# Review: Decision Tree Pruning: Biased or Optimal?

Sholom M. Weiss
Department of Computer Science,
Rutgers University
New Jersey 08903, USA

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

## 1 Introduction

The main objective of the paper is pruning the tree and selecting the ‘right sized’ tree from a series of sub-trees.So, the pruning of trees is mapped into a problem of tree selection. It is not that pruning always leads to improved results it may sometimes even degrade the performance. Pruning is done in order to avoid overfitting. Also one must keep in mind that too much pruning can cause underfitting, hence the selection of right sized trees plays an important role in pruning.

## 2 Summary

The task of pruning was done by selecting the tree with the lowest error estimate which is calculated by error estimation procedures. We are using 3 error estimators and comparing their performance.Error-Estimators used are - Ideal,NP and Cross-Validation(CV). The cross validation used here is most widely used 10 fold cross validation using 90 % of training and 10% testing cases with mutually exclusive test partitions. The authors for this paper used various datasets to be specific-10, representative of the typical real world problems.CART tree induction program was used for all experiments with slight modifications. We observed that for a decent large sample unbiased tree selection is near to the optimal solution but if the bias fits the characteristic of the population, then biased strategy is closer to truth. As mentioned in the paper,since we are using only 90% training, the error rate estimates are for 90% not for 100% which is a small weakness when the true answer is near the unpruned tree.

## 3 Results

From the results in paper, we observed that pruning by 10-CV is nearly unbiased and consistent which means results get better with increasing sample size. NP performs better only when its bias is close to the true answer that too is limited for small samples.NP may lead to problems due to its optimistic predictions as real world data is full with noisy features.10-CV performs well enough on all different datasets in all sample sizes. So, in total 10-CV performs much better than NP for samples of at least 200 cases.

# 4 Evaluation

## 4.1 PROS

- This paper addresses the major issues like bias of tree pruning by CV, effect of sample size, divergence from optimal solution and extent to which knowledge about overall population is needed for accurate results.

- Discussion in this paper is based on the most general case i.e moderate to large size samples, minimum being sample with 200 cases, so this makes the study a well generalised study for a real world analysis.

- This paper uses an optimal tree to measure the divergence from the optimal solution.This helps us to compare our results from different estimators to the hypothetically optimal tree size.By computing average tree sizes and comparing results to ideal/optimal trees, it has provided an objective basis to compare bias and accuracy of selecting the right-sized tree.

- This paper uses 10 varying datasets which helps us to examine a wide spectrum of true answers and these variety of datasets are representative of typical real-world applications.For example - Like by adding the noise dataset we came to know the how badly NP performs for noisy data.Hence these many datasets helped us to have a more generalised results.

- We used CART tree induction but we made some slight modifications that Each trial was initiated with a new random seed which helped to maximise randomness. Ties were broken in favour of larger trees,the 90% tree is actually being estimated, and therefore the larger tree is somewhat more likely for the full sample.

## 4.2 Cons

- Resampling is quite complex for decision trees when less independent test cases are available because in addition to generating multiple trees , these trees have to be pruned so that complexities of subtrees are matched for each subsample.

- When using CV, when fewer samples we ought to look for resampling. For 10-fold cross validation we use 90% of training, so the error rate estimates are for 90% not for 100% which is a small weakness when the true answer is near the unpruned tree.

- For smaller samples ,10 fold CV involves some interpolation for matching of trees.And also this paper is somewhat ignoring the small sized samples i.e less than 200 as it causes problems. Hence the results we get here are not for smaller size samples and are limited for moderate to larger sizes.

- In this paper we only focused on minimum error pruning which may generate a larger tree for a little improvement in accuracy.But in real world majority of the times, there is strong tendency to simplify results and for that we need a simple tree because larger trees may have some overfitting and may be computationally extensive.

### 4.3 Suggestions for Improvement

Instead of having a 90-10 split in CV we can instead have a 95-5 train-test split which will result in better training and will help us to reduce the error and will give us more accurate error estimates. Also, here we focused only on the minimum error tree, we could have also taken into consideration the simpler trees and then compared them. It would have given us a better idea of which estimation procedure works best. In this paper ties were broken in favor of larger trees instead we could have chosen the simpler tree instead which may have been a good choice because minimum error and simpler tree is what we want.

## 5 Conclusion

In conclusion, this paper was very well written and presented. This paper reconsidered some previous papers and produced more generalised results and also performed additional experiments. It addressed several major issues like 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. The modifications added to the CART method used also helped us to get a better understanding. The use of multiple datasets helped to gain better insight for real world data analysis.