Occam's Razor Just Got Sharper

Saher Esmeir and **Shaul Markovitch**
Computer Science Department, Technion—Israel Institute of Technology, Haifa 32000, Israel
{esaher, shaulm}@cs.technion.ac.il

Abstract

Occam's razor is the principle that, given two hypotheses consistent with the observed data, the simpler one should be preferred. Many machine learning algorithms follow this principle and search for a small hypothesis within the version space. The principle has been the subject of a heated debate with theoretical and empirical arguments both for and against it. Earlier empirical studies lacked sufficient coverage to resolve the debate. In this work we provide convincing empirical evidence for Occam's razor in the context of decision tree induction. By applying a variety of sophisticated sampling techniques, our methodology samples the version space for many real-world domains and tests the correlation between the size of a tree and its accuracy. We show that indeed a smaller tree is likely to be more accurate, and that this correlation is statistically significant across most domains.

1 Introduction

Occam's razor, attributed to the 14th-century English logician William of Ockham, is the principle that, given two hypotheses consistent with the observed data, the simpler one should be preferred. This principle has become the basis for many induction algorithms that search for a small hypothesis within the *version space* [Mitchell, 1982]. Several studies attempted to justify Occam's razor with theoretical and empirical arguments [Blumer *et al.*, 1987; Quinlan and Rivest, 1989; Fayyad and Irani, 1990]. But a number of recent works have questioned the utility of Occam's razor, and provided theoretical and experimental evidence against it.

Schaffer [1994] proved that no learning bias can outperform another bias over the space of all possible learning tasks. This looks like theoretical evidence against Occam's razor. Rao *et al.* [1995], however, argued against the applicability of this result to real-world problems by questioning the validity of its basic assumption about the uniform distribution of possible learning tasks.

Domingos [1999] argued that the disagreement about the utility of Occam's razor stems from the two different interpretations given to it: the first is that simplicity is a goal in and of itself, and the second is that simplicity leads to better accuracy. While accepting the first interpretation, Domingos questioned the second one.

Webb [1996] presented *C4.5X*, an extension to *C4.5* that uses similarity considerations to further specialize consistent leaves. Webb reported an empirical evaluation which shows that *C4.5X* has a slight advantage in a few domains and argued that these results discredit Occam's thesis.

Murphy and Pazzani [1994] reported a set of experiments in which all possible consistent trees were produced and their accuracy was tested. Their findings were inconclusive. They found cases where larger trees had, on average, better accuracy. Still, they recommend using Occam's principle when no additional information about the concept is available. The major limitation of their work is the exhaustive enumeration of the version space. Such an approach is only applicable to domains with very few features.

In this work we present an alternative approach that performs statistical testing of Occam's thesis on a sample of the version space. This approach allows us to use high-dimensional domains and complex concepts. One problem with random sampling of the version space is the rarity of small trees in the sample. We therefore use, in addition to random sampling, biased sampling methods based on modern anytime induction algorithms [Esmeir and Markovitch, 2004]. These methods produce samples with much higher concentrations of small trees.

The major contribution of this work is to provide convincing empirical evidence for Occam's razor in the context of classification trees. Furthermore, the various sampling techniques we applied help to better understand the space of consistent decision trees and how top-down induction methods explore it. Note that this empirical demonstration of the utility of Occam's principle does not pretend to provide a philosophical proof for Occam's thesis.

2 Occam's Empirical Principle

In the context of machine learning, the widely accepted interpretation of Occam's razor is that given two consistent hypotheses, the simpler one is *likely to have a lower error rate*. Fayyad and Irani [1990] have formally defined this notion: for two decision trees T_1 and T_2 and a fixed ϵ, $0 < \epsilon < 1$, T_1 is likely to have a lower error rate than T_2 if $Pr\left\{P\left(T_1, \epsilon\right) < P\left(T_2, \epsilon\right)\right\} > 0.5$, where $P(T, \epsilon)$ is the probability that T has an error rate greater than ϵ.

Procedure TDIDT(E, A)
  If $E = \emptyset$
    Return Leaf(nil)
  If $\exists c$ such that $\forall e \in E$ Class(e) $= c$
    Return Leaf(c)
  $a \leftarrow$ CHOOSE-ATTRIBUTE(A, E)
  $V \leftarrow$ domain(a)
  Foreach $v_i \in V$*
    $E_i \leftarrow \{e \in E \mid a(e) = v_i\}$
    $S_i \leftarrow$ TDIDT($E_i, A - \{a\}$)
  Return Node($a, \{\langle v_i, S_i \rangle \mid i = 1 \ldots |V|\}$)

* When a is numeric, a cutting point is chosen and and a is not filtered out when calling SID3.

Figure 1: Top-down induction of decision trees. E stands for the training set and A stands for the set of attributes.

Fayyad and Irani [1990] also provided theoretical support for favoring smaller trees. They showed that under a set of assumptions, given two trees T_1 and T_2 consistent with the observed data, T_1 is likely to have a lower error rate than T_2 if T_1 has fewer leaves. Berkman and Sandholm [1995], however, have questioned this set of assumptions and argued that the opposite conclusion can be drawn from it.

The main challenge we face in this work is to *empirically* test the validity of Occam's razor in decision tree induction. We therefore define the *Occam's empirical principle*:

Definition 1 *Let E_{train} and E_{test} be a training and a testing set respectively. Let H be a set of hypotheses consistent with E_{train}. We say that H satisfies Occam's empirical principle with respect to E_{train} and E_{test} if, for any h_1, h_2 drawn from $H \times H$,*

$$P\left(Acc\left(h_1, E_{test}\right) \geq Acc\left(h_2, E_{test}\right) \;\middle|\; |h_1| \leq |h_2|\right) \geq 0.5,$$

where $|h|$ is the size of hypothesis h and $0 \leq Acc(h, E) \leq 1$ is the accuracy of h on a test set E.

3 Sampling the Version Space

Given a learning problem, we would like to produce all the possible trees consistent with the observations and test whether their size and accuracy are correlated. Such an exhaustive enumeration, however, is not practical for most real-world domains. Therefore, in what follows we propose sampling as an alternative. First we define the population, i.e., the space of decision trees we sample from, and then we describe 3 different sampling techniques, each of which focuses on a different part of the sampled space.

3.1 Defining the Version Space

Given a set of attributes A, the hypothesis class we deal with is the set of decision trees over A, denoted by DT_A. Let E be a set of examples. Because Occam's razor is applicable only to hypotheses that can explain the observations, we limit our discussion to the *version space*—the set of all trees consistent with E, denoted by $DT_A(E)$. Furthermore, since most decision tree learners build a tree top-down, we focus on a subset of the consistent trees—the trees obtainable by top-down induction, denoted by $TDIDT_A(E)$. Under the TDIDT scheme, the set of examples is partitioned into subsets by testing the value of an attribute and then each subset is used to recursively build a subtree. The recursion stops when all the examples have the same class label. Figure 1 formalizes the basic procedure for top-down induction.

While $TDIDT_A(E)$ is a strict subset of $DT_A(E)$, we claim that the trees in $DT_A(E)$ that are not in $TDIDT_A(E)$ are not interesting for the purpose of model learning. There are two types of trees in $DT_A(E) - TDIDT_A(E)$:

1. A tree containing a subtree with all leaves marked with the same class. Obviously, such a subtree could have been replaced by a single node marked with the class.
2. A tree with an internal node that has no associated examples from E. The subtree rooted at this node is not supported by training examples and is therefore not interesting for induction.

Note that including the above trees could unjustly distort the results. In the first case, larger trees that are logically equivalent will be included, arbitrarily weakening the negative correlation. In the second case, the extra branches are not supported by any training example. Thus, their leaves must be labeled randomly, lowering the accuracy and hence arbitrarily strengthening the negative correlation.

While we restrict the definition of Occam's empirical principle to consistent hypotheses, in our experiments we also examine its applicability to pruned $TDIDT_A(E)$ trees. This allows us to draw conclusions for noisy datasets as well.

3.2 Sampling Techniques

Our goal is to sample the $TDIDT_A(E)$ space in order to test Occam's empirical principle. Our first proposed sampling technique uses TDIDT with a random selection of the splitting attribute (and cutting point, where the attribute is numeric). We refer to this method as the Random Tree Generator (*RTG*). Observe that although it has no bias with respect to generalization quality, *RTG* does not uniformly sample $TDIDT_A(E)$. For example, if the concept was $\overline{a_1}$ and the attributes were $\{a_1, a_2, a_3\}$, the probability of constructing the smallest tree (with a single split) is much higher than that of constructing a specific large tree. We will later show that the non-uniform sampling should not affect the validity of our conclusions.

One problem with random sampling is the rarity of small trees. Many induction methods, however, are likely to concentrate on small trees. Theoretically, the correlation could have been statistically significant when sampling the *TDIDT* space but not when sampling a subspace consisting of small trees. To test this hypothesis we need a sample of small trees. Such a sample could be obtained by repeatedly invoking *RTG* and keeping the smaller trees. Nevertheless, the number of *RTG* invocations needed to obatin a reasonable number of small trees is prohibitively high. Another alternative is to use *ID3*. Repeated invocations of *ID3*, however, result in similar trees that can vary only due to different tie-breaking decisions. Esmeir and Markovitch [2004] introduced *SID3*, a stochastic version of *ID3* that is designed

Procedure SID3-CHOOSE-ATTRIBUTE(E, A)
  Foreach $a \in A$
    $p(a) \leftarrow$ gain-1(E, a)
  If $\exists a$ such that entropy-1$(E, a) = 0$
    $a^* \leftarrow$ Choose attribute at random from $\{a \in A \mid \text{entropy-1}(E, a) = 0\}$
  Else
    $a^* \leftarrow$ Choose attribute at random from A; for each attribute a, the probability of selecting it is proportional to $p(a)$
  Return a^*

Figure 2: Attribute selection in *SID3*

Procedure LSID3-CHOOSE-ATTRIBUTE(E, A, r)
  If $r = 0$
    Return ID3-CHOOSE-ATTRIBUTE(E, A)
  Foreach $a \in A$
    Foreach $v_i \in \text{domain}(a)$
      $E_i \leftarrow \{e \in E \mid a(e) = v_i\}$
      $min_i \leftarrow \infty$
      Repeat r **times**
        $T \leftarrow \text{SID3}(E_i, A - \{a\})$
        $min_i \leftarrow \min(min_i, |T|)$
    $total_a \leftarrow \sum_{i=1}^{|\text{domain}(a)|} min_i$
  Return a for which $total_a$ is minimal

Figure 3: Attribute selection in *LSID3*

to sample the version space semi-randomly, with a bias to smaller trees. In *SID3*, instead of choosing an attribute that maximizes the information gain, we choose the splitting attribute semi-randomly. The likelihood that an attribute will be chosen is proportional to its information gain.[1] However, if there are attributes that decrease the entropy to zero, then one of them is picked randomly. The attribute selection procedure of *SID3* is listed in Figure 2.

For many hard learning tasks such as parity concepts, *ID3*'s greedy heuristic fails to correctly estimate the usefulness of the attributes and can mislead the learner to produce relatively large trees. In such cases, *SID3* can, in theory, produce significantly smaller trees. The probability for this, however, is low and decreases as the number of attributes increases. To overcome this problem, we use a third sampling technique that is based on the recently introduced *LSID3* algorithm for anytime induction of decision trees [Esmeir and Markovitch, 2004]. *LSID3* adopts the general TDIDT scheme, and invests more time resources for making better split decisions. For every candidate split, *LSID3* attempts to estimate the size of the resulting subtree were the split to take place, and favors the one with the smallest expected size. The estimation is based on a biased sample of the space of trees rooted at the evaluated attribute. The sample is obtained using *SID3*. *LSID3* is parameterized by r, the sample size. When r is greater, the sample is larger and the resulting estimate is expected to be more accurate. Therefore, *LSID3* is expected to improve with the increase in r. Figure 3 lists the procedure for attribute selection as applied by *LSID3*. Because our goal is to sample small trees and not to always obtain the smallest tree, we use *LSID3*($r = 1$) as a sampler. Observe that *LSID3* is stochastic by nature, and therefore we do not need to randomize its decisions.

4 Experimental Results

We tested Occam's empirical principle, as stated in Definition 1, on 20 datasets, 18 of which were chosen arbitrarily from the UCI repository [Blake and Merz, 1998], and 2 which are artificial datasets that represent hard concepts: XOR-5 with 5 additional irrelevant attributes, and 20-bit Multiplexer. Each dataset was partitioned into 10 subsets that were used to create 10 learning problems. Each problem consisted of one subset serving as a testing set and the union of the remaining 9 as a training set, as in 10-fold cross validation. We sampled the version space, $TDIDT_A(E)$, for each training set E using the three methods described in Section 3 and tested the correlation between the size of a tree (number of leaves) and its accuracy on the associated testing set. The size of the sample was ten thousand for *RTG* and *SID3*, and one thousand for *LSID3* (due to its higher costs). We first present and discuss the results for consistent trees and then we address the problem of pruned, inconsistent trees.

4.1 Consistent Decision Trees

Figure 4 plots size-frequency curves for the trees obtained by each sampling method for three datasets: Nursery, Glass, and Multiplexer-20 (for one fold out of the 10). Each of the three methods focuses on a different subspace of $TDIDT_A(E)$, with the biased sampling methods producing samples consisting of smaller trees. In all cases we see a bell-shaped curve, indicating that the distribution is close to normal. Recall that *RTG* does not uniformly sample $TDIDT_A(E)$: a specific small tree has a better chance of being built than a specific large tree. The histograms indicate, however, that the frequency of small trees in the sample is similar to that of large trees (symmetry). This can be explained by the fact that there are more large trees than small trees. To further verify this, we compared the distribution of the tree size in an RTG sample to that of all trees, as reported in [Murphy and Pazzani, 1994] (Mux-11 dataset). The size-frequency curves for the full space and the sampled space were found to be similar.

Occam's empirical principle states that there is a negative correlation between the size of a tree and its accuracy. To test the significance of the correlation, we used the nonparametric Spearman correlation test on each of the samples.[2] Spearman's coefficient ρ measures the monotonic association of two variables, without making any assumptions about their frequency distribution. For a paired sample of X and Y, ρ is defined as $1 - 6\sum d_i/(n(n^2-1))$, where d_i is the difference in the statistical rank of x_i and y_i. There is a special correction for this formula in the presence of ties.

Table 4.1 lists summarizing statistics for the *RTG*, *SID3*, and *LSID3* samplers. The validity of Occam's empirical prin-

[1] SID3 ensures that attributes with gain of zero will have a positive probability to be selected.

[2] All statistics were computed using the *The R Project* package [R Development Core Team, 2005].

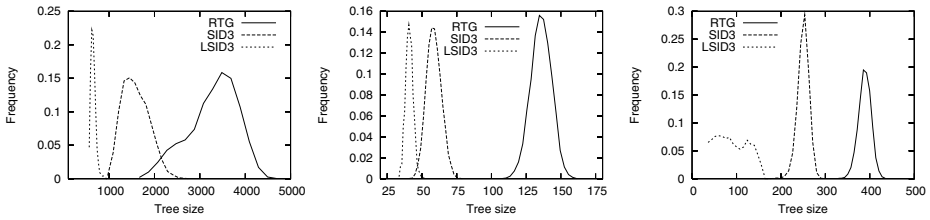

Figure 4: Frequency curves for the Nursery (left), Glass (middle), and Multiplexer-20 (left) datasets

	RTG				SID3				LSID3			
DATASET	ACC.	SIZE	ρ	$\surd$	ACC.	SIZE	ρ	$\surd$	ACC.	SIZE	ρ	$\surd$
BREAST-W	92.9±2.8	128±14	-0.1	7	93.1±2.7	108±11	-0.1	8	94.3±1.6	77±4	-0.1	5
BUPA	59.7±8.0	213±10	0	10	63.4±7.3	94±6	0	7	61.9±7.4	69±3	0	3
CAR	72.8±4.6	647±77	-0.8	10	79.7±5.8	520±95	-0.9	10	91.9±1.1	285±13	0	3
CLEVELAND	51.6±6.9	188±7	0	6	50.2±7.2	134±7	0	7	46.1±7.4	98±5	-0.1	7
CORRAL	73.3±22.9	15±3	-0.1	9	81.6±19.6	10±2	-0.2	10	89.8±8.3	7±1	0.2	NA
GLASS	55.6±9.9	135±8	-0.1	10	62.3±9.3	57±5	-0.2	10	68.0±8.4	39±3	-0.1	9
HUNGERIAN	72.8±7.4	125±10	-0.1	9	73.3±7.2	65±6	-0.1	8	69.9±7.5	47±3	-0.1	8
IRIS	88.9±7.3	39±9	-0.3	10	92.8±4.5	12±2	-0.1	8	93.8±2.7	8±0	-0.1	7
MONKS-1	91.1±4.6	203±42	-0.5	10	97.0±3.8	113±55	-0.7	10	100.0±0.0	28±4	NA	NA
MONKS-2	77.8±4.4	294±9	-0.3	10	75.5±4.6	289±8	-0.4	10	77.7±3.1	259±3	0.2	0
MONKS-3	88.3±5.4	171±46	-0.6	10	96.0±2.5	77±33	-0.5	10	96.7±0.4	38±2	0.1	NA
MUX-20	55.7±6.3	388±14	-0.1	10	56.6±6.5	249±13	-0.2	10	86.1±11.8	89±35	-0.9	10
NURSERY	77.2±5.9	3271±551	-0.9	10	93.1±1.8	1583±295	-0.8	10	98.1±0.5	656±54	-0.4	10
SCALE	72.2±4.2	394±11	0.1	0	71.7±4.1	389±11	0.1	0	70.1±3.8	352±5	0.1	3
SPLICE	61.1±3.7	1977±101	-0.6	10	60.5±4.2	1514±112	-0.7	10	89.3±1.8	355±23	-0.5	10
TIC-TAC	72.3±4.8	468±34	-0.4	10	80.2±4.5	311±30	-0.4	10	87.7±3.0	166±11	-0.1	9
VOTING	89.2±5.8	52±12	-0.3	10	92.8±4.3	26±5	-0.2	9	94.5±3.1	15±2	-0.2	8
WINE	78.6±10.2	73±12	-0.3	10	90.6±6.7	13±3	-0.2	9	91.7±4.3	7±1	-0.1	6
XOR-5	50.7±10.6	136±8	-0.1	10	51.9±11.8	108±11	-0.4	10	96.5±7.7	39±11	-0.8	10
ZOO	90.0±7.4	24±5	-0.2	10	91.8±6.2	18±4	-0.2	9	94.2±3.5	11±1	-0.1	NA

Table 1: Testing Occam's empirical principle using different sampling methods that produce consistent trees. For each method we report the accuracy, tree size, and Spearman's correlation coefficient (ρ) averaged over all 10 partitions. We also report the number of times (out of 10) that a negative correlation was found to be statistically significant with $p = 0.95$.

ciple, tested by Spearman's method, is listed in the rightmost column. For each sampling method, for each dataset, we count how many times, out of the 10 folds, the null hypothesis H_0 (which states that the variables are not correlated) can be rejected at an $\alpha = 5\%$ significance level, against the alternative hypothesis that the correlation is negative.

The results indicate that when random sampling (*RTG*) in used, Occam's empirical principle, as measured by Spearman's test, is valid for almost all problems, except for Scale. The results for the *SID3* sampling method indicate that even when focusing on smaller trees, simplicity is beneficial as a bias for accuracy. Again, except for the Scale dataset, there is a strong inverse correlation between size and accuracy. The numbers indicate that the correlation is weaker than the *RTG* case, yet still significant across most domains.

The *LSID3* samples focus on very small trees. In several cases, *LSID3* could reach the smallest tree possible. Again the negative correlation was found to be significant for most domains.[3] However, the number of cases where the null hypothesis could not be rejected is higher than in the previous samplers. One possible reason for this phenomenon is that *LSID3* samples trees from a much tighter size-range. Hence, there is an increased probability for finding two trees, $T1$ and $T2$, with the size and accuracy of $T1$ being greater than T_2.

As indicated by the frequency curves, the different sampling methods cover different portions of the space, but sometimes overlap. An interesting question is whether the conclusions hold if we analyze all samples together. Note that it is not statistically valid to merge the samples of *RTG*, *SID3*, and *LSID3* due to their different distributions. Nevertheless, we measured the correlation statistics for the combined samples and found that the results are very similar, with a strong negative correlation between the size of a tree and its accuracy.

To illustrate the correlation between the size of a tree and its accuracy, we grouped the trees from a single run into bins, according to their size, and calculated the average accuracy for each bin. We did this for each of the sampling methods. Bins with less than 5 observations were discarded. Figure 5 plots the results for the Nursery, Glass, and Multiplexer-20 datasets. The error bars represent confidence intervals, with $\alpha = 5\%$.

[3]Note that in some cases, significance tests could not be computed due to the large number of ties (indicated by NA in the table).

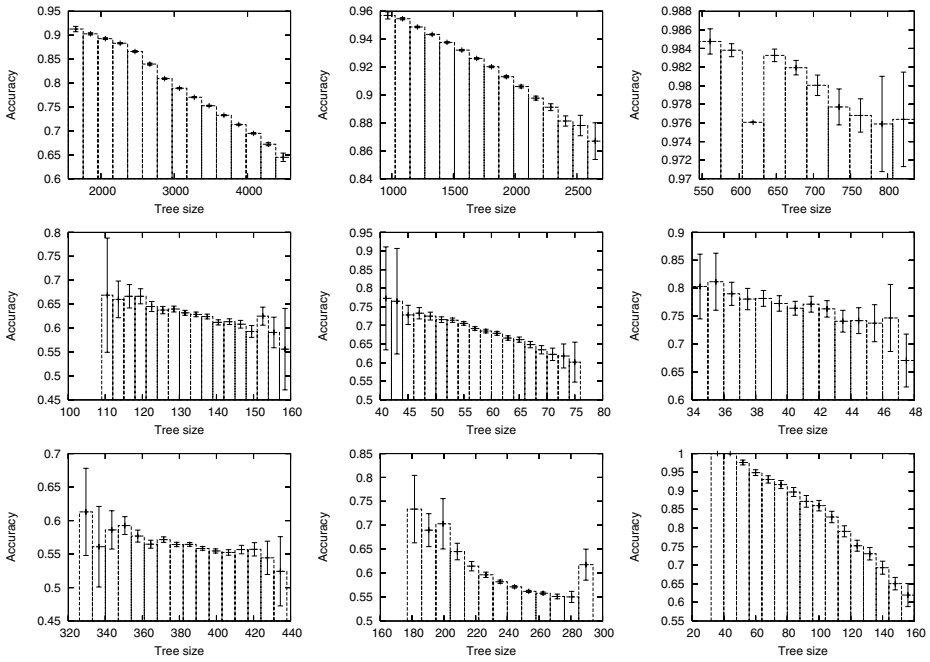

Figure 5: Correlation between size and accuracy using *RTG* (left-most), *SID3* (middle), and *LSID3* (right-most). The upper graphs represent the results for the Nursery dataset, the graphs in the middle row stand for the Glass dataset and the lower graphs stand for the Multiplexer-20 dataset.

Again, the graphs show a strong correlation between size and accuracy, confirming Occam's empirical principle. For the Nursery and Multiplexer-20 datasets the correlation is strong for all 3 samplers. For Glass, the correlation is weaker when the trees are very small. These graphs, which represent each size range in its own bin, indicate that the positive support we showed for Occam's principle is not a result of a size bias in our sampling methods.

4.2 Pruned Decision Trees

Formally, Occam's empirical principle, as defined in Section 2, is applicable only to consistent hypotheses. Most decision tree learners, however, do not necessarily produce consistent models because they output a pruned tree in attempt to avoid overfitting the data. This is usually done in two phases: first a tree is grown top-down and then it is pruned. In our second set of experiments, we examine whether taking simplicity as a bias in the first stage is beneficial even when the tree is later pruned. Therefore, we measure the correlation between the size of unpruned trees and their accuracy after pruning.

Table 4.2 summarizes the results. The same statistics were measured with a single change: the accuracy was measured after applying error-based pruning [Quinlan, 1993]. As in the case of consistent trees, examining the overall correlation between the size of a tree and its accuracy indicates that for most datasets the inverse correlation is statistically significant.

Table 4.2 also gives the percentages of pruned leaves. While the trees produced by *RTG* were aggressively pruned, the percentage of pruned leaves in *LSID3* was relatively low. This is due to the stronger support for the decisions made at the leaves of smaller consistent trees.

5 Conclusions

Occam's razor states that given two consistent hypotheses, the simpler one should be preferred. This principle has been the subject for a heated debate, with theoretical and empirical arguments both for and against it. In this work we provided convincing empirical evidence for the validity of Occam's principle with respect to decision trees. We state *Occam's empirical principle*, which is well-defined for any learning problem consisting of a training set and a testing set, and show experimentally that the principle is valid for many known learning problems. Note that our study is purely empirical and does not attempt to reach an indisputable conclusion about Occam's razor as an epistemological concept.

Our testing methodology uses various sampling techniques to sample the version space and applies Spearman's correlation test to measure the monotonic association between the size of a tree and its accuracy. Our experiments confirm Occam's empirical principle and show that the negative correlation between size and accuracy is strong. Although there were several exceptions, we conclude that in general, simpler trees are likely to be more accurate. Observe that our results do not contradict those reported by Murphy and Pazzani [1994], but complement them: we do not claim that a smaller tree is always more accurate, but show that for many domains smaller trees are likely to be more accurate.

We view our results as strong empirical evidence for the utility of Occam's razor to decision tree induction. It is important to note that the datasets used in our study do not necessarily represent all possible learning problems. However, these datasets are frequently used in machine learning research and considered typical tasks for induction algorithms.

	RTG					SID3					LSID3				
DATASET	ACC.	SIZE	%P	ρ	$\surd$	ACC.	SIZE	%P	ρ	$\surd$	ACC.	SIZE	%P	ρ	$\surd$
BREAST-W	92.8±2.9	128±14	82	-0.1	9	93.3±2.8	108±11	77	-0.1	7	94.1±1.3	77±4	70	-0.3	7
BUPA	62.7±7.2	213±10	86	0	7	64.9±7.0	94±6	46	0	5	62.2±7.3	69±3	11	0	2
CAR	80.8±3.5	647±77	93	-0.6	10	84.0±4.1	520±95	87	-0.7	10	91.6±2.0	285±13	56	-0.3	9
CLEVELAND	54.6±5.8	188±7	73	0	5	52.9±6.5	134±7	48	0	6	47.1±7.3	98±5	12	-0.1	7
CORRAL	74.6±20.3	15±3	67	-0.1	9	78.2±19.5	10±2	46	-0.3	10	86.0±13.0	7±1	9	0	7
GLASS	58.7±9.4	135±8	73	-0.1	10	63.5±9.1	57±5	25	-0.2	10	68.1±8.4	39±3	5	-0.1	9
HUNGERIAN	77.5±5.3	125±10	92	0	5	76.1±6.3	65±6	65	0	5	71.3±7.3	47±3	23	-0.1	5
IRIS	91.5±6.6	39±9	82	-0.2	10	93.5±3.8	12±2	47	-0.1	7	93.8±3.1	8±0	20	-0.1	5
MONKS-1	90.9±6.3	203±42	89	-0.6	10	95.2±6.2	113±55	77	-0.7	10	100.0±0.0	28±4	2	-0.1	NA
MONKS-2	65.5±2.1	294±9	97	0	6	65.5±2.2	289±8	97	0	5	67.3±4.1	259±3	78	0	4
MONKS-3	93.6±4.9	171±46	92	-0.6	10	97.0±2.6	77±33	84	-0.6	10	98.9±0.3	38±2	65	-0.2	NA
MUX-20	57.4±6.4	388±14	84	-0.1	10	58.2±6.7	249±13	61	-0.2	10	87.8±11.5	89±35	32	-0.9	10
NURSERY	88.4±2.3	3271±551	97	-0.7	10	92.1±1.7	1583±295	90	-0.7	10	96.1±0.7	656±54	61	-0.5	10
SCALE	69.4±4.2	394±11	90	0	3	69.2±4.2	389±11	90	0	4	68.6±3.8	352±5	89	0.2	0
SPLICE	63.3±4.8	1977±101	91	-0.6	10	66.1±5.4	1514±112	81	-0.6	10	91.9±1.4	355±23	58	-0.4	10
TIC-TAC	78.0±4.3	468±34	88	-0.3	10	81.0±4.1	311±30	77	-0.4	10	87.2±3.1	166±11	36	-0.1	6
VOTING	94.2±4.5	52±12	92	-0.3	10	96.2±2.2	26±5	89	-0.1	8	96.4±1.3	15±2	79	0	5
WINE	82.2±9.6	73±12	84	-0.2	10	91.1±6.5	13±3	22	-0.2	9	91.7±4.3	7±1	8	-0.1	6
XOR-5	54.8±11.7	136±8	85	-0.2	10	55.6±12.5	108±11	77	-0.4	10	97.1±7.2	39±11	18	-0.7	10
ZOO	90.0±8.0	24±5	53	-0.2	10	91.0±7.1	18±4	34	-0.2	9	94.3±3.5	11±1	5	0	NA

Table 2: Testing Occam's empirical principle when the trees are pruned. For each method we report the accuracy, tree size, percentage of pruned leaves, and Spearman's correlation coefficient (ρ) averaged over all 10 partitions. We also report the number of times (out of 10) that a negative correlation between the size of the unpruned trees and their accuracy after pruning was found to be statistically significant with $p = 0.95$.

We also examined the applicability of Occam's razor to pruned trees. For most domains we found the correlation between the size of a tree before pruning and its accuracy after pruning to be statistically significant. These results indicate that taking Occam's razor as a preference bias when growing a tree is useful even if the tree is post-pruned.

An interesting research direction that we plan to pursue is to test the correlation between a tree's accuracy and a variety of other measures, such as expected error and description length.

Acknowledgments

This work was partially supported by funding from the EC-sponsored MUSCLE Network of Excellence (FP6-507752).

References

[Berkman and Sandholm, 1995] N. C. Berkman and T. W. Sandholm. What should be minimized in a decision tree: A re-examination. Technical report, 1995.

[Blake and Merz, 1998] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998.

[Blumer *et al.*, 1987] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's Razor. *Information Processing Letters*, 24(6):377–380, 1987.

[Domingos, 1999] P. Domingos. The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999.

[Esmeir and Markovitch, 2004] S. Esmeir and S. Markovitch. Lookahead-based algorithms for anytime induction of decision trees. In *ICML'04*, pages 257–264, 2004.

[Fayyad and Irani, 1990] U. M. Fayyad and K. B. Irani. What should be minimized in a decision tree? In *AAAI'90*, pages 749–754, 1990.

[Mitchell, 1982] T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.

[Murphy and Pazzani, 1994] P. M. Murphy and M. J. Pazzani. Exploring the decision forest: an empirical investigation of Occam's Razor in decision tree induction. *Journal of Artificial Intelligence Research*, 1:257–275, 1994.

[Quinlan and Rivest, 1989] J. R. Quinlan and R. L. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80(3):227–248, 1989.

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

[R Development Core Team, 2005] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.

[Rao *et al.*, 1995] R. Rao, D. Gordon, and W. Spears. For every generalization action, is there really an equal or opposite reaction? In *ICML'95*, pages 471–479, 1995.

[Schaffer, 1994] C. Schaffer. A conservation law for generalization performance. In *ICML'94*, pages 259–265, 1994.

[Webb, 1996] G. I. Webb. Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, 4:397–417, 1996.