An Empirical Analysis of the Benefit of Decision Tree Size Biases as a Function of Concept Distribution

TECHNICAL REPORT 95-29

Patrick M. Murphy*

August 1995

*Correspondence should be directed via email to pmurphy@ics.uci.edu, or via post to UC-Irvine, Dept. of ICS, Irvine, CA 92717, USA.

An Empirical Analysis of the Benefit of Decision Tree Size Biases as a Function of Concept Distribution

Patrick M. Murphy

Department of Information & Computer Science
University of California, Irvine, CA 92717
pmurphy@ics.uci.edu
http://www.ics.uci.edu/~pmurphy
(714) 824-4035

Abstract

The results reported here empirically show the benefit of decision tree size biases as a function of concept distribution. First, it is shown how concept distribution complexity (the number of internal nodes in the smallest decision tree consistent with the example space) affects the benefit of minimum size and maximum size decision tree biases. Second, a policy is described that defines what a learner should do given knowledge of the complexity of the distribution of concepts. Third, explanations for why the distribution of concepts seen in practice is amenable to the minimum size decision tree bias are given and evaluated empirically.

Keywords: Induction, Decision Trees, Bias

1 Introduction

Top down induction of decision trees has been significantly studied by a number of researchers, e.g. (Breiman, Friedman, Olshen, & Stone, 1984) and (Quinlan, 1986). The majority of the algorithms that construct decision trees from examples use splitting heuristics that aim to minimize the size of the induced decision trees. Empirical evidence (Buntine, 1992) has suggested that, for real world problems, the bias for small trees tends to be useful for increasing predictive accuracy.

Unfortunately, there has been little explanation for why the distribution of concepts, seen in practice, is amenable to the bias of choosing the minimum sized decision tree. Why not choose from the largest trees? Why does size matter?

In the following, the conditions under which the minimum size and maximum size biases are beneficial will be shown. Also, explanations will be given for why the minimum size bias appears to be beneficial in practice.

The approach used in this research differs markedly from the approach used in previous research by the author, (Murphy & Pazzani, 1994). In the earlier research, a more detailed analysis of the effect of size biases on a small number of concepts was presented. In this research, the analysis is over distributions of concepts.

In Section 2, the experimental methodology used to form the results of this paper will be described; in Section 3, it will be shown under which concept distributions the decision tree size biases are beneficial and harmful; and in Section 4, explanations and supporting results will be presented that show why the minimum size bias appears to be beneficial in practice.

2 Experimental Methodology

For each experiment, a random sample of a specific distribution of concepts[1] was generated and then evaluated (relative to a specific training set size). For all experiments, each concept had five boolean features and was formed over an example space of either 16 or 32 unique examples.

Given a specific distribution to draw concepts from and a training set size, concepts were generated and evaluated in the following manner.

1. Draw a concept from the specific distribution.
2. Determine the concept's complexity (number of internal nodes in the smallest decision tree consistent with the example space).
3. Generate a single random train/test partition of the example space as specified by the training set size.
4. Generate all decision trees[2] consistent with the training set.
5. Record the size (number of internal nodes) and the number of errors on the test set for each consistent tree.
6. Partition the set of consistent trees by size.
7. Calculate the mean number of errors of all trees in each partition.

3 Bias Benefit & Concept Distribution

The experiments in this section will demonstrate under which concept distributions decision tree size biases are beneficial and harmful.

[1] Each sample contained at least 20,000 concepts.

[2] The generation of all consistent decision trees was done as described in (Murphy & Pazzani, 1994).

3.1 Bias Benefit Evaluation

The benefit of decision tree minimum and maximum size biases were evaluated relative to the approach of selecting the decision tree size randomly[3]. For each sample of concepts, two measures of benefit were calculated.

The first measure, "percent benefit", is the percentage of concepts for which the mean error of trees with the minimum (maximum) size was lower than the mean error of trees of a randomly selected size. Ties counted one half towards each approach. When using percent benefit, a value near 50% means that there is no benefit in using the size bias over the random approach. When the value is above (below) 50%, the size bias is beneficial (harmful) relative to the random approach.

The second measure, "accuracy benefit", is the difference in mean accuracy between using the minimum (maximum) size bias and the random approach. When using accuracy benefit, a value near zero means that there is no benefit in using the size bias over using the random approach. When the value is above (below) zero, the size bias is beneficial (harmful) relative to the random approach.

3.2 Concept Distribution Complexity

This section will show some conditions under which the minimum (maximum) size bias is of benefit to a learner. For this result, four samples of concepts, all drawn from the uniform distribution and each with a different training set size, were generated and evaluated. Uniform distribution concepts were generated by randomly assigning a true or a false value to the class attribute for each example in the example space of 32 examples. Each sample was then partitioned by concept complexity and the two benefit measures were calculated for each partition.

[3] It is assumed that once a learner has chosen a size from which to select a decision tree, the selection of that decision tree, from among all consistent decision trees of that size, is done randomly.

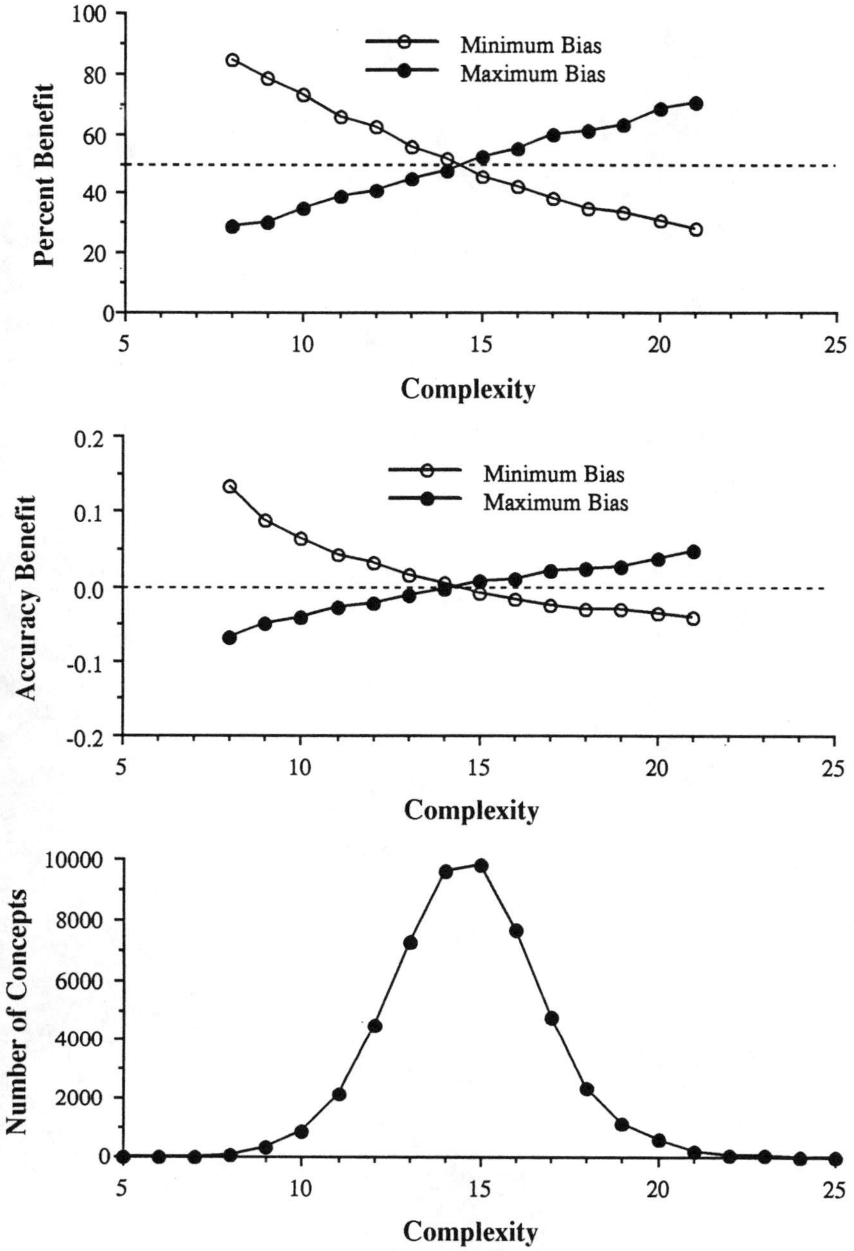

Figure 1: (Top) Percent benefit of the minimum and maximum size decision tree biases as a function of concept complexity. (Middle) Accuracy benefit of the minimum and maximum size decision tree biases as a function of concept complexity. (Bottom) Number of concepts generated as a function of concept complexity, when concepts are drawn for the uniform distribution.

Figure 1 (top) plots the percent benefit of the two size biases as a function of concept complexity[4]. Figure 1 (middle) plots the accuracy benefit of the two size biases as a function of concept complexity. For each measure of benefit, when concept complexity is small, the minimum size bias is most beneficial, and when concept complexity is large, the minimum size bias is most harmful. The opposite is true for the maximum size bias. Between complexities 14 and 15, neither bias has any benefit[5]. Note, the results presented in Figure 1 are based on the sample formed with a training set size of 20. The results were similar for the other training set sizes.

Figure 1 (bottom) plots the number of concepts used to form the benefit measures at the various complexities. This graph shows how rare very simple and very complex concepts are when generated uniformly.

3.3 Generalization Performance

At first glance, the results presented in Figure 1 might imply a policy that a learner could take, given knowledge of the complexities of the distribution of concept. For example,

- When the complexity of the distribution is small use the minimum size decision tree bias.
- When the complexity of the distribution is large, use the maximum size decision tree bias.

However, the results presented in Figure 1 are somewhat misleading.

In (Schaffer, 1994), the benefit of the bias used by a learner (generalization performance) is defined as the difference between the mean accuracy of the learner and the mean accuracy of a random guesser (0.5 for two class problems). A value near zero means that the learner does no different than

[4] Even though a few concepts with complexities below 8 and above 21 were randomly generated, the benefit measures for those complexities were not included. There were too few concepts to form accurate benefit measurements for those complexities.

[5] Empirically, 14.6 was the mean complexity for each of the four samples.

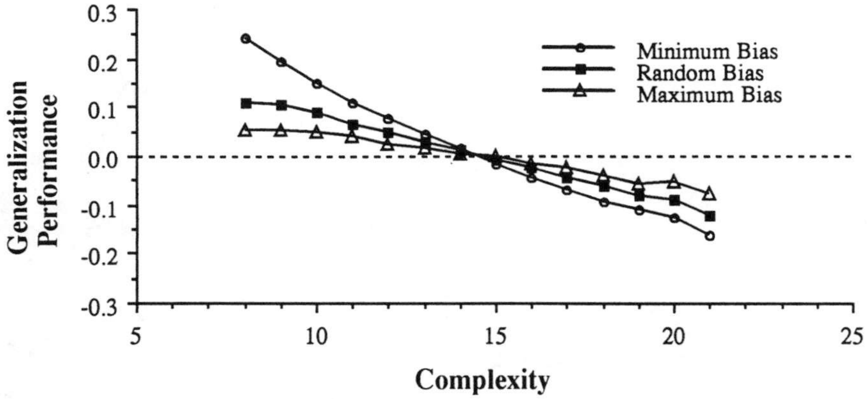

Figure 2: Generalization performance for the minimum, random and maximum size biases as function of concept complexity.

chance, and a value above (below) zero means that the learner does better (worse) than chance.

The reason the results presented in Figure 1 are misleading is because the approach that the minimum and maximum size biases were compared against, the approach of choosing the size of the decision tree randomly, is not a random guesser. This random approach is better described as a bias for "medium sized" trees[6].

Figure 2 presents generalization performance for the three biases as a function of concept distribution complexity. When the complexity of the distribution is low, the minimum size bias is more accurate than the random (or medium size) bias, and the random bias is more accurate than the maximum size bias. The reverse is true when the complexity of the distribution is large.

However, only when the complexity of the distribution is low, do any of the biases have positive generalization performance. When the complexity of the distribution is high, the generalization performance of the biases is negative. Therefore, a better strategy that a learner could take, when the complexity of the distribution of concepts is high, is to guess randomly.

[6] Actually, the random bias is a bias against extreme tree sizes.

4 Concept Distributions Benefited by the Minimum Size Bias

The results of Section 3 showed that when concepts are simple, the use of the minimum size decision tree bias is beneficial. Because this bias is successful in practice, it may be that the distribution of concepts that practical problems are being drawn from is, in some manner, simple.

One explanation for simple concept distributions may be that we are expecting our learners to remove irrelevant features. For example, when we as the knowledge engineers decide on the set of features that our learners use, we tend to err in favor of retaining irrelevant attributes. "Let the learner select the relevant features," we say to ourselves. Another explanation may be, that we, as a legacy from statistics, are using features that are highly correlated with the class attribute. We expect our learners to combine the features into an accurate model. The following two sections will show that the presence of irrelevant features or class correlated features do lead to simple concept distributions, under which the minimum size decision tree bias is beneficial. Note, percent benefit is used as the benefit measure in this section.

4.1 Irrelevant Features

The results in this section will show that when concepts, drawn from a uniform distribution, are given irrelevant attributes, the minimum size decision tree bias becomes beneficial. Under the uniform distribution, concepts with irrelevant features are very unlikely, only 1 in $2^{2^{d-1}}$ for d feature concepts (1 in 2^{16} for concepts with 5 features).

In Section 3, concepts were generated by randomly associating each example in the five boolean feature example space to either the true class or the false class. For this result, uniformly generated concepts were formed over a subset of the features with the remaining features added as irrelevant fea-

Training Set Size	*Number of Irrelevant Features* 0	1	2
5	50.4	52.8	56.9
10	49.8	64.5	77.3
15	50.3	77.1	89.5
20	49.7	86.0	93.9

Table 1: Benefit of the minimum size decision tree bias as a function of both the training set size and the number of irrelevant features used to create each concept distribution.

Number of Irrelevant Features	*Mean Concept Complexity*
0	14.6
1	7.42
2	3.66

Table 2: Mean concept distribution complexity as a function of the number of irrelevant features used to create each concept distribution.

tures. Note, all concept distribution samples generated for this experiment have five boolean features and an example space of 32 unique examples.

Table 1 presents the benefit of the minimum size decision tree bias as a function of the both the training set size and the number of irrelevant features used to create each concept distribution. When there are no irrelevant features, the minimum size bias has no benefit (all values are near 50%) because the concepts are drawn from a uniform distribution. As the number of irrelevant features increases, the benefit of the minimum size decision tree bias increases. In addition, from Table 2 (which presents mean concept distribution complexity as a function of the number of irrelevant features), as concept distribution complexity decreases, the benefit of the minimum size decision tree bias increases. This later result is consistent with the results shown in Figure 1.

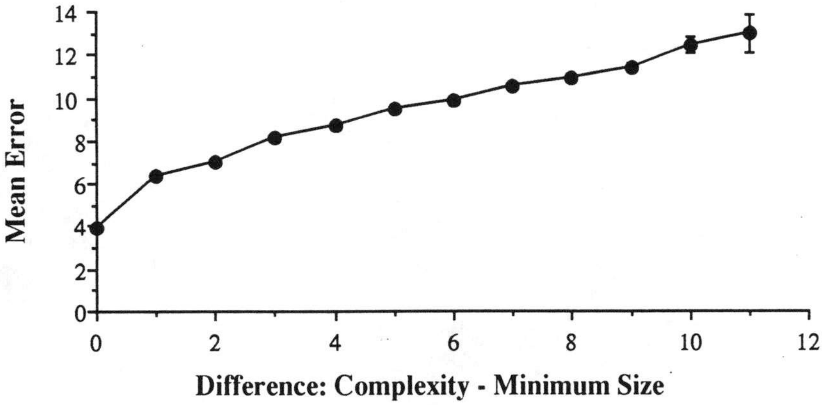

Figure 3: Mean number of errors for the smallest trees as a function of the difference between the complexity and the size of the smallest trees.

Also, when there are irrelevant features, as the number of training examples increases, the benefit of the minimum size decision tree bias increases. This can be understood by recognizing that small training sets typically generate small minimum size decision trees, and the farther the minimum size decision tree is from the complexity of the concept, the less accurate it tends to be. Figure 3 provides empirical evidence for this claim.

Figure 3 was created by partitioning concepts, from a particular distribution and training set size, by the difference between the complexity of the concept and the size of the smallest trees consistent with the training examples. After partitioning, a mean of the mean number of errors of the smallest tree sizes for the concepts in each group were calculated. Figure 3 plots the means (and 95% confidence intervals) as a function of the differences[7].

4.2 Class Correlated Features

The results in this section will show that the minimum size decision tree bias becomes beneficial when concepts, drawn from a uniform distribution, are

[7] This curve was formed from concepts generated with one irrelevant feature and 10 training examples. Results were similar for other concept distributions with irrelevant features.

Correlation (%)	*Mean Concept Complexity*	*Minimum Size Bias (%)*
50	6.73	50.0
75	5.83	56.5
100	1.00	92.0

Table 3: Mean concept complexity and minimum size decision tree bias benefit as a function of the correlation of the added feature used to form the concept distribution.

given a feature that is highly correlated with the class attribute.

For these results, a fifth correlated feature was added to randomly generated concepts over four boolean features (an example space of 16 examples). The mean correlation of the correlated feature was varied to form three different concept distribution samples. Correlation is defined as the percentage of examples in the example space where the correlated feature's value was the same as the class value[8]. Note, all concept distribution samples generated for these results have five boolean features and an example space of 16 unique examples.

Table 3 presents mean concept distribution complexity and the benefit of the minimum size decision tree bias as a function of the mean correlation of the added feature. As the mean correlation of the added feature increases, the mean complexity of the concept distribution decreases, and the benefit of the minimum size decision tree bias increases. Note, because there were only 16 examples in the example space, the maximum complexity for this distribution is 15.

[8] For uniformly distributed concepts over a boolean feature space, the mean correlation between any attribute and the class attribute is 50%.

5 Conclusion

Through a sampling of the space of uniformly distributed concepts of specific complexities (number of internal nodes in the smallest decision tree consistent with the example space), it is shown that the bias for small decision trees is beneficial. It is also shown, that while larger trees perform better than smaller trees when the complexity of the distribution of concepts is high, it is better to guess randomly than to use a size-based bias.

Explanations for why the distribution of concepts seen in practice are simple and amenable to the minimum size decision tree bias were given and evaluated empirically. It was shown that the use of highly class correlated features or the presence of irrelevant features can cause simpler concept distributions, under which the minimum size decision tree bias is beneficial.

Acknowledgments

I thank Rob Holte, Ross Quinlan, Mike Pazzani, Dennis Kibler, Chris Merz, Steve Hampson and Cliff Brunk for commenting on various aspects of this research. The research reported here was supported in part by NSF infrastructure grant number MIP-9205737, NSF Grant INT-9201842, AFOSR grant F49620-92-J-0430, and AFOSR AASERT grant F49620-93-1-0569.

References

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

Buntine, W. (1992). A further comparison of splitting rules for decision-tree induction. *Machine Learning, 8*(1), 75–85.

Murphy, P., & Pazzani, M. (1994). Exploring the decision forest: An empirical investigation of occam's razor in decision tree induction. *Journal of Artificial Intelligence Research*, *1*, 257–275.

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

Schaffer, C. (1994). A conservation law for generalization performance. In *Machine Learning: Proceedings of the Eleventh International Conference*, pp. 259–265.