that unbiased procedure is not necessarily optimal. When bias fits characteristics of population, then biased strategy works better and is closer to the truth.
When focusing on sources of error, one cannot overlook the fact that in cross-validation we are training on 90% data not on full data.Another source can be the sample size because with smaller sample sizes we have some errors with matching of tree complexities.
In order to ease out comparisons, same datasets reported in (Schaffer 1992b;1992a;1993)[1] were used . In addition to these we also used 4 other datasets namely:
1. Random noise for two classes with prevalence of approximately 75% for one class
2. A two class problem representing word counts in German Reuters new stories[2]
3. The Peterson/Barney Vowel Formant Dataset[3]
4. The Waveform data [4]
With these variety of dataset we were able to examine a wide range of true answers. All experiments used the CART tree induction program and

| Dataset | n | Ideal Err | Ideal Size | 10-cv Err | 10-cv Size | NP Err | NP 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 |

| Data | n | Ideal Err | Ideal Size | 10-cv Err | 10-cv Size | NP Err | NP 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 |

Figure 1:
(a)Table 1: Comparison of Ideal, 10-CV and NP
(b)Table 2: Comparison of Ideal, 10-CV and NP with very large samples

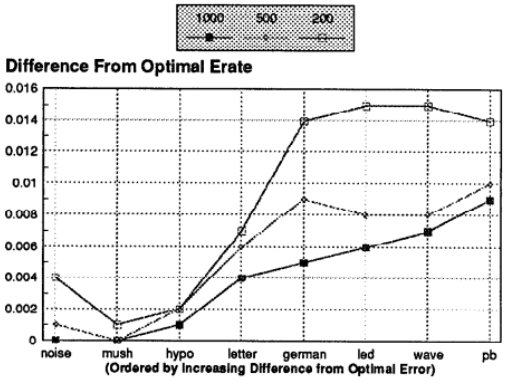


Figure 2: Consistency of 10-CV Performance for Varying Size Samples

10 fold cross-validation was used to select minimum error tree.Although some modification were made that, Each trial was initiated with a new random seed and Ties were broken in favor of the larger tree.New random seeds were used in order to maximise the randomness.Differentiating from original analysis by (Schaffer 1992b)[1], we determined the average error rates and hypothetically optimal tree-selection strategy. The results corresponding to 3 estimators we use for different sample sizes for different datasets are given in figure1 and figure2. Table 1 shows results for samples of moderate to large size, And table 2 are results for a very large sample size. From results we can say that pruning by 10-CV is nearly unbiased. NP strategy is totally biased strategy that means it only performs better when the sample size is small and bias is close to true answer.Also, NP leads to very unexpected bad results for noisy data and most of data always contains noise. Equation 1 explains well why unbiased model behaves near optimal for large sample and shows weak behaviour for small samples.Figure 2 also shows that 10-CV pruning is fairly consistent. In total, from our experiments we can say that 10-CV performs way better than NP for samples of at least 200 cases.

[1] 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.
[2] Apte,Damerau and Weiss, S. 1994. Automated Learning of Decison Rules for Text Categorization. Technical Report RC 18879, IBM T.J. Watson Research Center
[3] Watrous, R. 1991. Current status of Peterson-barney vowel formant data. Journal of the Acoustical Society of America 89(3)
[4] Breiman, L.; Friedman, J.; Olshen, R.; and Stone, C. 1984. Classification and Regression fiess. Monterrey, Ca: Wadsworth.