Adjusting for Multiple Testing in Decision Tree Pruning

David Jensen
Experimental Knowledge Systems Laboratory
Department of Computer Science
Box 34610 LGRC
University of Massachusetts
Amherst, MA 01003-4610
jensen@cs.umass.edu

July 1, 1996

1 Introduction

Overfitting is a widely observed pathology of induction algorithms. Overfitted models contain unnecessary structure that reflects nothing more than chance variations in the particular data sample used to construct the model. Portions of these models are literally wrong, and can mislead users. Overfitted models require more storage space and take longer to execute than their correctly-sized counterparts. Finally, overfitting has been shown to reduce the accuracy of induced models on new data [13, 6].

For induction algorithms that build decision trees [1, 12, 14], *pruning* is a common approach to correct overfitting. Pruning techniques take an induced tree, examine individual subtrees, and remove those subtrees deemed to be unnecessary. While pruning techniques can differ in several ways, their primary differences concern the criteria used to judge sub-trees. Many criteria have been proposed, including statistical significance tests [8], pessimistic error estimates [14], and minimum description length calculations [11].

Most common pruning techniques, however, do not account for one potentially important factor — multiple testing. Multiple testing occurs whenever an induction algorithm examines several candidate models and selects the one that best accords with the data. Any search process necessarily involves multiple testing, and most common induction algorithms involve implicit or explicit search through a space of candidate models. In the case of decision trees, search involves examining many possible subtrees and selecting the best one. Pruning techniques need to account for the number of subtrees examined, because such multiple testing affects the apparent accuracy of models on training data [7].

This paper examines the importance of adjusting for multiple testing. Specifically, it examines the effectiveness of one particular pruning method — *bonferroni pruning*. Bonferroni pruning adjusts the results of a standard significance test to account for the number of subtrees examined at a particular node of a decision tree. Evidence that bonferroni pruning leads to better models supports the hypothesis that multiple testing is an important cause of overfitting.

The next section briefly surveys several relevant approaches to decision tree pruning. Section 3 presents the results of an experiment comparing bonferroni pruning to other pruning techniques. The final section discusses the implications of the experiment and the content of the full paper.

2 Existing Methods of Pruning Decision Trees

One approach to pruning uses a statistical significance test to judge whether a particular subtree should be retained in the final model. Both Quinlan [12] and Kass [8] have devised pruning strategies based on significance tests. Clark and Niblett's CN2 rule induction algorithm [2] also uses statistical significance tests.

Unfortunately, conventional statistical significance tests are not an entirely satisfactory solution to overfitting. Empirical testing has shown that extremely high significance levels must be used in order to prevent overfitting, and even these do not prevent overfitting in some cases. For example, Clark and Niblett [2] note that relatively serious overfitting still occurs even when significance tests are used.

For these reasons, Quinlan rejects conventional statistical tests in favor of pessimistic error estimation, the technique still used as the default in C4.5. The initial version of pessimistic error estimation [13] was superseded by a more stringent approach [14] that estimates the error rate of a subtree based on the binomial distribution.

Most recently, approaches to pruning based on the Minimum Description Length (MDL) principle have been devised [11]. MDL characterizes both datasets and models by the number of bits needed to encode them. The best tree is the one with the smallest total "description length" for the data, that is, the smallest sum of model description and description of the exceptions to the model's predictions.

However, none of these approaches are intended to account for multiple testing. An occasional induction algorithm has taken account of its effects [6, 5], but pruning techniques have generally not. Several researchers have called attention to the effect of multiple testing on *empirical comparisons of learning algorithms* [3, 4, 15], but the algorithms themselves have only rarely been so analyzed.

3 Experiment

This experiment applies a decision tree algorithm to an artificially-created data set. The training set was created by randomly generating instances with 30 binary attributes. Next, a binary classification variable was created by applying a specified, tree-structured function to five of the attributes. For this experiment, the theoretically correct tree (shown in figure 1) has eleven nodes — five decision nodes and six leaf nodes. Finally, the values of this classification variable were systematically corrupted by complementing each of the variable's values with a probability of 0.1. Thus, on average, 10% of the values of the variable were incorrect, producing a theoretical upper limit of 90% on classification accuracy.

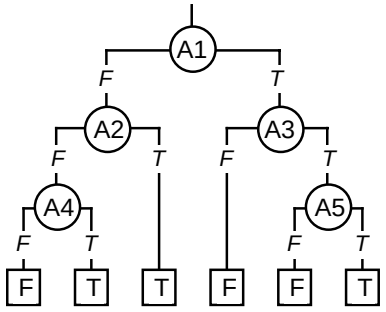

Figure 1: Tree Representing Correct Relationship

The experiment employs an ID3-like recursive splitting algorithm that constructs decision trees. The algorithm uses the information gain function to choose the best attribute for a particular decision node. A node is made into a leaf node when all examples are of a single class, no potential split improves accuracy, or no unused attributes remain. The class label of the leaf node is determined by the majority class of the training examples present at that node.

Pruned trees are created by applying several bottom-up pruning techniques to the original tree. The pruning techniques differ only by their decision criteria. The criteria is applied to all non-leaf subtrees of the original tree. If the subtree is judged to be valid, the node is retained. If not, the subtree is converted to a leaf node and labeled with the majority class of the training examples present at that node. For any given tree, pruning continues until all subtrees are judged to be valid.

The pruning criteria used in the experiments are:

- FISHERS Significance testing using Fisher's exact test [9] with $\alpha = 0.10$.
- PESSIMISTIC The technique used in C4.5 [14].
- MDL Minimum length encoding, using Utgoff's MDL formulation [16].
- BONFERRONI Fisher's exact test with α adjusted to account for the number of tests using the Bonferroni adjustment [9]. The significance level for each individual test is $\alpha_1 = 1-(1-\alpha_k)^{1/k}$, where α_k is the overall significance level desired and k is the number of tests.

Finally, these methods are compared to trees which have not been pruned at all (UNPRUNED).

The size of training sets was varied from 5 to 200 by increments of 5 instances. In each trial, a training set of appropriate size was generated, an unpruned tree was induced, that tree was provided to each pruning technique, and the accuracy of the resulting pruned trees was evaluated on a test set of 1000 freshly generated examples. In addition, the complexity of each pruned tree was determined by counting the total number of nodes. For each sample size, 100 trials were conducted and the results (accuracy and complexity) were averaged.

The results are shown in Figure 2. BONFERRONI produces the best trees if the number of instances in the training set exceeds 125. Prior to that threshold, there is a period where BONFERRONI creates trees whose accuracy is virtually indistinguishable, on average, from those of other pruning techniques ($75 < N < 125$), as well as a period where it produces trees that are less accurate ($N < 75$). Only BONFERRONI produces trees that converge to the correct tree size as the number of instances in the training set increases. The other techniques add unwarranted complexity.

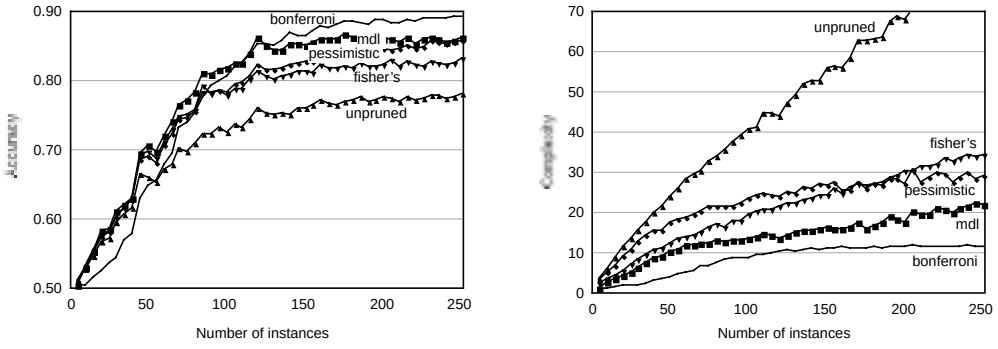

Figure 2: Accuracy and Complexity of Pruned Trees

4 Implications

The relative performance of the techniques tested in Section 3 suggests that only those techniques that adjust for multiple testing (e.g., BONFERRONI) avoid overfitting. Other techniques such as PESSIMISTIC and FISHERS lower the complexity of induced models somewhat, but they do not converge on the theoretically correct model. Instead, they continue to add complexity as the number of instances in the training set increases. This behavior has been found in other induction algorithms, including C4.5 [10].

The full paper will present a broader range of experiments using relationships of varying complexity. In addition, it will discuss the tradeoff between statistical significance and statistical power and its implications for induction algorithms. Finally, it will present a more detailed discussion of why multiple testing affects overfitting.

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees.* Pacific Grove, CA: Wadsworth and Brooks.

[2] Clark, P. and T. Niblett (1989). The CN2 induction algorithm. *Machine Learning* 3(4):261–283.

[3] Feelders, A. and W. Verkooijen (1995). Which method learns the most from data? Methodological issues in the analysis of comparative studies. *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics,* January 4-7, 1995, Ft. Lauderdale, Florida. 219–225.

[4] Gascuel, O. and G. Caraux (1992). Statistical significance in inductive learning. *ECAI92: Proceedings of the 10th European Conference on Artificial Intelligence.* B. Neumann (Ed.), Chichester: Wiley. 435–439.

[5] Gaines, B. (1989). An ounce of knowledge is worth a ton of data: Quantitative studies of the trade-off between expertise and data based on statistically well-founded empirical induction. *Proceedings of the Sixth International Workshop on Machine Learning.* 156–159.

[6] Jensen, D. (1992). Induction with Randomization Testing: Decision-Oriented Analysis of Large Data Sets, Doctoral Dissertation, Sever Institute of Technology, Washington University, St. Louis Missouri.

[7] Jensen, D. and P.R. Cohen (1996). Overfitting in Induction Algorithms: A Synthesis and New Theory. In preparation.

[8] Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29(2):199–127.

[9] Kotz, S. and N.L. Johnson (Eds.) (1982-1989). *Encyclopedia of Statistical Sciences.* New York: Wiley.

[10] Oates, T. and D. Jensen (1996). The effects of training set size on decision tree complexity. Submitted to the Sixth International Workshop on Artificial Intelligence and Statistics.

[11] Quinlan, J.R. and R. Rivest 1989. Inferring decision trees using the minimum description length principle. *Information and Computation* 80:227–248.

[12] Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning* 1(1):81–106.

[13] Quinlan, J.R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies* 27:221–234.

[14] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* San Mateo: CA: Morgan Kaufmann Publishers.

[15] Salzberg, S. (1995). On Comparing Classifiers: A Critique of Current Research and Methods. Technical Report JHU-95/06, Department of Computer Science, Johns Hopkins University, May 1995.

[16] Utgoff, P. (1995). Decision tree induction based on efficient tree restructuring. Technical Report 95-18. Department of Computer Science, University of Massachusetts, Amherst. March 17, 1995.