What should be minimized in a decision tree: A re-examination

Neil C. Berkman and Tuomas W. Sandholm
Computer Science Department
University of Massachusetts at Amherst
CMPSCI Technical Report 95-20
September 6, 1995

What should be minimized in a decision tree: A re-examination

Neil C. Berkman and Tuomas W. Sandholm *
{berkman,sandholm}@cs.umass.edu
Tel. +1-413-545-0675, Fax. +1-413-545-1249

University of Massachusetts at Amherst
Computer Science Department
Amherst, MA 01003

Abstract

This paper examines a recent attempt to justify an inductive bias toward decision trees with few leaves. It is shown that this argument is invalid because it rests upon questionable assumptions, and can be used to deduce contradictory conclusions. Specifically, it can be used to prescribe any inductive bias. In general, it is shown that one cannot justify a preference for any inductive bias over another without making *a priori* assumptions about the distribution of target concepts. These results refute one common justification for Occam's Razor, which recommends preferring simple hypotheses over complex ones when both are consistent with a set of observations.

Keywords: Occam's Razor, inductive bias, decision tree

*Neil Berkman was supported by the National Science Foundation under Grant No. IRI-9222766. Tuomas Sandholm was supported by ARPA contract N00014-92-J-1698. The content does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

1 Introduction

The notion that the accuracy of an explanation is associated with its simplicity dates back at least as far as the 14th century, when William of Occam first posited his famous razor. More recently, machine learning researchers have followed this principle in biasing their algorithms toward finding hypotheses with simple representations. Two approaches have been used to justify Occam's Razor:

- **Simplicity of Nature:** This is the notion that nature exhibits regularity, i.e. that natural phenomena are more often simple than complex. Alternatively one could argue that the phenomena humans choose to study tend to have simple explanations. For either reason, simple theories would better describe the phenomena under study.
- **Rarity of simple theories:** An alternative justification is expressed: "There are far fewer simple hypotheses than complex ones, so that there is only a small chance that *any* simple hypothesis that is wildly incorrect will be consistent with all observations. Hence, other things being equal a simple hypothesis that is consistent with the observations is more likely to be correct than a complex one." (Russell & Norvig, 1995)

In this paper, we do not dispute Occam's Razor. We believe that the first justification may well hold; that is, the phenomena that are studied in practice exhibit simplicity due to either inherent simplicity of natural problems or simplicity-directed biased sampling from among natural problems. Thus, this argument for simple theories relies on an *a priori* distribution of target concepts. Such a distribution has to be taken as a conjecture and is not provable as a theorem. The second justification for Occam's Razor assumes no such distribution. We claim that the second justification is questionable. We first discuss this in general and then analyze a specific example of an argument along these lines, a recent attempt (Fayyad, 1991) to justify formally a bias toward decision trees with few leaves. This line of reasoning relies on unsubstantiated assumptions and can be used to arrive at contradictory conclusions.

In Section 2 we argue that to state a preference for any inductive bias, one must make *a priori* assumptions about the distribution of target concepts. Section 3 reviews Fayyad's argument in favor of small trees. In Section 4 we

show that the assumptions of this argument are unfounded, and necessarily untrue in some cases. Section 5 shows how this "proof" can be used to justify a preference for any decision tree over any other, for example trees with large numbers of leaves over trees with few leaves. The fact that the same argument can produce results diametrically opposed to each other indicates that it cannot be used to justify any specific inductive bias. Section 6 demonstrates that, although Fayyad's argument relies on counting representations of hypotheses rather than hypotheses themselves, our objections would hold even if his argument were modified to count hypotheses directly. Section 7 presents conclusions.

2 Need for assumptions on priors

A general issue (Wolpert, 1994b; Wolpert, 1995; Schaffer, 1994; Schaffer, 1993; Wolpert, 1993) that has received much attention in the machine learning community lately is that under certain conditions, no inductive bias can validly be preferred over another without making certain *a priori* assumptions. Specifically such assumptions have to be made regarding the distribution of *target concepts*. We define a target concept as the combination of a *labeling* and a *distribution of feature vectors*. Consider a classification problem with k binary features, each of which can take on value 0 or 1. A labeling f is a mapping from feature vectors $V = \{0,1\}^k$ to classes C. Similarly, π is a probability distribution of feature vectors if $\pi : V \rightarrow [0,1]$, and $\sum_{\vec{v} \in V} \pi(\vec{v}) = 1$.

The fact that no inductive bias can be justified over another holds, for example, under the conditions that the hypotheses under consideration are consistent with the training data, that no noise is present, and that the training and test sets are drawn independently according to the same fixed distributions of f and π. This point can be illustrated by the following simple example. Let us consider an inductive bias that prescribes preferring one consistent hypothesis (H_1) over another (H_2). By the fact that $H_1 \neq H_2$, there is at least one labeling f^* such that for a uniform π, the hypothesis H_1 has a greater error rate than H_2. If no *a priori* assumptions are made about π and the distribution of f's, an adversary can choose a uniform π and a distribution of f's such that f^* has probability one and all other f's have probability zero. For these priors, the prescribed inductive bias is clearly worse than its complement. But the inductive bias was chosen arbitrarily.

Thus no inductive bias can validly be preferred over another without assuming any properties of the aforementioned priors [1]. This argument casts severe doubts upon the second (rarity of simple theories) justification of Occam's Razor.

Wolpert has presented similar results (the No Free Lunch theorems (Wolpert, 1992)) concerning *generalization performance*, accuracy on examples not found in the training set. In the case where all hypotheses under consideration are consistent with the training set (which is usually the case for decision tree induction algorithms when no noise is present), any error must result from generalization error. Thus, Wolpert's results verify the above finding.

Similarly, Schaffer's *Conservation Law* (Schaffer, 1994), a restatement of some of the No Free Lunch theorems, demonstrates that expected generalization performance over all learning situations is zero. In other words, the existence of a concept for which a particular inductive bias leads to good generalization implies the existence of a different concept for which the bias leads to poor generalization. Schaffer's analysis weights all possible labelings equally; thus, it corresponds to a situation in which the f's are uniformly distributed [2]. Again, the existence of a distribution for which all biases have equal generalization performance necessitates that assumptions be made about π and the distribution of f's in order to prescribe any inductive bias.

These results call into question several previously published results. For example, Wolpert has shown that the Bayesian "Occam factors" argument for Occam's Razor is incorrect (Wolpert, 1994a). The remainder of our paper is an examination of Fayyad's attempt to theoretically justify a bias toward small decision trees [3]. This argument makes no assumptions about π or the

[1] Note that this does not mean that one must assume a particular π or distribution of f's, but that one must assume some properties of these.

[2] While the Conservation Law itself is sound, we do not agree with one of the conclusions Schaffer draws from it. Specifically, he points out several examples of real-world datasets for which many common learning algorithms perform surprisingly poorly. While this phenomenon is interesting in itself, it is not a necessary consequence of the Conservation Law. The Law concerns only the *existence* of difficult problems for a given algorithm, but says nothing about the *likelihood* that they occur in the real world. To make an argument that such difficult problems occur in the real world for any algorithm, one needs to know something *a priori* about π and the distribution of f's, as pointed out in Rao et al. (Rao, Gordon & Spears, 1995).

[3] The bias toward small decision trees has been questioned before. Murphy and Pazzani (1994) present example target concepts for which larger decision trees are more accurate. Webb (1995) shows that by systematically increasing the complexity of trees found by

distribution of f's. A preference for small decision trees is a form of inductive bias; therefore, Fayyad's conclusion is inconsistent with the above results. In the remainder of this paper, we resolve this contradiction by pointing out flaws in Fayyad's argument. We present his proof in detail, in order to be able to examine later exactly how this argument fails. We feel that such a detailed analysis is valuable, not only to point out that the specific conclusion is erroneous, but also to help other researchers avoid similar flaws in the process of constructing arguments about induction algorithms.

3 The argument for small trees: a review

In his dissertation "On the induction of decision trees for multiple concept learning," Fayyad attempts to show [4] that given two decision trees consistent with a set of data, the tree with fewer leaves is likely to have a lower error rate (as defined below). In order to present our objections, it will be necessary to summarize the proof as presented by Fayyad. First we present some preliminary assumptions made in the argument.

> [...] it is assumed that the learning algorithm generates a decision tree that classifies all examples in the *training set* correctly. Another assumption we make is that the sets of training and test examples are *noise-free* and *not ambiguous* [5]. Furthermore, we assume that the examples in the training and test sets are drawn independently and at random according to some fixed (unknown) probability distribution. Finally, the training examples are expressed in terms of attribute-value pairs [...] These assumptions admit most decision tree algorithms in the literature [...]

Fayyad also presents a theorem that shows that for any decision tree T with n leaves, there exists a binary decision tree T' with n leaves that represents the same hypothesis. Using this fact, Fayyad confines his analysis to binary trees.

C4.5 (Quinlan, 1993) in such a way that performance on the training set is unchanged, trees more accurate than the original usually result.

[4] The argument in the dissertation is an updated version of an argument appearing earlier (Fayyad & Irani, 1990).

[5] A data set is said to be *ambiguous* when it contains at least two examples that agree on the values of all the attributes but belong to different classes. [Fayyad's footnote]

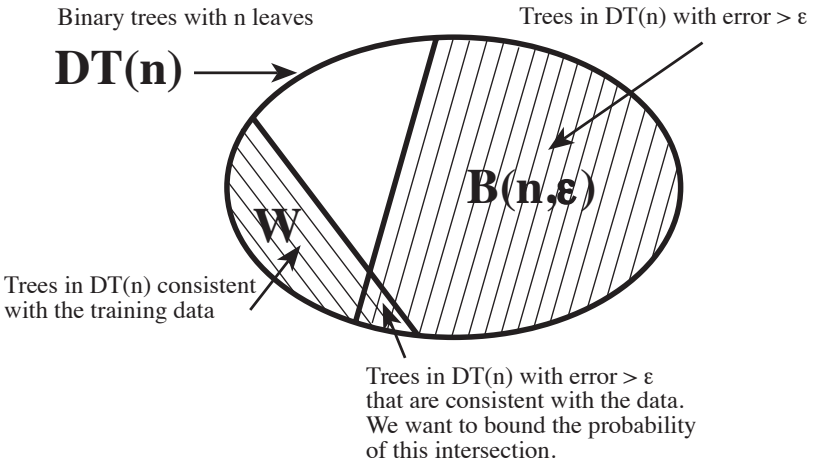


Figure 1: *Venn diagram of binary decision trees with n leaves.*

The learning situation consists of a training set of m examples consistent with some target concept Q that we wish to approximate. Define the *error rate* of a decision tree T as the probability that T will misclassify (i.e. assign a classification different from that assigned by Q) a test instance that is drawn randomly from our fixed probability distribution.

Let $DT(n)$ be the set of binary decision trees with n leaves. Call a tree *bad* if it has error rate greater than ϵ, $0 < \epsilon < 1$. Let $B(n, \epsilon) \subseteq DT(n)$ denote the set of bad binary trees with n leaves. In inductive learning, one generally does not know the target concept Q, and thus will not be able to determine $B(n, \epsilon)$.

Call a tree *consistent* if it correctly classifies all instances in a given training set. Let $W \subseteq DT(n)$ be the set of consistent trees with n leaves. Figure 1 depicts the relationship between $DT(n)$, $B(n, \epsilon)$, and W.

The first step in the proof is to bound the probability that a given consistent tree with n leaves is bad.

Theorem 3.3.2 [Fayyad [6]] *Let T be a binary tree having n leaves that classifies a set of N randomly chosen training examples correctly. Then $|B(n, \epsilon)| \cdot (1 - \epsilon)^N$ is an upper bound on the probability that T has an error rate greater than ϵ, $0 < \epsilon < 1$.*

[6]This theorem is similar to the more general Theorem 1 in Blumer et al. (1987).

Let $p_{bc} = \text{Prob}\{T \in B(n,\epsilon) \mid T \in W\}$, the probability that a tree T is bad given that it is consistent. One would like to know p_{bc}, but to calculate it exactly would require additional information that is unavailable (for example, it would be sufficient to know the cardinalities of W and $B(n,\epsilon) \cap W$). Instead, the theorem uses the fact that the *particular* consistent tree in which we are interested can be bad only if *some* tree exists that is both bad and consistent. Thus, p_{bc} is bounded above by $\text{Prob}\{W \cap B(n,\epsilon) \neq \emptyset\}$, the probability that at least one tree is both consistent and bad. Now, $\text{Prob}\{W \cap B(n,\epsilon) \neq \emptyset\}$ can be computed by Prob $\left\{\bigcup_{T' \in B(n,\epsilon)}\{E_{T'}\}\right\}$, where $E_{T'}$ is the event that $T' \in W$. The probability of this union cannot be computed directly, but an upper bound $|B(n,\epsilon)| \cdot Prob\{E_{T'}\}$ may be determined by subadditivity. Now, $Prob\{E_{T'}\} < (1-\epsilon)^N$ because the probability that T' correctly classifies one instance drawn from the fixed distribution is less than $(1-\epsilon)$ (since $T' \in B(n,\epsilon)$) and there are N training instances. So we have obtained an upper bound on an upper bound of the probability in which we are interested (p_{bc}) as follows:

$$|B(n,\epsilon)| \cdot (1-\epsilon)^N > \text{Prob}\{W \cap B(n,\epsilon) \neq \emptyset\} \geq p_{bc}$$

Note that unless $N > -\frac{\log|B(n,\epsilon)|}{\log(1-\epsilon)}$, this upper bound will be greater than 1 and will not bound the probability at all.

The next step in Fayyad's argument is to show that the number of trees with error rate greater than ϵ increases with the number of leaves:

Lemma 3.3.6 [Fayyad] *For any fixed ϵ, $0 < \epsilon < 1$, and any $n_2 > n_1 \geq 2$ the following property holds:*

$$\frac{|B(n_2,\epsilon)|}{|B(n_1,\epsilon)|} > 2^{2(n_2-n_1)}.$$

The proof of this lemma relies on the fact that for $n \geq 2$, any bad tree, say T, with n leaves can be augmented to produce at least 4 bad trees with $n+1$ leaves, each logically equivalent to T (i.e., they represent the same hypothesis). The general case where $n_2 > n_1 \geq 2$ is shown by induction on $n_2 - n_1$.

Fayyad presents the following corollary to Lemma 3.3.6:

Corollary 3.3.1 [Fayyad] *Let T_1 and T_2 be two decision trees consistent with a fixed training set of N examples. Let n_1 and n_2 be the*

number of leaves in T_1 and T_2 respectively. For a fixed ϵ, $0 < \epsilon < 1$, let b_1 and b_2 be the bounds derived in Theorem 3.3.2 for T_1 and T_2 respectively.

If $n_2 > n_1 \geq 2$ then $b_1 < b_2$. Furthermore

$$\frac{b_2}{b_1} > 2^{2(n_2 - n_1)}.$$

Following his presentation of Corollary 3.3.1, Fayyad discusses its implications:

> Thus for a fixed training set, given two decision trees T_1 and T_2 with n_1 and n_2 leaves respectively, let $P_1 = \text{Prob}\{T_1 \text{ has error rate } > \epsilon\}$ and $P_2 = \text{Prob}\{T_2 \text{ has error rate } > \epsilon\}$. Note that if $n_1 < n_2$ it follows from Theorem 3.3.2 that: $P_1 < b_1$ and $P_2 < b_2$. Corollary 3.3.1 states that b_1 is smaller than b_2 by a factor of $2^{2(n_2 - n_1)}$. However, Corollary 3.3.1 *does not* imply that $P_1 < P_2$. Proving this would be desirable but is not possible because the trees T_1 and T_2 were derived by an induction process that examined the same finite subset of the set of all possible examples. The corollary *does* state, however, that the *upper bound* on the probability that T_1 has an error rate that exceeds ϵ is always lower than the corresponding upper bound for T_2. [...]
>
> We shall denote the probability that a tree T has error greater than ϵ by $P(T, \epsilon)$, i.e.
>
> $$P(T, \epsilon) = \text{Prob}\{T \text{ has error rate } > \epsilon\}$$
>
> **Definition 3.3.4:** We say that a tree T_1 is *likely to have a lower error rate* than a tree T_2 if, for a fixed ϵ, $0 < \epsilon < 1$,
>
> $$\text{Prob}\{P(T_1, \epsilon) < P(T_2, \epsilon)\} > \frac{1}{2}$$
>
> i.e. when it is more likely that $P(T_1, \epsilon) < P(T_2, \epsilon)$ than otherwise.
>
> Given two decision trees, we should, at least on an intuitive level, prefer the one that is *likely to have a lower error rate* than the other. Assume that, for a fixed training set of N examples, we are given two decision trees T_1 and T_2 having n_1 and n_2 leaves respectively, with

$n_2 > n_1$. If we have no special knowledge of the data or the trees, we only have one more piece of information, namely that for any fixed ϵ, Corollary 3.3.1 states that the bound b_1 on $P(T_1, \epsilon)$ is exponentially smaller than the corresponding bound b_2 for $P(T_2, \epsilon)$. How may we use this piece of information? Is there a formally justifiable strategy that justifies our intuitive tendency to prefer T_1 over T_2?

Having no further knowledge, it seems reasonable to assume that $P(T_1, \epsilon)$ could be any value less than b_1, i.e. that $P(T_1, \epsilon)$ is uniformly distributed over the range $[0, b_1)$. [...] Making the same assumption for T_2 and its bound b_2, the following corollary will formally justify our intuitive tendency to prefer T_1 over T_2.

Corollary 3.3.2 [Fayyad] *Given two decision trees T_1 and T_2 having n_1 and n_2 leaves respectively, if $n_1 < n_2$ then assuming that $P(T_1, \epsilon)$ and $P(T_2, \epsilon)$ are uniformly distributed below their respective bounds b_1 and b_2, then T_1 is likely to have a lower error rate than T_2.*

Given the assumptions that $P(T_1, \epsilon)$ and $P(T_2, \epsilon)$ are uniformly distributed below their respective bounds, it is only necessary to show that $b_1 < b_2$ to show that T_1 is "likely to have a lower error rate" than T_2. In Section 4 we will examine the assumption that the probabilities are uniformly distributed below their bounds.

After presenting Corollary 3.3.2, Fayyad discusses its implications:

Corollary 3.3.2 justifies the strategy that prescribes preferring the tree with the smaller number of leaves. Note that Corollary 3.3.2 *does not* state that a tree with the minimal number of leaves is the best tree. The key condition is that the tree must be consistent with the training examples. For example, the tree consisting of a single node is not consistent with the training set (unless all training examples are of one class, in which case it would be the best tree). Furthermore, the corollary does not state that $P(T_1, \epsilon) < P(T_2, \epsilon)$, it just states that, under the uniform distribution assumption, the event that $\{P(T_1, \epsilon) < P(T_2, \epsilon)\}$ is a much more likely event than its complement: $\{P(T_2, \epsilon) < P(T_1, \epsilon)\}$.

This finishes the review of the argument for small decision trees. In the following sections we present our criticism of the argument. First, the assumptions are questioned.

4 Questionable assumptions

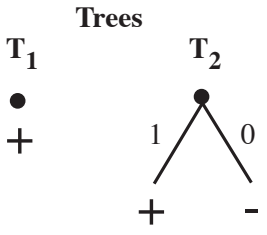
The argument for small decision trees is based on the assumption that P_1 and P_2 are uniformly distributed below their respective bounds. This assumption that p_{bc} is uniformly distributed below the bound is not true in general. It would require that the bound be tight (i.e. that the upper end of the uniform distribution be at the bound) and that the distribution of p_{bc} be uniform (wherever the distribution is located). In Section 4.1 we discuss the events that generate the distribution, in Section 4.2 we discuss the assumption that the bound on p_{bc} is tight, and in Section 4.3 we discuss the assumption that this probability is uniformly distributed.

4.1 What is the event space?

A probability distribution is generated by events. Each event corresponds to one point in the distribution, but not necessarily *vice versa*. When making arguments concerning probabilities, it is crucial to make clear what the event space is. Although Fayyad's argument deals with probability distributions, the event space is not specified. Wolpert (1992) has pointed out the same problem in the PAC learning framework, from which Fayyad's argument is derived. Since the training set and the hypothesis function (determined by the decision tree under consideration) are fixed, the only things left to vary are the labeling f and the probability distribution of feature vectors π (defined in Section 2). Therefore, each event corresponds to a combination of some f_i and some π_j; denote the event by $\langle f_i, \pi_j \rangle$ [7]. The labeling f_i allows us to determine which instances a given tree misclassifies. The distribution π gives a weighting on the instances, allowing us to calculate the error rate for a given tree. Thus, each $\langle f_i, \pi_j \rangle$ corresponds to one point in the distribution of p_{bc}. The probability mass of each point p_{bc} is the sum of the probabilities of the $\langle f_i, \pi_j \rangle$'s that correspond to that point. Therefore, making assumptions regarding the tightness of the bounds or the shape of the distribution of p_{bc} amounts to making assumptions about the distribution of $\langle f_i, \pi_j \rangle$'s.

Because Fayyad's argument does not explicitly make any assumptions about the target concept distribution, the argument should apply to all such distributions. Here we construct a very simple example distribution of $\langle f_i, \pi_j \rangle$'s for which the argument does not hold (see Figure 2). Let us look at

[7] Note that we do not assume that the distribution π of feature vectors is statistically independent of the labeling f.



Training set: { <<1>, +> }

	π*	0.5	0.5	Error rate	
Prob(f_i)		<1>	<0>	T_1	T_2
0	Labeling 1	+	+	0	0.5
1	Labeling 2	+	−	0.5	0

Figure 2: *Example distribution of $\langle f_i, \pi_j \rangle$'s where a larger tree is "likely to have a lower error rate" than a smaller tree.*

a classification problem with two classes ($+$ and $-$), and one binary feature. Let the training set consist of only one instance where the feature has value 1 and the correct classification is $+$. Say that T_1 has only one leaf and classifies all instances as $+$. Say that T_2 has two leaves and classifies instances with feature value 1 as $+$ and instances with feature value 0 as $-$. Let $0 < \epsilon < \frac{1}{2}$. Let us define π^* as follows: $\pi^*(\langle 0 \rangle) = \frac{1}{2}$, $\pi^*(\langle 1 \rangle) = \frac{1}{2}$, and let us define f^* as follows: $f^*(\langle 0 \rangle) = -$, $f^*(\langle 1 \rangle) = +$. Now let us choose a distribution of $\langle f_i, \pi_j \rangle$'s where $\langle f^*, \pi^* \rangle$ has probability one and all other $\langle f_i, \pi_j \rangle$'s have probability zero. Now T_1 has error rate $\frac{1}{2} > \epsilon$, so $P(T_1, \epsilon) = 1$. Similarly, T_2 has error rate $0 < \epsilon$, so $P(T_2, \epsilon) = 0$. Thus, $\text{Prob}\{P(T_1, \epsilon) < P(T_2, \epsilon)\} = 0 < \frac{1}{2}$. This contradicts Fayyad's conclusion that $\text{Prob}\{P(T_1, \epsilon) < P(T_2, \epsilon)\} > \frac{1}{2}$. More complex examples of settings where trees with more leaves outperform trees with fewer leaves have been presented in (Murphy & Pazzani, 1994).

4.2 Tightness of the bounds

In Fayyad's argument, the purpose of the assumption that the bounds b_1 and b_2 are tight is to ensure that $\text{Prob}\{P(T_1, \epsilon) < P(T_2, \epsilon)\} > \frac{1}{2}$. This same property could also be assured in other, less restrictive ways. For example, if $P(T_1, \epsilon)$ and $P(T_2, \epsilon)$ are uniformly distributed, the desired property holds if $P(T_1, \epsilon)$ is distributed right below a *true bound* b_1' and $P(T_2, \epsilon)$ is distributed right below a *true bound* b_2', where $b_1' < b_1$, $b_2' < b_2$, and $b_1' < b_2'$. In order to conclude anything about the actual probabilities being bounded, the true bounds b_1' and b_2' must be in the same order as b_1 and b_2. In general, there is no reason to believe that some arbitrary bounds are in the same order as the true bounds. In fact, in Section 5.1, we show that one can use Fayyad's

argument to obtain arbitrary bounds that give a different ordering.

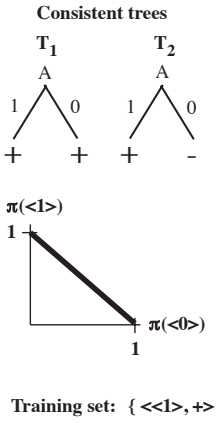
Secondly, when $N \leq -\frac{\log|B(n,\epsilon)|}{\log(1-\epsilon)}$, the bound b will be greater than one. However, because the bounded quantity is a probability, it is already known to be less than or equal to one. Therefore, the bound is not tight in this situation.

4.3 Uniformity of the distribution

Fayyad justifies the uniform distribution assumption by appeal to the "principle of indifference" which states that, absent prior knowledge, the uniform distribution is the most unbiased. This principle must be applied with care because it can easily lead to contradictions. A classic example is Bertrand's paradox (Li & Vitányi, 1993), where assuming a uniform distribution of the values of a quantity, and assuming a uniform distribution of the inverses of these values, leads to a contradiction. This paradox points out that uniformity assumptions applied to related distributions may be inconsistent with each other.

By similar reasoning, it turns out that assuming a uniform distribution of p_{bc} is equivalent to making an assumption of a non-uniform distribution of $\langle f_i, \pi_j \rangle$'s (target concepts).

Consider a situation where there is one binary attribute A, two classes $+$ and $-$, $\epsilon = 0.25$, and the training set consists of one instance: $\langle\langle 1\rangle, +\rangle$. Let us consider binary trees with two leaves. There are four such trees, and two of them are consistent with the training example (see Figure 3). Let us assume a uniform distribution of $\langle f_i, \pi_j \rangle$'s and consider the distribution of p_{bc} that this induces. There are four possible f_i's (labelings), of which two are consistent with the training example and are thus considered. There are infinitely many distributions π satisfying the constraints that $\pi(\langle 0\rangle) + \pi(\langle 1\rangle) = 1$, $0 \leq \pi(\langle 0\rangle) \leq 1$, and $0 \leq \pi(\langle 1\rangle) \leq 1$. Each possible pair $\langle f_i, \pi_j \rangle$ is an event, and all such pairs are considered equally probable. For each event, we would like to know the probability that a consistent tree is bad. For labeling 1, T_1 is never bad. T_2 is bad whenever $\pi(\langle 0\rangle) \geq \epsilon = 0.25$, i.e. with probability 0.75. Put together, there is a 0.25 probability that $p_{bc} = 0$ and a 0.75 probability that $p_{bc} = 0.5$. For labeling 2, T_2 is never bad. T_1 is bad whenever $\pi(\langle 0\rangle) \geq \epsilon = 0.25$, i.e. with probability 0.75. Put together, there is a 0.25 probability that $p_{bc} = 0$ and a 0.75 probability that $p_{bc} = 0.5$. Now, averaging over labelings, there is a 0.25 probability that $p_{bc} = 0$ and a 0.75 probability that $p_{bc} = 0.5$ (see Figure 4). Clearly p_{bc} is not uniformly



Column weights		π(<1>)	π(<0>)		
Prob(f_i)		<1>	<0>	T_1	T_2
0.5	Labeling 1	+	+	Never bad	Bad if π(<0>) ≥ 0.25
0.5	Labeling 2	+	−	Bad if π(<0>) ≥ 0.25	Never bad

Figure 3: *Example showing the non-uniform distribution of p_{bc} when $\langle f_i, \pi_j \rangle$'s are distributed uniformly.*

distributed. Because the uniform distribution of $\langle f_i, \pi_j \rangle$'s implies a non-uniform distribution of p_{bc}'s, it follows that assuming a uniform distribution of p_{bc}'s amounts to assuming a non-uniform distribution of $\langle f_i, \pi_j \rangle$'s!

5 Contradictory results

One could argue that even though the assumptions do not hold exactly, they are a sufficiently good approximation to prove that small decision trees are "likely to be more accurate" than large trees. However, in this section we show that using the same assumptions, but grouping trees by different criteria, we can use the same argument to justify any other inductive bias as well. Because there is no reason to prefer any grouping of trees over any other grouping, the argument is of no use in justifying any inductive bias.

One crucial point to note about the argument for small trees is that the *only* role played by the number of leaves is in partitioning the space of decision trees into subsets. The only property of this partitioning that is used in the proof is that some of the subsets (in this case, those with many leaves) are guaranteed to contain more bad trees than others (in this case, those with few leaves). Thus, we can make the same argument for *any* partitioning of decision tree-space, as long as the partitioning has this property. In this manner, we "justify" several inductive biases which give contradictory prescriptions to the bias toward small trees.

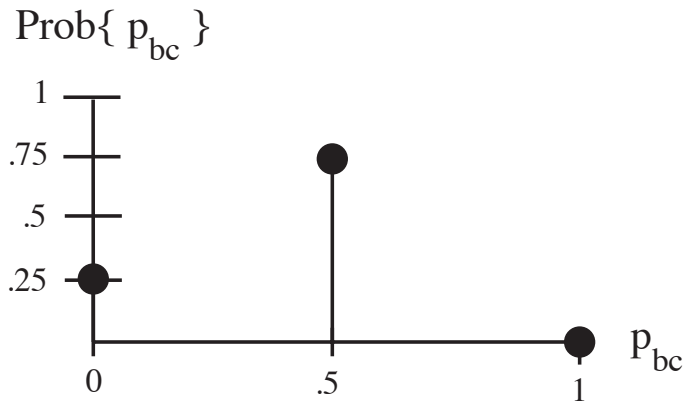


Figure 4: *Example showing the non-uniform distribution of the probability when concepts are distributed uniformly.*

5.1 The more the leaves, the better the tree?

Consider two arbitrary consistent trees T_1 and T_2 with n_1 and n_2 leaves respectively such that $n_2 > n_1$. For any fixed ϵ, $0 < \epsilon < \frac{1}{2}$, [8] assuming that $P(T_1, \epsilon)$ and $P(T_2, \epsilon)$ are uniformly distributed below their bounds, Corollary 3.3.2 tells us to prefer T_1 over T_2. Now let us consider the following alternate partitioning of decision tree-space:

- For all $n \neq n_1, n_2$, $DT'(n) = DT(n)$
- $DT'(n_1) = DT(n_1) \cup \{T_2\} - \{T_1\}$ [9]
- $DT'(n_2) = DT(n_2) \cup \{T_1\} - \{T_2\}$.

Define $B'(n, \epsilon)$ as $\{T \mid T \in DT'(n)$ and Error$(T) > \epsilon\}$. Now, $|B'(n_1, \epsilon)| \leq |B(n_1, \epsilon)| + 1$, and $|B'(n_2, \epsilon)| > |B(n_2, \epsilon)| - 1$. Assuming $|B(n_1, \epsilon)| > 1$, by Corollary 3.3.6 $|B(n_2, \epsilon)| > 4|B(n_1, \epsilon)| \geq 4$. Hence, $\frac{|B'(n_2,\epsilon)|}{|B'(n_1,\epsilon)|} \geq \frac{4-1}{1+1} \geq \frac{3}{2}$. This is sufficient to ensure that Prob$\{P(T_1, \epsilon) < P(T_2, \epsilon)\} > \frac{1}{2}$, and using Fayyad's terminology, we can say that T_2 is "likely to have a lower error rate" than T_1. Thus, using Fayyad's argument and changing only the partitioning

[8] In Fayyad's argument, $\epsilon \in (0, 1)$. This is slightly erroneous because for $\epsilon > \frac{1}{2}$ and small n, it is possible that $|B(n, \epsilon)| = 0$ which would invalidate Lemma 3.3.6. This problem can be fixed by choosing $\epsilon \in (0, \frac{1}{2})$.

[9] Although this partitioning may not be as intuitive as the original partitioning by number of leaves, it is equally valid for the purposes of the proof.

of decision tree-space, we have arrived at a justification of a bias directly opposed to his: a preference for the tree with the *greater* number of leaves!

5.2 The lexicographically earlier the root test name, the better the tree?

We can also use a partitioning that is unrelated to size. Consider a situation in which the number of attributes k is greater than 2. First, number the attributes from 1 to k by lexicographic order of their names. Let $RT(n)$ be the set of binary decision trees with root test of attribute n for $1 \leq n \leq k$. Note that each of these groups contains the same number of trees of each size, and thus the average tree size is the same in each group. Let $RT(0)$ be the set of binary decision trees with no root test (i.e., with one leaf). We can construct a partitioning of decision-tree space that has the property that some of the subsets are guaranteed to contain more bad trees than others as follows.

- For $1 \leq m \leq \lfloor \log_3 (k+1) \rfloor - 1$, let $RT'(m) = \bigcup_{n=\frac{3^m-1+1}{2}}^{\frac{3^m-1}{2}} RT(n)$.
- For $m = \lfloor \log_3 (k+1) \rfloor$, let $RT'(m) = RT(0) \cup \left[\bigcup_{n=\frac{3^m-1+1}{2}}^{k} RT(n) \right]$.

For example, in a situation with 50 attributes, $RT'(1) = RT(1)$, $RT'(2) = RT(2) \cup RT(3) \cup RT(4)$, and $RT'(3) = RT(5) \cup \ldots \cup RT(13)$, and $RT'(4) = RT(0) \cup RT(14) \cup \ldots \cup RT(50)$.

Let $RB(m, \epsilon)$ denote the set of bad trees in $RT'(m)$. In appendix A we show that, for $0 < \epsilon < \frac{1}{2}$, at least half of the trees in each group $RT'(m)$ are bad. Thus, $\frac{|RT'(m)|}{2} \leq |RB(m, \epsilon)| \leq |RT'(m)|$. Similarly, $\frac{|RT'(m+1)|}{2} \leq |RB(m+1) \leq |RT'(m+1)|$. Because $|RT'(m+1)| \geq 3 \cdot |RT'(m)|$, it follows that $|RT'(m)| < \frac{|RT'(m+1)|}{2}$. Therefore, $|RB(m+1, \epsilon)| > |RB(m, \epsilon)|$.

Now consider two arbitrary trees T_1 and T_2 where T_1 has more leaves than T_2, each tree has more than one leaf, and T_1's root test is sufficiently lexicographically earlier than T_2's so that $T_1 \in RT'(i)$, $T_2 \in RT'(j)$, and $i < j$. The fact that $|RB(m+1, \epsilon)| > |RB(m, \epsilon)|$ ensures that Prob$\{P(T_2, \epsilon) < P(T_1, \epsilon)\} > \frac{1}{2}$. T_1 is "likely to have a lower error rate" than T_2. Thus, again using Fayyad's argument and changing only the partitioning of decision tree-space, we have arrived at a justification of a preference for the tree with the lexicographically earlier root test, but the *greater* number of leaves!

5.3 Any arbitrarily chosen tree is better than any other?

Taken to its logical extreme, this same argument can be used to justify a preference for *any* given tree over any other. Consider any two consistent trees T_1 and T_2, each of any size. Consider T_1 to be a member of the set $\{T_1\}$ and T_2 to be a member of U, the set of all decision trees for this learning situation. Define $B(U, \epsilon)$ as $\{T \mid T \in U \text{ and } \text{Error}(T) > \epsilon\}$ and $B(\{T_1\}, \epsilon)$ as the set $\{T_1\}$ if $\text{Error}(T_1) > \epsilon$ and $\emptyset$ otherwise. For any ϵ such that $0 < \epsilon < \frac{1}{2}$, $|B(U, \epsilon)| > 1 \geq |B(\{T_1\}, \epsilon)|$. Thus the bound $b_U \geq b_{\{T_1\}}$. Using the assumption that the true probabilities are uniformly distributed below these bounds, $\text{Prob}\{P(T_1, \epsilon) < P(T_2, \epsilon)\} > \frac{1}{2}$. Thus we have "proven" that any arbitrarily chosen tree is "likely to have a lower error rate" than any other arbitrarily chosen one.

We have shown that by changing only the partitioning of tree-space while leaving the rest of the argument intact, we can justify any bias. Since we are given no reason to prefer the partitioning of tree-space by size over any other, it is clear that the original result cannot be used to justify a bias toward decision trees with fewer leaves.

6 Counting trees vs. counting hypotheses

Fayyad's argument for small trees differs from other examinations of the simplicity bias (Pearl, 1978) in that representations (in this case, decision trees) for hypotheses are counted rather than hypotheses themselves. One might suspect that the problems we demonstrate with Fayyad's analysis could be remedied by modifying the argument to count hypotheses, or equivalently, logically non-equivalent trees. We now show that this is not the case.

The fact that logically equivalent trees are counted is crucial to the proof of Lemma 3.3.6, which states that the number of bad trees with n leaves grows with n. The proof uses the fact that for every tree T with n leaves, there are four trees logically equivalent to T with $n + 1$ leaves. These are constructed by adding tests to T for which both children are leaves of the same class (call such tests *spurious*). The growth in the number of bad trees constructed in this manner is exponential, and since this number is a factor in the formulation of the bounds, these grow exponentially as well. Decision tree induction algorithms typically do not produce trees containing spurious

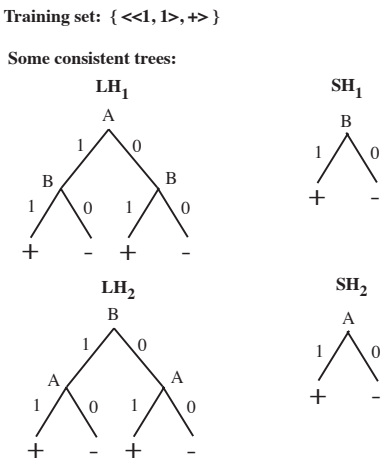


Figure 5: *Example trees where the argument for small decision trees leads to contradictions.*

tests. [10]

In the next two sections we present examples where the argument for small decision trees results in paradoxical prescriptions of hypotheses. In Section 6.3 we show that the argument for small decision trees cannot be corrected even by analyzing hypotheses rather than trees.

6.1 Preferring a hypothesis to itself

The fact that logically equivalent trees are counted rather than hypotheses allows the argument to give contradictory results in certain situations.

Consider the trees SH_1 and LH_1 shown in Figure 5. Both trees represent the same hypothesis. However, because SH_1 has fewer leaves than LH_1, Fayyad's argument states that SH_1 is "likely to be more accurate" than SH_2, even though they have the same error rate!

[10]Another source of logically equivalent trees is repeated tests, i.e. testing the same attribute more than once on a path from the root to a leaf. Repeated tests necessarily occur in trees with more than 2^a leaves, where a is the number of binary attributes. Decision tree algorithms typically do not produce trees containing repeated tests. Even if one disallows spurious and repeated tests, there may be many logically equivalent trees (even with the same number of leaves).

6.2 Contradictory prescriptions of hypotheses

Consider the four trees LH_1, SH_1, LH_2, and SH_2 shown in Figure 5. LH_1 and SH_1 represent the same hypothesis, call it $H1$. Similarly, LH_2 and SH_2 represent the same hypothesis, call it $H2$. Now the argument for small decision trees prescribes preferring SH_1 over LH_2, i.e. preferring $H1$ over $H2$. But the argument for small decision trees also prescribes preferring SH_2 over LH_1, i.e. preferring $H2$ over $H1$, which is contradictory.

The next section shows that the argument for small decision trees cannot be corrected even by analyzing hypotheses rather than trees.

6.3 Will counting hypotheses fix the argument?

To change the argument for small decision trees to count logically non-equivalent trees, it would only be necessary to show that the number of bad non-equivalent trees grows in n. It is not necessary to show that growth is exponential in order to replace Lemma 3.3.6. The growth of the number of such trees implies that $b_2 > b_1$. Now we can again use the argument of Section 5.3 (which did not use the assumption of exponential growth) to reach a contradictory conclusion. Thus the objections do not depend upon the counting of logically equivalent trees.

7 Conclusions

It should be emphasized that these results do not deny Occam's Razor. Neither do they deny the possibility that for data sets occurring in the real world, an inductive bias toward decision trees with fewer leaves may be appropriate. They do show, however, that the argument intended to prove this claim in general is invalid. We demonstrated this by first questioning the assumptions it relies upon, and then using the same argument with different partitionings of decision tree-space to deduce contradictory conclusions. This shows that the argument cannot be used to prescribe any inductive bias. More generally, we showed that no bias can be justified without making *a priori* assumptions about the distribution of target concepts. This refutes the second (rarity of simple concepts) justification of Occam's Razor.

Our analysis also points out several pitfalls to avoid when constructing arguments about induction algorithms. First, we point out that when making arguments about probability distributions, confusion may arise if one

does not make explicit the event spaces over which these distributions are defined. Second, we show that when modeling a state of no knowledge with a uniform probability distribution, one must be very aware of the implications of such an assumption. In many cases, uniformity assumptions of two quantities are inconsistent. Thus, one should keep in mind that making a uniformity assumption for some quantity may amount to making an assumption of non-uniformity for related quantities about which we also have no knowledge. Third, one cannot conclude that two quantities are likely to be in a particular order based on upper bounds of unknown tightness. Finally, in making counting arguments, one should be aware of the distinction between hypotheses and representations of hypotheses, e.g. trees. In a given representation schemes, a hypothesis may have several representations, and different hypotheses may have different numbers of representations. Thus, counting arguments regarding representations will give results different from those regarding hypotheses. Nevertheless, even if one considers hypotheses rather than representations, one cannot justify any inductive bias without making a priori assumptions about the target concept distribution.

References

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters, 24*, 377-380.

Fayyad, U. M., & Irani, K. B. (1990). What should be minimized in a decision tree? *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 749-754). Boston, MA: Morgan Kaufmann.

Fayyad, U. M. (1991). *On the induction of decision trees for multiple concept learning.* Doctoral dissertation, Computer Science and Engineering, University of Michigan.

Li, M., & Vitányi, P. M. B. (1993). *An introduction to Kolmogorov complexity and its applications.* New York: Springer-Verlag.

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.

Pearl, J. (1978). On the connection between the complexity and credibility of inferred models. *International Journal of General Systems, 4*, 116-126.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* Morgan Kaufmann.

Rao, R.B., Gordon, D. , & Spears, W. (1995). For every generalization action, is there really an equal and opposite reaction? Analysis of the Conservation Law for Generalization Performance. *Machine Learning: Proceedings of the Twelfth International Conference* (pp. 115-121). Tahoe City, CA: Morgan Kaufmann.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach.* Englewood Cliffs, NJ: Prentice-Hall.

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning, 10*, 153-178.

Schaffer, C. (1994). A conservation law for generalization performance. *Machine Learning: Proceedings of the Eleventh International Conference* (pp. 259-265). New Brunswick, NJ: Morgan Kaufmann.

Webb, G. I. (1995). *An experimental disproof of Occam's razor*, (TR-C95-07), Geelong, Australia: Deakin University, School of Computing and Mathematics.

Wolpert, D. H. (1992). On the connection between in-sample testing and generalization error. *Complex Systems, 6*, 47-94.

Wolpert, D. H. (1993). *On overfitting avoidance as bias*, (SFI TR 93-03-016), Santa Fe, NM: The Santa Fe Institute.

Wolpert, D. H. (1994a). On the Bayesian "Occam factors" argument for Occam's Razor. In Hanson, Drastal & Rivest (Eds.), *Computational learning theory and natural learning systems: Volume III Natural learning systems.* Cambridge, MA: M.I.T. Press.

Wolpert, D. H. (1994b). *The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework*, (SFI TR 94-03-123), Santa Fe, NM: The Santa Fe Institute.

Wolpert, D. H. (1995). *Off-training set error and a priori distinctions between learning algorithms*, (SFI TR 95-01-003), Santa Fe, NM: The Santa Fe Institute.

A Proof: At least half of the trees in $RT'(m)$ are bad

In this appendix we prove that, for $0 < \epsilon < \frac{1}{2}$, at least half of the trees in each group $RT'(m)$ are bad. This fact was used in Section 5.2. Let T be an arbitrary tree in $RT'(m)$ that is not bad. For each leaf i of T, $1 \leq i \leq n$, let V_i be the set of feature vectors $\vec{v}$ that fall into i. Associated with each leaf i is an error rate e_i determined by V_i, the weight π of each $\vec{v} \in V_i$, the correct classifications of these $\vec{v}$'s, and the class label of i. Specifically, $e_i = \frac{\sum_{\vec{v} \in V_i} \pi(\vec{v}) \cdot \text{MISCLASSIFIES}(i,\vec{v})}{\sum_{\vec{v} \in V_i} \pi(\vec{v})}$, where MISCLASSIFIES$(i, \vec{v})$ is 1 if leaf i misclassifies feature vector $\vec{v}$, and 0 otherwise.

From the good tree $T \in RT'(m)$ we can construct a bad tree T' as follows. Consider the classes of T as a vector. Now T' is the same as T except that the classes are rotated by one position (say to the left). For example, in a problem with classes 1, 2, and 3, every leaf of T that has class 1 will have class 2 in T', every leaf that has class 2 in T will have class 3 in T' and every leaf that has class 3 in T will have class 1 in T'. Because the root test of T' is the same as that of T, clearly $T' \in RT'(m)$. The error rates of the leaves of T' will be at least $(1-e_1)\ldots(1-e_n)$ because T' will misclassify all the feature vectors that T classified correctly. Let E_T be the overall error rate of T and let $E_{T'}$ be the overall error rate of T'. These are computed from the error rates of the leaves by weighting each leaf by the sum of the π's of the feature vectors that fall into the leaf. For example, $E_T = \sum_{i=1}^{n} \left[\sum_{\vec{v} \in V_i} \pi(\vec{v})\right] e_i$. It follows from the above that $E_{T'} \geq \sum_{i=1}^{n} \left[\sum_{\vec{v} \in V_i} \pi(\vec{v})\right] (1 - e_i) = 1 - \sum_{i=1}^{n} \left[\sum_{\vec{v} \in V_i} \pi(\vec{v})\right] e_i = 1 - E_T > 1 - \epsilon > \frac{1}{2} > \epsilon$, and thus T' is bad.

Note that T is the only tree that transforms into T' using our rotation mechanism. Therefore every good tree in $RT'(m)$ has a *unique* bad tree in $RT'(m)$. Thus at least half of the trees in $RT'(m)$ are bad. □