Machine Learning, 28, 257–291 (1997)

An Exact Probability Metric for Decision Tree Splitting and Stopping

J. KENT MARTIN jmartin@ics.uci.edu
Department of Information and Computer Science, University of California, Irvine, Irvine, CA 92692

Editor: Doug Fisher

Abstract. ID3's information gain heuristic is well-known to be biased towards multi-valued attributes. This bias is only partially compensated for by C4.5's gain ratio. Several alternatives have been proposed and are examined here (distance, orthogonality, a Beta function, and two chi-squared tests). All of these metrics are biased towards splits with smaller branches, where low-entropy splits are likely to occur by chance. Both classical and Bayesian statistics lead to the multiple hypergeometric distribution as the exact posterior probability of the null hypothesis that the class distribution is independent of the split. Both gain and the chi-squared tests arise in asymptotic approximations to the hypergeometric, with similar criteria for their admissibility. Previous failures of pre-pruning are traced in large part to coupling these biased approximations with one another or with arbitrary thresholds; problems which are overcome by the hypergeometric. The choice of split-selection metric typically has little effect on accuracy, but can profoundly affect complexity and the effectiveness and efficiency of pruning. Empirical results show that hypergeometric pre-pruning should be done in most cases, as trees pruned in this way are simpler and more efficient, and typically no less accurate than unpruned or post-pruned trees.

Keywords: pre-pruning, post-pruning, hypergeometric, Fisher's Exact Test

1. Introduction and Background

Top-Down Induction of Decision Trees, TDIDT (Quinlan, 1986), is a family of algorithms for inferring classification rules (in the form of a decision tree) from a set of examples. TDIDT makes a greedy choice of a candidate split (decision node) for a data set and recursively partitions each of its subsets. Splitting terminates if all members of a subset are in the same class or the set of candidate splits is empty.

Some algorithms, e.g., ID3 (Quinlan, 1986), have included criteria to stop splitting when the incremental improvement is deemed insignificant. These stopping criteria are sometimes collectively referred to as pre-pruning criteria. Other algorithms have added recursive procedures for post-pruning (replacing a split with a terminal node). Some procedures described as post-pruning go beyond mere pruning by replacing a split with some other split, typically with a child of the replaced node, as in C4.5 (Quinlan, 1993).

Note that there are more than 10^{13} ways to partition a set containing only 20 items. Practical algorithms can explore only a small portion of such a vast space. Greedy hill-climbing is a general strategy for reducing search, but here it must operate in the context of exploring only a tiny subset of the possible splits. TDIDT builds complex trees by recursive refinement of simpler trees, and it explores only simple splits at each decision node. At each decision node, split selection is addressed as two separate but interdependent subproblems:

1. choosing a set of candidate splits
2. selecting a split (or, perhaps, none of them, if pre-pruning is used)

The earliest TDIDT algorithms such as CART (Breiman, Friedman, Olshen, & Stone, 1984), ID3 (Quinlan, 1986), and C4.5 (Quinlan, 1993) restricted the candidates to splits on the values of a single attribute having a small number of distinct values and only binary splits for continuous attributes. More recent algorithms extend the candidate space in various ways, including lookahead (e.g., Elder, 1995; Murthy & Salzberg, 1995; Quinlan & Cameron-Jones, 1995), multi-way splits for continuous attributes (e.g., Fayyad & Irani, 1992b, 1993; Fulton, Kasif, & Salzberg, 1995), combinations of two discrete attributes (Murphy & Pazzani, 1991), and linear combinations of continuous attributes (e.g., John, 1995; Murthy, Kasif, Salzberg, & Beigel, 1993; Park & Sklansky, 1990).

Choosing a split from among the candidates takes place in the context of, and may interact strongly with, the choice of a set of candidates. At each decision point, both of these processes take place in the context of all of the choices made at higher levels in the tree. The interactions between the two phases of split selection, between the two phases and the context created by earlier choices, and between the two phases, the context, and the greedy search strategy create a very complex environment; one in which it is very difficult to determine what the impact would be of changing some aspect of a procedure. It is equally difficult to determine which aspects of a procedure may be responsible for poor or good performance on any particular problem.

An important facet of the changing context for split selection is that the mean subset size decreases with the depth of the decision node. A fundamental principle of inference is that the degree of confidence with which one is able to choose between alternatives is directly related to the number of examples. There is thus a strong tendency for inferences made near the leaves of a TDIDT decision tree to be less reliable than those made near the root.

The strong interaction between the choice of the set of candidates and the selection among candidates is exemplified by pre-pruning the exclusive-or (XOR) of two Boolean attributes. Neither attribute, taken alone, appears to have any utility in separating the classes; yet the combination of the two will completely separate the classes. If only single-attribute splits are allowed, and pre-pruning based on apparent local utility is used, the resulting tree will have a single leaf of only 50% accuracy (assuming equally frequent classes).

This example is often cited as an argument against pre-pruning. The difficulty is actually the result of the interaction of pre-pruning and allowing only single-attribute splits, and one could easily argue against a very restricted choice of a candidate set. For any given set of candidates, pre-pruning will tend to preclude discovering a significantly better tree for problems where the correct concept definition contains compound features similar to XOR[1]. There are, however, at least two approaches which might lead to discovering a better decision tree. One approach is not to pre-prune but, rather, to post-prune as appropriate. The other approach is to expand the set of candidates. Both of these approaches increase the learning time — if both ultimately discover equivalent trees, we should prefer the approach entailing the least additional work.

Though we have mentioned expanding the candidate set as a possible means of dealing with XOR and other difficulties arising from exploring only single-attribute splits, and will touch on it again at the end of the paper, this paper does not explore this phase of split selection experimentally. The main focus of this paper is on the second phase of split selection, the use of heuristic functions to select a split from among a set of candidates. Another objective is to explore causes (other than the XOR difficulty) of the poor performance of

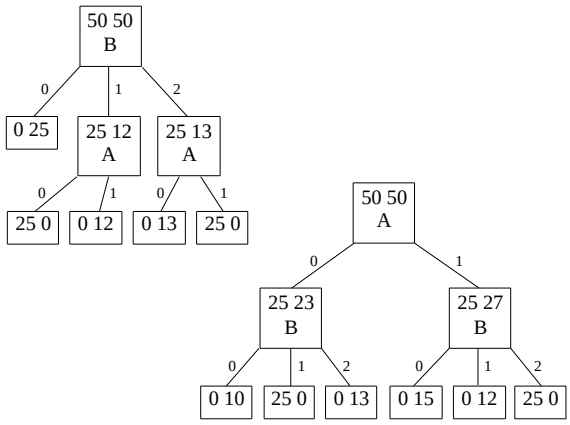

Figure 1. Alternative Splits

pre-pruning in early empirical studies (Breiman, et al., 1984; Fisher & Schlimmer, 1988; Quinlan, 1986).

Evaluation criteria for split selection involve tradeoffs of accuracy and complexity. There is no single measure which combines these appropriately for every application. Measures of complexity include the number of leaves and their average depth (weighted according to the sample fraction covered by each leaf), and the time complexity of the algorithm. The following terms will be used to distinguish between these: complexity $\equiv$ number of leaves; efficiency $\equiv$ average depth (expected classification cost); and practicality $\equiv$ tree building, pruning, and cross-validation time.

In referring to classifier accuracy, an important distinction is made between the *population* (all of the instances in the problem's domain) and the *sample* (the classifier's training/testing data). The dominant goal is usually to infer trees where the population instances covered by each leaf are, as nearly as possible, members of the same class. If each leaf is labeled with some predicted class, the accuracy of the leaf is defined as the percentage of the covered population instances for which the class is correctly predicted. The accuracy of the tree is defined as the average accuracy of the leaves, weighted according to the fraction of the population covered by each leaf. In most cases, accuracy can only be estimated, and it is important to report a variance or confidence interval as well as the point estimate. Typically, cross-validation (Breiman, et al., 1984) is used to estimate accuracy.

2. Impact of Different Choices Among Candidate Splits

Figure 1 shows two different decision trees for the same data set, choosing a different split at the root. In this case, the accuracy of the two trees is the same (100%, if this is the entire population), but one of the trees is more complex and less efficient than the other. For this problem, the set of candidate splits is sufficient to fully separate the classes, and each of the

candidate splits is necessary. The choice of one split over another is a matter of complexity and efficiency, rather than of accuracy.

A set of candidate splits might be insufficient because of missing data, noise, or some hidden feature. After introducing noise[2] into the population of Figure 1, the average results of splitting on A first versus splitting on B first are shown in Table 1a (averaged over 100 independent samples randomly drawn from this noisy population, each sample of size 100). Here also, the difference between the alternative split orderings is a matter of complexity and efficiency, not accuracy.

Returning to the noise-free population of Figure 1, if we add an irrelevant variable[3] X and split on A first then B, or on B first then A, we get the same trees shown in Figure 1 (the first two lines in Table 1b) and attribute X will not be used. The effects of splitting on attribute X first, or splitting on X between the splits on A and B are also shown in Table 1b. Again, the difference between the alternative split orderings is a matter of complexity and efficiency, not accuracy.

Rather than the irrelevant attribute X, suppose that we added a binary attribute Y, which is equal to the classification 99% of the time, but opposite to the class 1% of the time,

Table 1. Effects of Split Order

a. Effects of Noise

	Error Rate	No. of Nodes	No. of Leaves	Wtd. Avg. Depth
A first	2.5%	9	6	2
B first	2.5%	8.7	5.4	1.8

b. Effects of An Irrelevant/Redundant Attribute

	Error Rate	No. of Nodes	No. of Leaves	Wtd. Avg. Depth
A, B	0	9	6	2
B, A	0	8	5	1.8
X, A, B	0	19	12	3
X, B, A	0	17	10	2.8
$X, AB/BA$	0	18	11	2.9
$X, BA/AB$	0	18	11	2.9
A, X, B	0	19	12	3
B, X, A	0	16	9	2.5

c. Combined Effects of Noise and Redundancy

	37.5% Noise Level				10% Noise Level			
	Error %	No. of Nodes	No. of Leaves	Avg. Depth	Error %	No. of Nodes	No. of Leaves	Avg. Depth
A, B, Z	2.6	13.3	8.2	2.4	2.8	13.3	8.2	2.4
B, A, Z	2.8	13.0	7.5	2.8	2.8	13.0	7.5	2.2
A, Z, B	2.8	19.0	12.0	3.0	4.0	18.1	11.4	3.0
B, Z, A	3.0	17.2	9.6	2.6	3.9	16.3	9.2	2.5
Z, A, B	2.8	19.0	12.0	3.0	4.3	18.1	11.4	3.0
Z, B, A	2.8	17.7	10.3	2.8	3.9	16.6	9.8	2.7
$Z, AB/BA$	2.9	18.3	11.1	2.9	3.9	17.3	10.6	2.8
$Z, BA/AB$	2.8	18.3	11.1	2.9	3.9	17.7	10.8	2.9

randomly. Splitting on this attribute alone would give 99% accuracy, so it is clearly relevant, but redundant (since the pair of attributes A and B give 100% accuracy). The results for splitting on A, B, and Y in different orders are identical to those given in Table 1b for A, B, and X.

As a final example in this vein, consider the effects of adding both noise and irrelevant or redundant attributes. Add a third attribute Z to the noisy population of Table 1a, one that is just a noisier version of the original attribute underlying the noisy attribute A. If the level of noise in this attribute is varied, its behavior ranges from being irrelevant at a 50% noise level to being redundant as its noise level approaches that of A (1%). (Note that, even at 1% noise, attribute Z taken alone is less predictive of the class than was the redundant attribute Y in the previous paragraph). The effects of splitting on A, B, and Z in various orders are shown in Table 1c. When attribute Z is more nearly irrelevant (37.5% noise), the order of the attribute splits is largely a matter of complexity and efficiency, rather than accuracy. As Z becomes more relevant, but redundant (10% noise), splitting on attribute Z before or between the splits on attributes A and B has a significant negative impact on accuracy as well as on efficiency and complexity.

From the foregoing examples, for unpruned trees, the order in which various splits are made is largely a matter of complexity and efficiency, rather than of accuracy. Accuracy may be significantly affected when attributes are noisy and strongly correlated (i.e., redundant). Insofar as the accuracy of unpruned trees is concerned, the ordering of the splits is not a significant factor in most cases. This is one of the factors underlying the frequent observations (e.g., Breiman, et al., 1984; Fayyad & Irani, 1992a) that various heuristic functions for choosing among candidate splits are largely interchangeable.

It is important to note that if significant differences in accuracy occur, the difference in accuracy would typically be of overriding importance. When the accuracies of various trees are equivalent, however, there is certainly a preference for simpler and more efficient trees. The differences in complexity and efficiency in the examples given above, and indeed in most of the applications in the UCI data depository (Murphy & Aha, 1995), are relatively minor. For more complex applications involving scores of attributes and thousands of instances, these effects will be compounded, and may have a much greater impact. It should also be noted that all of these differences in accuracy and complexity are being explored in the context of having severely restricted the set of candidate splits for the sole purpose of reducing an intractable problem to manageable proportions. Differences in complexity and efficiency may be greatly magnified as the set of candidate splits is expanded.

Liu and White (1994) discuss the importance of discriminating between attributes which are truly 'informative' and those which are not. The examples in Figure 1 and Table 1 do not consider the possible effects of pruning. Consider the effects of pruning in Table 1c. From Table 1a, we know that splitting on the noisy attributes A and B alone (and ignoring attribute Z) achieves an error rate of 2.5%. Subsequently splitting on attribute Z does not improve accuracy (it appears to be harmful), and adds significantly to the complexity of the trees. There is strong evidence that the final split on attribute Z overfits the sample data and should be pruned.

When the split on attribute Z does not come last, then simple pruning will not correct the overfitting (it would, in fact, be very harmful). The pruning strategy used in Quinlan's (1993) C4.5 algorithm, replacing the split with one of its children and merging instances

from the other children, would be beneficial here. This kind of tree surgery is by far less practical than simple pruning (Martin & Hirschberg, 1996a), and could be avoided if the candidate selection heuristic chose to split on Z last. The presence of this kind of tree surgery in an algorithm suggests that the algorithm's heuristic does not choose splits in the best order from the point of view of efficient pruning.

This strong dependence of the effectiveness and efficiency of either pre- or post-pruning on the order in which the splits are made appears to have been overlooked in previous machine-learning studies comparing various split selection metrics, e.g., CART (Breiman, et al., 1984) and Fayyad and Irani (1992a), which have consistently found various metrics to be largely interchangeable with regard to the resulting tree's accuracy. The examples given here indicate that different split orderings can profoundly affect how effective a simple pre- or post-pruning algorithm will be, and whether more elaborate and expensive algorithms such as C4.5's can be avoided.

Thus, we suggest that the following three criteria should be considered in choosing a heuristic:

1. It should prefer splits that most improve accuracy and avoid those which are harmful.
2. For equivalent accuracies, it should prefer splits leading to simpler and more efficient trees.
3. It should order the splits so as to permit effective simple pruning.

3. Functions for Selection Among Candidates

A natural approach is to label each of the split subsets according to their largest class and choose the split which has the fewest errors. There are several problems with this approach, see, e.g., CART (Breiman, et al., 1984), the most telling being that it simply has not worked out well empirically.

Various other measures of split utility have been proposed. Virtually all of these measures agree as to the extreme points, i.e., that a split in which the classes' proportions are the same in every subset (and, thus, the same as in the parent set) has no utility, and a split in which each subset is pure (each contains only one class) has maximum utility. Intermediate cases may be ranked differently by the various measures. Most of the measures fall into one of the following categories:

1. Measures of the difference between the parent and the split subsets on some function of the class proportions (such as entropy). These measures emphasize the purity of the subsets, and CART (Breiman, et al., 1984) terms these *impurity* functions.
2. Measures of the difference between the split subsets on some function of the class proportions (typically a distance or an angle). These measures emphasize the *disparity* of the subsets.
3. Statistical measures of independence (typically a χ^2 test) between the class proportions and the split subsets. These measures emphasize the weight of the evidence, the *reliability* of class predictions based on subset membership.

Suppose, for instance, that we randomly choose 64 items from a population and observe that 24 items are classified positive and 40 negative. If we then observe that 1 of the positive items is red and all other items (positive or negative) are blue, how reliable is an inference that all red items are positive, or even a weaker inference that red items tend to have a different class than blue ones?

Fayyad and Irani (1992a) cite several studies showing that various impurity measures are largely interchangeable, i.e., that they result in very similar decision trees, and CART (Breiman, et al., 1984) finds that the final (unpruned) tree's properties are largely insensitive to the choice of a splitting rule (utility measure).

A great variety of differing terminology, representations, and notation for splits is used in the machine learning literature. To facilitate comparisons of the different metrics, only one representation and notation is used here. A convenient representation for splits is a contingency, or cross-classification, table:

	sub-1	$\cdots$	sub-V	Total
cat-1	f_{11}	$\cdots$	f_{1V}	n_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
cat-C	f_{C1}	$\cdots$	f_{CV}	n_C
Total	m_1	$\cdots$	m_V	N

C	is the number of categories
V	is the number of subsets in the split
m_v	is the no. of instances in subset v
f_{cv}	is the no. of those which are in class c
N	is the total no. in the parent
n_c	is the total no. in class c

3.1. Approximate Functions for Selection Among Candidates

Variants of the information gain impurity heuristic used in ID3 (Quinlan, 1986) have become the *de facto* standard metrics for TDIDT split selection. Information gain is the difference (decrease) between the entropy at the parent and the weighted average entropy of the subsets.

$$\begin{aligned} \text{gain} \;=\; & \left(\sum_{c=1}^{C} \left[-\left(\frac{n_c}{N}\right) \log_2 \left(\frac{n_c}{N}\right) \right] \right) \\ - \; & \left(\sum_{v=1}^{V} \left(\frac{m_v}{N}\right) \sum_{c=1}^{C} \left[-\left(\frac{f_{cv}}{m_v}\right) \log_2 \left(\frac{f_{cv}}{m_v}\right) \right] \right) \end{aligned} \tag{1}$$

The gain ratio function used in C4.5 (Quinlan, 1993) partially compensates for the known bias of gain towards splits having more subsets (larger V).

$$\text{gain ratio} = \text{gain} \Bigg/ \sum_{v=1}^{V} \left[-\left(\frac{m_v}{N}\right) \log_2 \left(\frac{m_v}{N}\right) \right] \tag{2}$$

Lopez de Mantaras (1991) proposes a different normalization, a distance metric $(1 - d)$

$$1 - d = \text{gain} \Bigg/ \sum_{v=1}^{V} \sum_{c=1}^{C} \left[-\left(\frac{f_{cv}}{N}\right) \log_2 \left(\frac{f_{cv}}{N}\right) \right] \tag{3}$$

Fayyad and Irani (1992a) give an orthogonality (angular disparity) metric for binary attributes

$$ORT = 1 - \left(\sum_{c=1}^{C} f_{c1} \cdot f_{c2}\right) \Big/ \left[\left(\sum_{c=1}^{C} f_{c1}^2\right)\left(\sum_{c=1}^{C} f_{c2}^2\right)\right]^{1/2} \tag{4}$$

Unlike gain ratio and $1 - d$, ORT is not a function of gain.

Buntine (1990) derives a Beta-function splitting rule

$$e^{-W(\alpha)} = \frac{\Gamma(C\alpha)^V}{\Gamma(\alpha)^{CV}} \prod_{v=1}^{V} \frac{\prod_{c=1}^{C} \Gamma(f_{cv} + \alpha)}{\Gamma(m_v + C\alpha)} \tag{5}$$

The parameter, α, is user-specified (typically $\alpha = 0.5$ or $\alpha = 1$), and describes the assumed prior distribution of the contingency table cells. Information gain appears as part of an asymptotic approximation to this function[4].

In addition to the above heuristics from the machine learning literature, the analysis of categorical data has long been studied by statisticians. The Chi-squared statistic (Agresti, 1990)

$$X^2 = \sum_{c=1}^{C}\sum_{v=1}^{V} \frac{(f_{cv} - e_{cv})^2}{e_{cv}}, \quad \text{where } e_{cv} = (n_c\, m_v/N) \tag{6}$$

is distributed *approximately* as χ^2 with $(C-1) \times (V-1)$ degrees of freedom[5]. The quantities e_{cv} are the expected values of the frequencies f_{cv} under the *null hypothesis* that the class frequencies are independent of the split. This test is admissible[6] only when *Cochran's criteria* (Cochran, 1952) are met (all of the e_{cv} are greater than 1 and no more than 20% are less than 5). We note that because of the recursive partitioning inherent in TDIDT, Cochran's criteria **cannot** be satisfied by all splits in a tree of depth $> \log_2(N_0/5)$ (where N_0 is the size of the tree's training set), and the criteria are unlikely to be satisfied even in shallower trees with unbalanced splits.

The Likelihood-Ratio Chi-squared statistic (Agresti, 1990)

$$G^2 = 2 \sum_{c=1}^{C}\sum_{v=1}^{V} f_{cv}\ \ln\left(\frac{f_{cv}}{e_{cv}}\right) \tag{7}$$

is also *approximately* χ^2 with $(C-1) \times (V-1)$ degrees of freedom. The convergence of G^2 is slower than X^2, and the χ^2 approximation for G^2 is usually poor whenever $N < 5\,CV$ (Agresti, 1990), as was also the case for X^2.

Replacing e_{cv} by $(n_c\ m_v/N)$ in Equation 7 and rearranging the terms leads to $G^2 = 2\ln(2)N$ gain. In the arguments supporting adoption of information gain, minimum description length (MDL), and general entropy-based heuristics, the product of the parent set size and the information gain from splitting ($N \times$ gain) is approximately the number of bits by which the split would compress a description of the data. The gain approximation is closely related to conventional maximum likelihood analysis, and message compression has a limiting χ^2 distribution that converges less quickly than the more familiar X^2 test.

Mingers (1987) discusses the G^2 metric, and White and Liu (1994) recommend that the χ^2 approximation to either G^2 or X^2 be used instead of gain, gain ratio, etc. We note again that Cochran's criteria for applicability of the χ^2 approximation are seldom satisfied for all splits in a decision tree.

3.2. An Exact Test

Fisher's Exact Test for 2×2 contingency tables (Agresti, 1990) is based on the hypergeometric distribution, which gives the exact probability of obtaining the observed data under the null hypothesis, conditioned on the observed marginal totals (n_c and m_v).

$$P_0 \equiv \left(\begin{array}{c} n_1 \\ f_{11} \end{array} \right) \left(\begin{array}{c} n_2 \\ f_{21} \end{array} \right) \Big/ \left(\begin{array}{c} N \\ m_1 \end{array} \right) \tag{8}$$

The achieved level of significance, α (the confidence level of the test is $1 - \alpha$), is the sum of the hypergeometric probabilities for the observed data and for all hypothetical data having the same marginal totals (n_c and m_v) which would have given a lower value for P_0. Fisher's test is uniformly the most powerful unbiased test (Agresti, 1990), i.e., in the significance level approach to hypothesis testing, no other test will out-perform Fisher's exact test (the power of a test is the probability that the null hypothesis will be rejected when some alternative hypothesis is really true).

White and Liu (1994) note that Fisher's exact test should be used for small e_{cv} instead of the χ^2 approximation, and suggest that a similar test for larger tables could be developed. The extension of Fisher's test for tables larger than 2×2 is the multiple hypergeometric distribution (Agresti, 1990)

$$P_0 = \left(\frac{\prod_{c=1}^{C} n_c!}{N!} \right) \prod_{v=1}^{V} \left(\frac{m_v!}{\prod_{c=1}^{C} f_{cv}!} \right) \tag{9}$$

This exact probability expression can be derived either from classical statistics, as the probability of obtaining the observed data given that the null hypothesis is true (Agresti, 1990), or from Bayesian statistics (Martin, 1995), as the probability that the null hypothesis is true given the observed data. The Bayesian derivation of P_0 differs from Buntine's Beta derivation primarily by conditioning on both the row and column totals of the contingency table, and by eliminating the α parameter.

For choosing among several candidate splits of the same set of data, P_0 is a more appropriate metric than the significance level. If we are seeking the split for which it is least likely that the null hypothesis is true, that is measured directly by P_0, whereas significance measures the cumulative likelihood of obtaining the given split or any more extreme split. (This is consistent with Minger's (1987) suggested use of G^2).

The following approximate relationships can be derived (Martin, 1995), showing that X^2, G^2, and gain arise as terms in alternative approximations to the statistical reliability of class predictions based on split subset membership (*split reliability*):

$$\begin{aligned} 2\ln(2)\ N\ \text{gain} &\approx -2\ln(P_0) - (C-1)(V-1)\ln(2\pi N) \\ &+ \text{(terms increasing as the interaction sum of squares)}^* \end{aligned} \tag{10}$$

$$X^2 \approx -2\ln(P_0) - (C-1)(V-1)\ln(2\pi N) + \text{(terms increasing as the main-effects sum of squares)}^* \quad (11)$$

*In neither case should it be assumed that these terms vanish, even as $N \to \infty$. Both factors are positive, indicating that these measures tend to overestimate the reliability of very non-uniform splits. The sum of squares terminology used here arises in analysis of variance (ANOVA) — main-effects refers to the variances of the marginal totals, m_v and n_c, and interaction refers to the additional variance of the f_{cv} terms over that imposed by the m_v and n_c totals.

3.3. Some Other Measures of Attribute Relevance

This section reviews two noteworthy measures of attribute relevance which are not intended to be used for selecting a split attribute or for stopping, but rather to screen out irrelevant attributes or to predict whether stopping would result in reduced accuracy.

Fisher and Schlimmer (1988) propose a variation of Gluck and Corter's (1985) *category utility* measure (which is itself the basis of Fisher's (1987) COBWEB incremental learning system):

$$\text{F-S} = \text{average of } \sum_v \left\{ \frac{m_v}{N} \sum_c \left[\left(\frac{f_{cv}}{m_v} \right)^2 - \left(\frac{n_c}{N} \right)^2 \right] \right\}$$

Category utility expresses the extent to which knowledge of one attribute's value predicts the values of all of the other attributes (including the class). This variant (F-S) focuses on predicting only the class, and averages the utility of the other attributes in this regard. F-S is not proposed for choosing the split feature, nor for stopping, but to determine the average relevance of a candidate set as a predictor of whether a stopped tree would be less accurate than an unpruned tree (using information gain for splitting and χ^2 for stopping).

In that same context, Fisher (1992) proposes using the average value of Lopez de Mantaras (1991) $(1-d)$ distance measure (Equation 3), rather than F-S, as the predictor for $(1-d)$-splitting and χ^2-stopping. Though d has the mathematical properties of a distance metric, it is perhaps easier to understand in information-theoretic terms. Maximizing information gain minimizes the average number of bits needed to specify the class once it is known which branch of the decision tree was taken. A similar question asks how many bits are needed on the average to specify the branch given the class. d is the sum of the number of bits needed to specify the class knowing the branch and the number needed to specify the branch knowing the class, normalized to the interval [0,1]. Maximizing $(1-d)$ minimizes this collective measure.

Kira and Rendell's (1992) RELIEF algorithm proposes a somewhat different measure (K-R), again not for choosing the split feature, nor for stopping, but for eliminating irrelevant attributes from the candidate set.

$$\text{K-R} = \text{average over } m \text{ randomly chosen items, } i, \text{ of: } \sum \left[\text{diff}(i, M_i)^2 - \text{diff}(i, H_i)^2 \right] \quad (12)$$

where H_i is the Euclidean nearest neighbor which has the same class as instance i; M_i is the nearest neighbor which has the opposite class from instance i; for a nominal attribute

$$\text{diff}(i,j) = \begin{cases} 0, & \text{if instances } i \text{ and } j \text{ have the same value} \\ 1, & \text{if instances } i \text{ and } j \text{ have different values} \end{cases}$$

and, for a numeric attribute

$$\text{diff}(i,j) = |\text{value}(i) - \text{value}(j)| \;/\; |\text{range}|$$

This K-R measure is nondeterministic because of the random choice of m instances from the size N sample, and because of tie-breaking when choosing H_i and M_i. This nondeterminism can lead to a very high variance of K-R for small data sets, and in some cases the decision whether to exclude an attribute from the candidate set can change depending on the random choices made.

Kononenko (1994) proposes extensions of K-R as metrics for choosing the split attribute, primarily letting $m = N$ for small datasets and averaging over k nearest instances H_i and k nearest M_i. Unfortunately, Kononenko's paper mis-states Kira and Rendell's formula (compare Equation 12) as:

$$\begin{aligned} \text{K-R} \;=\; & \text{average over } m \text{ randomly chosen items, } i, \text{ of:} \\ & \sum [\text{diff}(i, M_i) - \text{diff}(i, H_i)] \end{aligned}$$

which could have a profound effect for numeric attributes (though not for the binary attributes tested by Kononenko). Kononenko's results indicate that a localized metric such as K-R may have an advantage in problem domains such as parity, where XOR-like features are common. However, neither Kononenko nor Kira and Rendell seem to have tested their proposed measures on numeric data, so that it is not clear how well these measures will work for numeric data.

4. Correlations Among the Various Measures

Values of each of the primary measures (P_0, gain, gain ratio, distance, orthogonality, chi-squared, and Beta) were calculated for over 1,000 2×2 tables[7]. These data (see Martin, 1995) confirmed the analyses given above (see the discussions around Equations 6 through 11 in Sections 3.1 and 3.2):

- when Cochran's criteria are satisfied (in this case, all $e_{ij} \geq 5$), $G^2 \approx X^2$ and $X^2 \approx -2.927 - 2\ \ln(P_0)$; when they are not, X^2 and G^2 tend to be spuriously high, and overestimate split reliability
- a similar linear relation to $\ln(P_0)$ is found for the other measures, with an even stronger tendency to overestimate split reliability when $X^2 \approx \chi^2$ is not valid
- very high values of information gain and the other measures occur frequently when the null hypothesis cannot be rejected ($P_0 \geq 0.5$) — occurrence of these high values is strongly correlated with circumstances under which the $X^2 \approx \chi^2$ approximation is invalid

Table 2. A Troublesome Data Set

Cat	Attr A		Attr B		Attr C		Attr D		Attr E		Attr F		Attr G		Attr H	
	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
P	23	1	17	7	19	5	3	21	11	13	18	6	15	9	9	15
N	40	0	34	6	35	5	10	30	25	15	35	5	31	9	21	19
Total	63	1	51	13	54	10	13	51	36	28	53	11	46	18	30	34

Attr	min e_{ij}	info gain	gain ratio	$1-d$	ORT	W(1) $\div N$	G^2	X^2	P_0
A	¶ 0.4	.022	.193	.021	.502	.686	1.99	1.69	.375
B	4.9	.020	.028	.012	.078	.693	1.81	1.86	.101
C	3.8	.009	.014	.006	.041	.700	.77	.79	.185
D	4.9	.017	.024	.010	.051	.697	1.53	1.45	.131
E	10.5	.019	.019	.010	.045	.697	1.69	1.69	.090
F	4.1	.018	.027	.011	.079	.694	1.60	1.65	.119
G	6.8	.018	.022	.010	.056	.696	1.64	1.67	.099
H	11.3	.015	.015	.008	.034	.700	1.37	1.36	.106

¶ The X^2 and G^2 tests are unreliable here.

Attr	Normalized Rank (apparent best = 1, worst = 8)							
	info gain	gain ratio	$1-d$	ORT	W(1)	G^2	X^2	P_0
A	1	1	1	1	1	1	2.1	8
B	2.0	7.4	5.0	7.3	4.4	2.0	1	1.3
C	8	8	8	7.9	7.8	8	8	3.3
D	3.6	7.6	5.7	7.8	6.5	3.6	3.7	2.0
E	2.7	7.8	6.1	7.8	6.6	2.7	2.1	1
F	3.2	7.5	5.5	7.3	4.8	3.2	2.4	1.7
G	3.0	7.7	5.9	7.7	5.9	3.0	2.2	1.2
H	4.6	7.9	6.9	8	8	4.6	4.3	1.4

- when $X^2 \approx \chi^2$ (Cochran's criteria) is valid, all of the measures converge (rank splits in roughly the same order, though differing in detail) — when $X^2 \approx \chi^2$ is invalid, the split rankings can be quite divergent

Consider a data set which produces the trial splits shown at the top of Table 2. The middle portion of the table gives the values of each of the split selection metrics (for $W(1)$ and P_0, the lower the value the better the split is taken to be; for the other metrics, the higher the value the better). For easier comparison, the bottom portion of the table gives a normalized rank, 1 indicating the heuristic's best split and 8 the poorest split. The rank, R, is defined as $R_i = a + bX_i$, where X_i is the split's heuristic value and

$$b = \begin{cases} -7/\left(\max(X_i) - \min(X_i)\right) & \text{if } \max(X_i) \text{ is best} \\ +7/\left(\max(X_i) - \min(X_i)\right) & \text{if } \min(X_i) \text{ is best} \end{cases}$$

$$a = \begin{cases} 1 - b\max(X_i) & \text{if } \max(X_i) \text{ is best} \\ 1 - b\min(X_i) & \text{if } \min(X_i) \text{ is best} \end{cases}$$

i.e., a and b assign the value 1 to the best split and 8 to the poorest split, and rank the other splits linearly between them.

Note the strong correlation (0.999) between gain ratio and orthogonality. Likewise, $(1-d)$ and $W(1)$ are strongly correlated (0.965), as are gain, G^2, and X^2 (1.000 for G^2 vs. gain and 0.950 for X^2 vs. either gain or G^2). The correlation of gain, G^2, and X^2 follows from their definitions and was noted in the previous section. The strength of the correlation between gain ratio and orthogonality and that between $(1 - d)$ and $W(1)$ was unexpected, as it is not obvious in their definitions that this should be the case.

Information gain chooses attribute A for the first split, as do all of the metrics except X^2 and P_0 (gain, G^2, and X^2 differ primarily over the question of which of the attributes A and B is best and which second best).

There is but a single instance of $A = 1$ in these data. Intuitively, splitting off single instances in this fashion is hardly efficient. Suppose there were *no* instances of $A = 1$, either because of noise or random chance in drawing the sample. Then, clearly, attribute A would be of no use in separating the data and would have had the lowest gain (zero). Likewise, if there were two instances of $A = 1$, one in each class, attribute A would again have the lowest gain. Apparently, when the relative frequencies of the attribute values are very non-uniform, as here, information gain is hyper-sensitive to noise and to sampling variation.

Gain ratio, distance, orthogonality, and the Beta function all *emphatically* choose attribute A for these data (especially gain ratio and orthogonality), evidence that these measures also suffer (even more) from this hyper-sensitivity. Mingers (1989b) has previously noted and expressed concern about this tendency to favor unbalanced splits.

This attribute (A) is clearly more suited to making subtle distinctions at the end of a chain of other tests than to making coarser cuts near the root of the tree. Only P_0 is qualitatively different from the other metrics, ranking attribute A dead last and clearly a poorer choice than the other attributes.

Hypothesis 1 — The chi-squared statistics, information gain, gain ratio, distance, and orthogonality all implicitly assume an infinitely large sample — i.e., that continuous population parameters are adequately approximated by their discrete sample estimates (e.g., substituting n_c/N for p_c, the proportion of class c in the population), and that a discrete (e.g., binomial) distribution is adequately approximated by a continuous normal distribution.

When Cochran's criteria are not satisfied, these assumptions may be incorrect, and these heuristics inadmissible. For such ill-conditioned data, use of these metrics entails a high likelihood of rejecting the null hypothesis when it is really true. (A data set is ill-conditioned for an analysis when slight changes in the observations would cause large perturbations of the estimated quantities.)

Hypothesis 2 — Buntine's Beta function derivation explicitly assumes that the class distributions in the subsets of a split are *a priori* independent of one another. While this assumption can be admitted for a single split considered in isolation, it is not appropriate when comparing alternative splits of a given population.

For example, given a population where each item has 3 binary attributes:

$$\text{class} = (\text{pos, neg})$$

$$
\begin{aligned}
\text{color} &= \text{(blue, red)} \quad \text{size} = \text{(large, small)} \\
\text{Let} \quad \alpha(i,j) &= \text{Prob}\{\text{class} = i \mid \text{color} = j\} \quad \gamma(j) = \text{Prob}\{\text{color} = j\} \\
\beta(i,k) &= \text{Prob}\{\text{class} = i \mid \text{size} = k\} \quad \delta(k) = \text{Prob}\{\text{size} = k\} \\
\text{and} \quad \theta(i) &= \text{Prob}\{\text{class} = i\} \\
\text{Now,} \quad \theta(i) &= \alpha(i,\text{blue}) \times \gamma(\text{blue}) + \alpha(i,\text{red}) \times \gamma(\text{red}) \qquad (13) \\
&= \beta(i,\text{small}) \times \delta(\text{small}) + \beta(i,\text{large}) \times \delta(\text{large})
\end{aligned}
$$

Because of Equation 13, the two statements "$\alpha(i, \text{blue})$ is independent of $\alpha(i, \text{red})$" and "$\beta(i, \text{small})$ is independent of $\beta(i, \text{large})$" cannot both be true of the same population.

Hypothesis 3 — The null hypothesis probability function P_0 appears to be a measure which properly incorporates all these factors (finite sample size, a discrete, non-normal distribution, Cochran's criteria, and the non-independence of split subsets), and may be a more suitable split selection metric than gain, gain ratio, distance, orthogonality, Beta, or chi-squared.

Viewing these hypotheses in terms of our three criteria for choosing a heuristic function (prefer splits which improve accuracy, prefer splits leading to simpler and more efficient trees, and order the splits to permit practical pruning), since none of the metrics directly measures either accuracy or complexity, the conjecture in Hypothesis 3 must be tested empirically, rather than analytically. Because pruning is a very complex (and often controversial) subject, we chose to do a partial evaluation at this point for the first two criteria in terms of unpruned decision trees, and to defer evaluation of the third criteria (the effectiveness and efficiency of pruning) until later in the paper, after a more extensive discussion of pruning issues.

5. Empirical Comparisons of the Measures for Unpruned Binary Trees

A Common Lisp implementation of ID3 obtained from Dr. Raymond Mooney was used in the experiments described here, substituting different split selection and pruning methods for ID3's information gain and χ^2 tests.

Sixteen data sets were used to evaluate the split metrics. They are described in a technical report (Martin, 1995), and were chosen to give a good variety of application domains, sample sizes, and attribute properties. None of the data sets has any missing values. Two issues arise with respect to the attributes:

- Numeric attributes must be converted to a form having only a few distinct values, i.e., cut into a small number of sub-ranges. Various procedures have been proposed for this, differing along dimensions of
  1. arbitrary vs. data-driven cuts
  2. once-and-for-all vs. re-evaluating cut-points at every level in the tree
  3. *a priori* (considering only the attribute's distribution) vs. *ex post* cut-points (also considering the classification)
  4. multi-valued vs. binary cuts
  5. the function used to evaluate potential cut-points

The particular method used may have important consequences for both efficiency and accuracy, and can interact with selection and stopping criteria in unpredictable ways.

- Orthogonality is defined (Equation 4) only for binary splits, and each attribute having $V > 2$ values must be converted to binary splits for this measure. This can be done most simply by creating V binary attributes. Quinlan (1993) describes a procedure for iteratively merging branches of a split using gain or gain ratio. Other procedures are given by Breiman, et al. (1984) and Cestnik, Kononenko, and Bratko (1987).

In order to control the splitting context and to avoid bias in comparing the selection metrics, two *a priori*, once-and-for-all, multi-valued strategies were used here to convert a numeric attribute to a discrete attribute:

1. 'natural' cut-points determined by visual examination of smoothed histograms
2. arbitrary cut-points at approximately the quartiles (approximate because the cut-points are not allowed to separate instances with equal values — quartiles because the 'natural' cut-points typically give about 4 subsets per attribute).

The resulting cut-points are not intended to be optimum (and may not even be "good"), merely *a priori*, consistent, and unbiased. Results obtained here should be compared only to one another, and not to published results using other (especially *ex post*) strategies on the same data set.

These two procedures were applied to every attribute in each of the ten datasets which contained numeric attributes, resulting in 20 new datasets which contained only discrete attributes. With the 6 original datasets which had no numeric attributes, there were 26 discrete-attribute datasets to be evaluated.

All but two of these datasets (Word Sense and natural cut-points WAIS, for which all attributes are binary) have some attributes with arity $V > 2$. For all 24 of these datasets, a new dataset was created in which all attributes having $V > 2$ values were replaced with V binary attributes. The 26 all-binary datasets permitted a fair comparison of the orthogonality measure to the other measures using exactly the same binary candidate sets. This binarization imposes a significant time penalty for tree-building and post-pruning relative to trees built from the un-binarized data set (see Section 8 and Martin and Hirschberg (1996a)).

In each experiment, a tree was grown using all of the instances, and the complexity and efficiency of this tree were determined. The accuracy of this tree was then estimated by 10-fold cross-validation (Martin and Hirschberg (1996b) show that 10-fold or greater cross-validation usually gives a nearly unbiased estimate of the accuracy of *the* classifier inferred from the entire sample).

It is hard to make a rigorous statement about the significance of the difference in accuracy between any two of the trees because the trees and accuracy estimates repeatedly sub-sample the same small dataset and are not statistically independent (see Martin & Hirschberg (1996c) for a full discussion of this question). In this study, we use the 2-SE heuristic test for significance proposed by Weiss and Indurkhya (1994) — the difference in two accuracy estimates, A_1 and A_2, is heuristically significant at the 95% confidence level if $|A_2 - A_1| \geq 2$ SE, where SE $= \sqrt{A(1-A)/N}$ and $A = (A_1 + A_2)/2$.

The results for unpruned trees using the various metrics are shown in Table 3. Only summaries of complexity, efficiency, and practicality are shown here, full results are given in a technical report (Martin, 1995). Two values are shown for the Beta metric's α parameter: 1, the uniform prior; and 0.5, the Jeffreys prior recommended by Buntine. The G^2 and X^2 trees were built without regard to significance or to admissiblity (Cochran's criteria).

None of the differences in accuracy between split metrics is statistically significant at the 95% level. The accuracy summary figures are averages weighted by the sample sizes. These averages are sensitive to systematic differences between the metrics. The 'overall' figure includes the six datasets which do not have a natural/quartiles distinction, and is therefore not simply the average of the natural and quartiles averages. The difference in accuracy between arbitrary and 'natural' nominalizations is very small, except for the Glass and WAIS data, and sometimes positive, sometimes negative.

Except for two of the datasets, the trees inferred using the various metrics all have about the same number of leaves. For the BUPA liver disease and Pima diabetes datasets, the number of leaves varies more widely between the metrics, with the G^2 and X^2 trees consistently having the fewest leaves for these two datasets. Note the large difference between the overall total number of leaves and the sum of the natural and quartiles totals — the trees for the Word Sense and the Solar Flare C and M datasets are very complex (around 250 leaves for Word Sense, and 60-80 leaves each for the Solar Flare trees).

All of the measures build shallower trees with more leaves for the quartiles splits than for the natural splits. This reflects the fact that a classifier must be more complex to deal with an arbitrary (quartiles) division into subsets. The quartile trees are all about the same depth. For the natural cut-points, the P_0 trees are consistently shallower and the gain ratio and orthogonality trees are consistently deeper than those of the other metrics for all of the datasets, reflecting a tendency for all the metrics except P_0 to be 'fooled' into using splits with one or more very small subsets (which occur frequently in the natural subsets data).

With a single exception (the WAIS data, where the natural subsets are binary), the quartile subsets reduce training time, 40-50% on the average. This time savings is directly attributable to the reduced dimensionality (number of attribute-value pairs) of the quartile subsets. The large difference between the overall total time and the sum of the natural and quartile totals is due almost entirely to the Word Sense dataset — in every case, this one dataset (which has 100 binary attributes) took longer than all of the others combined.

P_0 is more practical (faster) in virtually every case (the sole exception being the quartiles BUPA dataset, where the X^2 and G^2 metrics were slightly faster). P_0 reduced training time by 30% on the average over the nearest competitor (X^2) and by 60% over the least practical (gain ratio). Martin and Hirschberg (1996a) give a theoretical analysis of the worst-case and average-case time complexity of TDIDT, which was confirmed empirically using the detailed data underlying Table 3.

These data support the conjecture that in virtually every case unpruned trees grown using P_0 are less complex (in terms of the number of leaves), more efficient (expected classification time) and more practical (learning time), and no less accurate than trees grown using the other metrics. They also reinforce the conclusion that for unpruned trees, the choice of metric is largely a matter of complexity and efficiency, and has little effect on accuracy.

Table 3. Unpruned Trees, Binary Splits

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
					Cross-Validation Accuracy, %					
BUPA	Nat	54	56	51	52	58	59	54	55	60
	Qua	66	59	60	59	64	62	61	62	62
Finance 1	Nat	77	75	75	71	71	71	73	79	75
	Qua	72	77	75	77	65	73	77	77	75
Finance 2	Nat	91	92	95	94	94	91	91	95	92
	Qua	86	95	91	91	94	92	92	94	92
Solar Flare C		87	87	87	86	87	85	85	86	86
Solar Flare M		85	85	84	82	83	85	87	83	85
Solar Flare X		97	96	96	97	96	97	97	97	97
Glass	Nat	51	50	50	50	50	47	51	51	53
	Qua	72	70	69	72	69	72	70	66	70
Iris	Nat	95	95	93	96	96	95	94	94	95
	Qua	91	91	91	92	91	89	91	91	90
Obesity	Nat	56	58	56	60	51	49	53	47	58
	Qua	40	51	49	47	56	44	44	51	51
Pima	Nat	72	71	73	70	70	71	72	71	70
	Qua	65	67	68	64	66	67	69	67	65
Servo Motors		95	95	95	96	95	96	93	95	95
Soybean		98	98	98	98	98	98	98	98	98
Thyroid	Nat	91	91	90	90	91	90	91	91	89
	Qua	93	93	93	92	93	92	93	94	93
WAIS	Nat	84	84	84	84	82	84	84	84	80
	Qua	61	67	61	65	57	63	67	71	65
Wine	Nat	91	92	89	86	91	92	90	94	91
	Qua	93	90	89	89	89	94	92	94	89
Word Sense		64	64	64	63	64	63	64	65	64
Overall		75.1	74.9	74.8	73.7	74.7	74.8	75.3	75.3	75.0
Natural		72.8	72.6	72.1	70.9	72.5	72.4	72.5	72.9	73.0
Quartiles		73.2	73.1	73.1	71.9	73.0	73.5	74.3	73.8	72.6
					Total Number of Leaves					
Overall		1295	1267	1191	1351	1318	1272	1070	1061	1213
Natural		371	365	371	369	324	311	262	259	298
Quartiles		531	488	424	562	536	527	421	421	502
					Weighted Average Depth					
Overall		9.9	14.0	10.1	15.8	10.3	9.1	8.1	9.1	7.3
Natural		13.6	16.6	13.9	17.1	10.6	9.6	8.9	9.0	6.6
Quartiles		6.5	6.7	5.9	7.5	6.7	6.5	5.8	5.8	6.2
					Total Run Time (sec)					
Overall		5852	7883	6585	7394	5004	4956	4490	4410	3028
Natural		1236	1711	1606	1242	1036	1049	998	943	746
Quartiles		778	936	959	655	621	652	606	536	428

6. Stopping Criteria

A characteristic of these kinds of inductive algorithms is a tendency to overfit noisy data (noise in the form of sampling variance, incorrect classifications, errors in the attribute values, or the presence of irrelevant attributes). Breiman, et al. (1984) initially searched for a minimum gain threshold to prevent this overfitting. Since $(N \times \text{ gain})$ has approximately a χ^2 distribution (which has very complex thresholds), setting a simple threshold for gain was not successful.

Quinlan (1986) originally proposed that the $X^2 \approx \chi^2$ significance test (Equation 6) be used to prevent overfitting in ID3 by stopping the process of splitting a branch if the 'best' split so produced were not statistically significant. Besides the unfortunate interaction exemplified by the XOR problem, there are two reasons that this strategy does not work well:

1. the χ^2 approximation to X^2 should not be used for splits with small e_{cv} components (there are similar difficulties with G^2 and with gain) — the divide-and-conquer strategy of TDIDT creates ever smaller subsets, so that this difficulty is certain to arise after at most $\log_2(N_0/5)$ splits have been made (N_0 is the size of the entire data set)

2. X^2 and gain converge at different rates and may rank splits in different orders — gain probably does not order the splits correctly for pre-pruning by X^2

Both of these approaches were abandoned in favor of some form of post-pruning, such as cost-complexity pruning (CART, Breiman, et al., 1984), reduced-error pruning (Quinlan, 1988), or pessimistic pruning (C4.5, Quinlan, 1993). There have been a number of studies in this area (e.g., Buntine & Niblett, 1992; Cestnik & Bratko, 1991; Fisher & Schlimmer, 1988; Mingers, 1989a, 1989b; Niblett, 1987; Niblett & Bratko, 1986; Schaffer, 1993). Among the notable findings are:

- in general, it seems better to post-prune using an independent data set than to pre-prune as originally proposed in ID3

- k-fold cross-validation seems to work better for pruning than point estimates such as X^2

- the decision to prune is a form of bias — whether pruning will improve or degrade performance depends on how appropriate the bias is to the problem at hand

- pruning, whether by X^2 or cross-validation, may have a negative effect on accuracy when the training data are sparse (i.e., ill-conditioned)

Note — A decision to prune the data opens the possibility of committing *Type II* errors (accepting the null hypothesis when some alternative hypothesis is really true, as in pre-pruning in the XOR problem). A decision not to prune when using real data almost certainly introduces *Type I* error (overfitting — rejecting the null hypothesis when it is really true).

Consider, for example, the following potential splits:

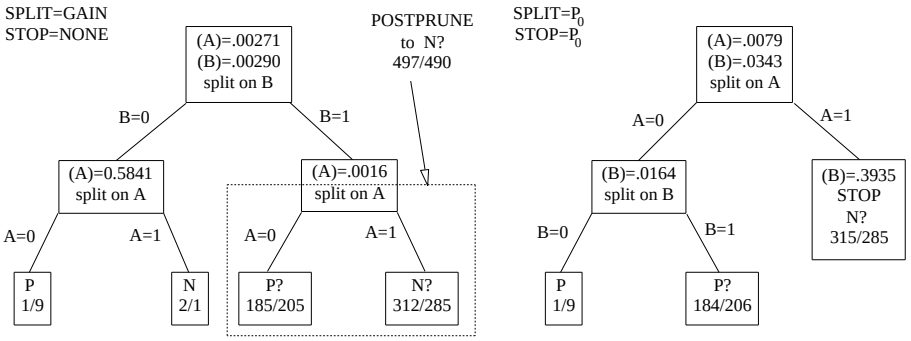

Figure 2. Alternative Split/Stop Strategies

Attribute A			
	$A=0$	$A=1$	Total
Class N	185	315	500
Class P	215	285	500
Total	400	600	1000
info. gain=.00271			
X^2=3.750		P_0=.0079	

Attribute B			
	$B=0$	$B=1$	Total
Class N	3	497	500
Class P	10	490	500
Total	13	987	1000
info. gain=.00290			
X^2=3.819		P_0=.0343	

Attribute B has a slightly larger gain (.00290) than does attribute A (.00271), and so attribute B would be chosen for the first split of an unpruned ID3 tree (gain ratio, orthogonality, $(1-d)$, and $W(1)$ all also would choose attribute B). X^2 for this split (3.819) is slightly below the 95% confidence cut-off (3.841), as is X^2 for attribute A (3.75), and so both splits are disallowed by this criterion and ID3 (splitting on gain and stopping on X^2) stops without generating any tree.

Splitting on gain and post-pruning (by C4.5's pessimistic method) leads to the rule $[(B = 0) \wedge (A = 0) \Rightarrow (\text{Class } P)]$ (see the left-hand tree in Figure 2) . $P_0 < 0.05$ for both splits, and the more balanced attribute A is the better choice (.008 for A versus .034 for B). Splitting and stopping using P_0 (see the right-hand tree in Figure 2) leads to the more general rule $[(A = 0) \Rightarrow (\text{Class } P)]$ more directly, without generating and later pruning a subtree on the right hand branch of the root.

Hypothesis 4 — The previous negative results concerning pre-pruning (e.g., Breiman, et al., 1984; Quinlan, 1988; Schaffer, 1993) may be due to use of different inadmissible statistics for split selection and stopping, and to interaction with the restricted split candidates set, rather than to any inherent fault of pre-pruning. Use of the P_0 function for both selection and stopping might permit more practical construction of simpler and more efficient decision trees without loss of predictive accuracy (except for problems such as parity, where the XOR problem might require an expanded candidate set if the tree is to be stopped).

7. Empirical Studies of Pre- and Post-Pruning

As was the case for our earlier conjecture (see Hypothesis 3, p. 270), the conjecture in Hypothesis 4 must be tested empirically. In this section we present and contrast results from post-pruning using C4.5's pessimistic pruning method (Quinlan, 1993); splitting and pre-pruning using P_0; and splitting using one of the other metrics and pre-pruning using a χ^2 criterion.

7.1. Effects of Post-Pruning

Quinlan's (1993) pessimistic post-pruning method (C4.5) was used at the default 0.25 confidence factor level. The results are summarized in Table 4. Some of the noteworthy features of these data are:

1. Using the 2-SE criterion (Section 5), there are no significant differences in accuracy between unpruned and post-pruned trees, nor between the various metrics. That is, the choice of a splitting heuristic and whether or not to post-prune are largely matters of complexity, efficiency, and practicality, not of accuracy.

2. The differences in complexity and efficiency between metrics are much smaller after post-pruning. The trees that 'overfit' most benefit most from post-pruning, though some 'overfitting' remains after post-pruning.

3. Post-pruning had virtually no effect on the complexity and efficiency of the trees built using P_0 as the splitting metric, and very little effect on the trees built using G^2 and X^2. Post-pruning these trees was largely wasted effort.

4. Post-pruning can be very expensive, and is not cost-effective. Trees with comparable accuracy, complexity, and efficiency can be obtained at half the run-time (or less) by using P_0 as the splitting metric without post-pruning, rather than using another metric and post-pruning.

It was somewhat surprising that C4.5's pessimistic pruning method was not more effective. Pessimistic post-pruning had little effect (less than 6% reduction in the total number of leaves) on the quartiles cut-points trees for any of the metrics. For the natural cut-points, by contrast, the number of leaves was reduced by 28-34% for the gain, gain ratio, $1 - d$, and orthogonality trees (the reduction was lower for the other metrics, and only 2% for the P_0 trees).

It appears that the pessimistic post-pruning method is fairly effective in dealing with the very unbalanced splits which are common in the natural cut-points data (especially when using gain or gain ratio and similar metrics), but is less effective in pruning the more balanced trees built using P_0, G^2, or X^2. That is, while gain and gain ratio, etc. are biased towards choosing very unbalanced splits, pessimistic post-pruning has the opposite bias (against the unbalanced splits) and offsets the bias of these metrics to achieve in the end a tree with roughly the same complexity as the unpruned G^2, X^2, and P_0 trees. This cancelling of the biases illustrates our third criterion (see Section 2) for choosing a split selection heuristic, i.e., that the heuristic should order the splits so as to permit effective simple pruning.

Table 4. Post-Pruned Trees, Binary Splits

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
					Cross-Validation Accuracy, %					
BUPA	Nat	59	57	54	59	59	58	53	55	59
	Qua	60	63	59	61	63	62	60	59	64
Finance 1	Nat	71	75	67	67	73	71	71	71	73
	Qua	71	83	69	69	69	69	75	77	71
Finance 2	Nat	94	94	94	95	94	94	91	94	94
	Qua	89	92	92	89	92	94	89	91	91
Solar Flare C		87	85	86	87	87	89	86	85	86
Solar Flare M		85	86	86	85	87	85	85	86	84
Solar Flare X		95	96	96	96	98	97	98	97	97
Glass	Nat	48	54	52	52	52	50	51	50	51
	Qua	70	69	68	72	69	66	73	70	66
Iris	Nat	95	95	95	95	95	95	95	95	96
	Qua	90	92	93	91	91	91	92	92	91
Obesity	Nat	51	49	53	51	53	58	58	64	51
	Qua	53	38	53	44	56	51	56	58	58
Pima	Nat	71	72	72	71	71	72	73	72	71
	Qua	65	67	65	65	64	68	67	68	66
Servo Motors		95	95	93	96	95	94	95	96	95
Soybean		98	98	98	98	98	98	98	98	98
Thyroid	Nat	90	89	90	88	91	91	90	90	91
	Qua	94	93	93	92	92	94	95	93	92
WAIS	Nat	84	84	84	84	84	84	80	84	84
	Qua	69	59	59	61	65	63	65	65	59
Wine	Nat	90	93	92	84	90	91	89	90	91
	Qua	89	90	92	89	91	92	91	92	89
Word Sense		65	64	63	64	63	64	65	64	64
Overall		74.7	75.3	74.5	74.6	75.0	75.5	75.4	75.2	75.1
Natural		72.5	73.4	72.7	72.2	73.1	73.3	72.4	72.9	73.2
Quartiles		72.2	73.3	72.1	72.3	72.5	73.8	74.0	73.8	73.0
					Total Number of Leaves					
Overall		1155	1051	1039	1142	1141	1171	1025	1019	1203
Natural		267	240	253	263	257	262	243	244	292
Quartiles		504	462	410	529	519	521	406	409	499
					Weighted Average Depth					
Overall		7.6	10.0	8.1	10.3	8.4	8.1	7.7	8.6	7.2
Natural		7.0	9.9	8.5	9.5	7.7	7.5	7.9	8.1	6.4
Quartiles		6.4	6.2	5.7	7.0	6.5	6.5	5.7	5.8	6.2
					Total Run Time (sec)					
Overall		7448	26868	9165	43506	9301	7787	6676	9794	5261
Natural		1769	4669	2539	5783	2189	1960	1727	1717	1086
Quartiles		1244	1340	1255	1135	1017	1016	1000	866	867

The run-times for tree building and post-pruning are roughly proportional to the data set size and exponential in the (weighted average) unpruned tree depth[8]. The post-pruned trees built using P_0 had the shortest run-times in all but one case (the quartiles BUPA data, as in Table 3). The gain ratio and orthogonality metrics consistently have above-average run-times, though this is somewhat exaggerated in the summary figures due to extremely long run-times for these two metrics on the Word Sense and Pima (natural cut points) data sets — both are large samples with many attributes and several very small split subsets.

7.2. Effects of Stopping

The effects of stopping based on P_0 are summarized in Table 5. Though accuracy for the Servo and Obesity problems is reduced by pruning at the 0.05 level, the differences are not statistically significant by our 2-SE heuristic (see Section 5). The improved accuracy of the pre-pruned quartiles Pima data is highly significant (the difference is approximately 5-SE, where $\geq$ 2-SE is deemed significant).

The decreased accuracy for the Servo data is largely due to pruning an XOR-like subtree. As mentioned earlier, pre-pruning when the candidate set is univariate is subject to this XOR difficulty. Post- rather than pre-pruning, or lookahead, or some other scheme for expanding the candidate set (see Section 10) would be beneficial for this dataset.

For the Obesity data, linear discriminant analysis fails, suggesting that the classes are not homogeneous (this will be discussed further in Section 10, and see Figure 7). The Obesity attributes are very noisy and correlated, and the data are very sparse (only 45 instances) relative to the concept being studied. A pruning strategy of varying the pruning threshold according to sample size would be beneficial for this dataset (this strategy will be discussed later, in Section 9).

The overall accuracy in Table 5 is mildly concave, peaking at around the $P_0 = 0.05$ level. Growing and stopping decision trees using P_0 at the 0.05 level usually does no harm and may, in fact, be mildly beneficial to accuracy.

At the 0.05 level, the number of leaves is reduced by 75% from the unpruned P_0 and gain trees. The average depth is reduced by 35% over the unpruned P_0 trees, and by 50% over unpruned gain trees. Training time is reduced by 30% over unpruned P_0 trees, by 60% over unpruned gain trees, and by 75% over post-pruned gain trees.

The overall effects of splitting and stopping using P_0 versus splitting using the various metrics and then post-pruning by the pessimistic method are shown in Figure 3. The x-axis labels in Figure 3 indicate which metric was used to select splits in building the trees.

The overall summary figures from Tables 3, 4, and 5 are plotted on the y-axes of Figure 3 as three bars for each metric, showing the results for unpruned, post-pruned, and pre-pruned trees. Note that the y-axis scale in Figure 3d is logarithmic, and equal increments on this axis represent a doubling of the learning time.

The P_0 trees in Figure 3 were pre-pruned using P_0 at the 0.05 level. Trees for the other metrics were pre-pruned using the χ^2 criterion provided in Mooney's ID3 implementation (disallow a split if Cochran's criteria are satisfied and X^2 is less than the 95% critical value of χ^2 for the split, but stop *iff* all candidates are disallowed). This χ^2 rule rarely resulted in a tree different from the unpruned tree — χ^2-stopping was largely ineffective because the χ^2 test is rarely admissible (Cochran's criteria are rarely met) after the first few splits (i.e., the

Table 5. Effects of Stopping, Binary Splits

data set	Nominalize	Unpruned	conf. limits	Pruning Threshold Level 0.5	0.1	0.05	0.01	0.005	0.001
				Cross-Validation Accuracy, %					
BUPA	Natural	60	53-67	58	61	57	54	58	58
	Quartiles	62	55-69	65	59	57	64	62	54
Finance 1	Natural	75	57-90	73	77	77	65	69	60
	Quartiles	75	57-90	79	79	79	71	64	¶ 44
Finance 2	Natural	92	80-99	97	94	94	94	94	94
	Quartiles	92	80-99	88	97	97	92	97	94
Solar Flare C		86	80-91	89	88	88	89	89	89
Solar Flare M		85	79-90	85	89	90	90	90	90
Solar Flare X		97	94-99	98	98	98	98	98	98
Glass	Natural	53	44-62	50	52	52	52	44	46
	Quartiles	70	61-78	68	67	70	65	61	63
Iris	Natural	95	89-99	95	95	95	96	96	96
	Quartiles	90	82-96	91	92	92	94	94	92
Obesity	Natural	58	37-77	47	44	49	40	¶ 33	¶ 13
	Quartiles	51	31-71	42	49	49	40	¶ 29	36
Pima	Natural	70	65-74	70	72	72	70	73	71
	Quartiles	65	60-70	68	§ 73	§ 73	§ 74	§ 74	§ 75
Servo Motors		95	89-98	93	91	89	89	90	¶ 81
Soybean		98	88-100	98	98	96	98	98	98
Thyroid	Natural	89	82-94	91	93	93	91	91	90
	Quartiles	93	87-97	94	93	92	92	91	91
WAIS	Natural	80	62-93	82	84	84	78	76	76
	Quartiles	65	45-82	67	65	63	63	74	76
Wine	Natural	91	84-96	92	93	92	92	86	87
	Quartiles	89	82-95	90	89	88	86	89	85
Word Sense		64	59-68	65	66	66	67	65	64
Overall		75.0	73.4-76.4	75.3	76.4	76.4	75.9	75.5	74.1
Natural		73.0	70.2-75.6	72.5	74.0	73.4	71.7	71.7	70.7
Quartiles		72.6	69.8-75.2	74.0	74.9	75.1	75.0	74.8	73.1
¶ below the 95% confidence limits				§ above the 95% confidence limits					
				Total Number of Leaves					
Overall		1213		895	406	295	192	164	125
Natural		298		231	122	82	58	48	39
Quartiles		502		394	151	109	68	60	44
				Weighted Average Depth					
Overall		7.30		6.71	5.21	4.78	3.83	3.57	2.96
Natural		6.67		5.78	4.29	4.12	3.16	2.95	2.40
Quartiles		6.22		6.06	4.77	4.19	3.30	2.99	2.42
				Total Run Time (sec)					
Overall		3028		2799	2387	2242	1960	1889	1733
Natural		746		684	569	520	461	437	386
Quartiles		428		409	335	307	259	252	221

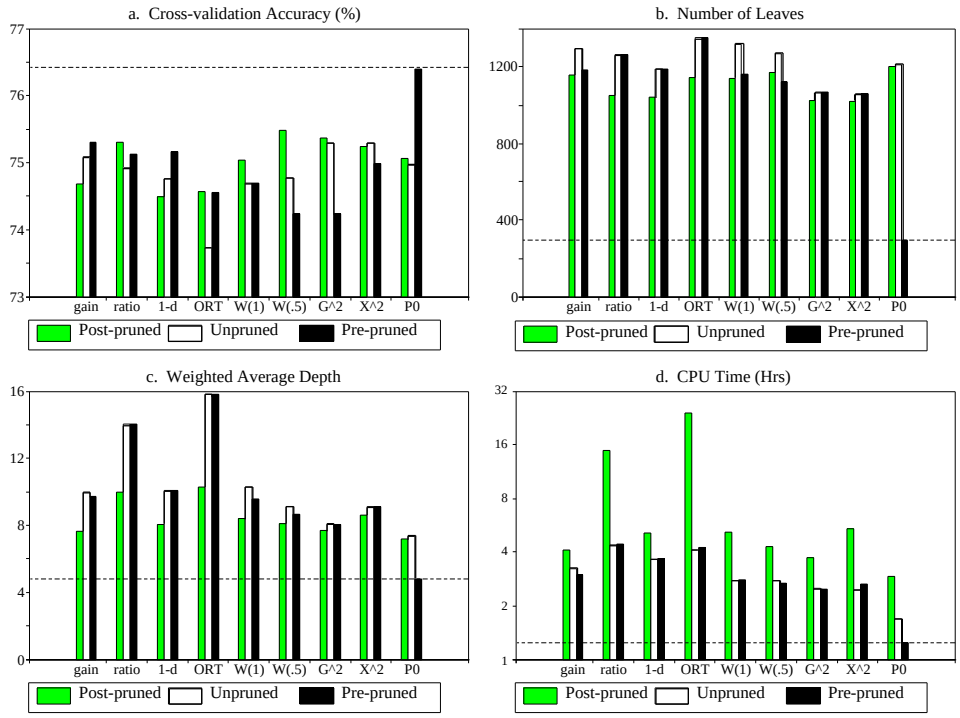

Figure 3. Stopping vs. Post-Pruning

most beneficial splits) have been made. χ^2-stopping was effective only for the trees built using Buntine's Beta function and the quartiles trees (but not the natural cut-points trees) built using information gain. The fact that P_0 is always admissible but χ^2 rarely is accounts for the large difference in the effectiveness of P_0-stopping versus χ^2-stopping in Figure 3. χ^2-stopping likely would have resulted in smaller trees if the χ^2 criterion informed pruning irrespective of Cochran's criteria, although probably not as small as when using P_0 since splits tend to look more informative when χ^2 is not admissible. Another option to consider would be to always prune in cases where Cochran's criteria are not satisfied.

Splitting and stopping using P_0 was more practical, and resulted in trees which were simpler, more efficient, and typically no less accurate than splitting using any of the other metrics and either χ^2-stopping or pessimistic post-pruning.

8. Binary vs. Multi-way Splits

An additional set of experiments was conducted to determine the effects of having used binary as opposed to multi-way splits. These data are summarized in Table 6. The multi-way trees have two or three times as many leaves as the binary trees, are only one-half to one-third as deep, and reduce training and validation time by 80-85%. The time savings is

a straightforward consequence of the increased branching factor reducing the height of the tree and of roughly halving the dimensionality.

A substantial time penalty is incurred when V-ary attributes are forced into V binary splits. Overall, learning time increases quadratically in the dimensionality of the data. Approaches such as those suggested by Weiss and Indurkhya (1991) to reduce dimensionality and optimum binarization techniques such as those used in C4.5 (Quinlan, 1993) and ASSISTANT (Cestnik, et al., 1987) should be pursued. With the *caveat* that the method of handling numeric attributes and steps to reduce dimensionality can influence accuracy and interact with stopping in unpredictable ways.

There is a slight decrease in accuracy for the multi-way splits which becomes smaller as the stopping threshold level decreases (and, in fact, is sometimes reversed below the 0.01 level). The effect is more pronounced for data sets with lower accuracy. Shavlik, Mooney, and Towell (1991) report a similar increase in accuracy of ID3 for binary encoding of the attributes.

The loss in accuracy and the better performance of pre-pruning when using the multi-way splits can be explained by taking note of the very large number of leaves typically found in the multi-way trees. Each multi-way split has more subsets and, on the average, smaller subsets than a binary split. Subsequently splitting the smaller subsets is more likely to overfit due to chance attribute/class association, and P_0-stopping is effective in preventing overfitting in these over-fragmented trees.

These findings suggest that a strategy such as that available as an option in C4.5 (Quinlan, 1993) for merging the values of an attribute to reduce its arity and produce more nearly balanced splits would be beneficial. Finding an optimum or near-optimum strategy based on P_0 rather than on C4.5's gain ratio test is a promising topic for future research.

9. When and How Strongly to Pre-prune

Fisher and Schlimmer (1988) and Fisher (1992) report that χ^2-stopping tends to improve accuracy if the average relevance of a candidate set's attributes is very low, and tends to be detrimental otherwise with an increasing detriment as the average relevance increases. In those studies, average relevance was measured either by the average of Lopez de Mantaras's $(1-d)$ measure or by the F-S measure of the candidate set (see Section 3.3), and the χ^2 test was applied without regard to Cochran's criteria.

Figures 4a and b show the results of a similar study for P_0-stopping. The y-axis in Figure 4 is the ratio of the unpruned P_0 tree's accuracy to the accuracy of the P_0-stopped tree (data taken from Table 5). A y value of 1 indicates no difference in accuracy; a value < 1 that stopping improved accuracy; and a value > 1 that stopping reduced accuracy. The x-axis value in Figure 4 is the average of $(1-d)$ over each split considered at the root of the tree for each data set.

The intercept of the regression line in Figure 4a is > 1, suggesting that P_0-stopping is harmful when the average attribute relevance is low. The negative slope of the regression line indicates that P_0-stopping slowly becomes more effective as the average relevance increases.

In Fisher's (1992) study for trees built using $1-d$ and pruned using χ^2, the corresponding regression line was $0.944 + 1.085x$, with $r^2 = 0.438$ (the positive slope $+1.085$ was

Table 6. Binary vs. Multi-way Splits

Data Set		Binary					Multi-way				
		unpruned		P_0 pruned			unpruned		P_0 pruned		
		Gain	P_0	0.05	0.01	0.005	Gain	P_0	0.05	0.01	0.005
Cross-Validation Accuracy %											
BUPA	Nat	54	60	57	54	58	55	57	59	51	57
	Qua	63	62	57	63	62	57	58	57	58	57
Finance 1	Nat	77	75	¶ 77	65	69	67	65	¶ 58	67	54
	Qua	¶ 72	75	79	¶ 71	64	¶ 52	71	71	¶ 52	56
Finance 2	Nat	91	92	94	94	94	95	92	91	91	92
	Qua	86	92	97	92	97	89	89	97	97	97
Flare C		87	76	88	89	89	85	87	88	88	88
Flare M		85	85	90	89	89	85	85	89	89	90
Flare X		97	97	98	98	98	97	97	97	97	97
Glass	Nat	51	¶ 53	52	52	44	47	¶ 44	51	50	50
	Qua	72	70	70	65	61	66	69	62	62	57
Iris	Nat	95	95	95	96	96	95	93	97	97	97
	Qua	91	90	92	94	94	92	90	90	93	93
Obesity	Nat	56	58	49	§ 40	§ 33	51	58	56	§ 62	§ 56
	Qua	40	51	49	40	§ 29	49	40	44	58	§ 63
Pima	Nat	72	70	72	70	73	70	70	71	72	72
	Qua	68	65	73	74	74	67	67	71	74	74
Servo		95	95	89	89	90	96	95	93	92	93
Soybean		98	98	96	98	98	98	98	96	98	98
Thyroid	Nat	91	89	93	91	91	91	91	91	91	91
	Qua	93	93	92	92	91	92	93	92	92	92
WAIS †	Qua	61	65	63	63	73	65	67	73	71	71
Wine	Nat	90	90	92	92	86	89	90	88	91	90
	Qua	93	89	89	§ 86	93	91	92	90	§ 93	91
Overall		77.1	76.7	78.0	77.4	77.4	75.5	76.1	77.1	77.5	77.4
Natural		72.6	72.8	73.2	71.6	71.6	71.0	71.1	72.6	71.9	72.2
Quartiles †		74.1	72.7	75.4	75.3	74.9	71.9	72.7	73.1	75.0	74.1
¶ binary is better (95% confidence level)							§ multi-way is better (95% confidence level)				
Number of Leaves											
Overall		944	955	216	140	120	2155	2384	717	487	425
Natural		366	292	80	56	46	880	911	311	237	199
Quartiles †		422	502	109	68	60	975	1136	327	210	186
Weighted Average Depth											
Overall		8.8	6.2	3.9	3.0	2.7	3.5	3.6	2.6	2.3	2.1
Natural		13.9	6.8	4.2	3.2	3.0	4.3	4.2	3.0	2.7	2.5
Quartiles †		5.8	6.2	4.2	3.3	3.0	3.1	3.3	2.6	2.3	2.2
Training & Validation Time (sec)											
Overall		2074	1354	944	809	771	323	201	170	159	156
Natural		1234	745	519	460	436	127	83	72	68	67
Quartiles †		585	428	307	259	252	127	72	64	58	57

† Not included in Quartiles summary. WAIS (Natural) and Word Sense are binary.

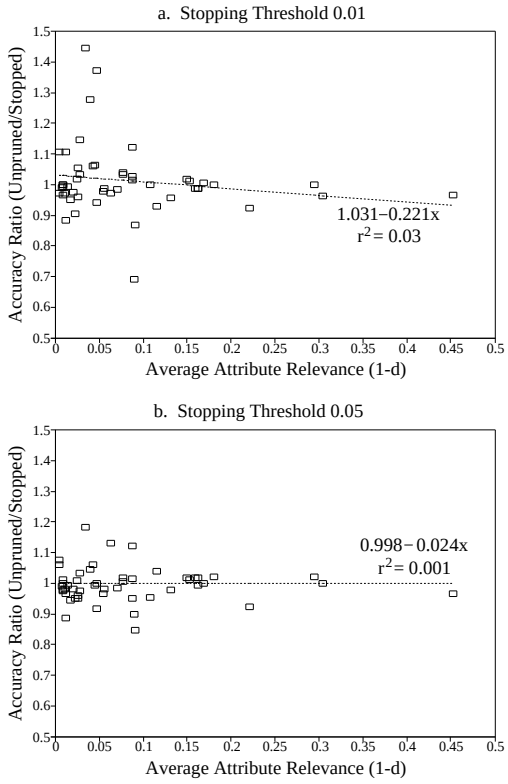

Figure 4. Relative Accuracy vs. Average Relevance

significant). For the P_0 stopping shown in Figure 4, the slope is slightly, but not significantly, negative. Very similar results were found for other stopping thresholds and for other measures of average relevance (F-S and $-\ln P_0$).

These results provide further evidence that using P_0 both to choose the split and to stop growing the tree is qualitatively different from the older, e.g., ID3 (Quinlan, 1986), strategy of splitting based on information gain or a similar measure and stopping based on χ^2. In contrast to Fisher and Schlimmer's results, there is no significant evidence here of a systematic relationship between the average attribute relevance and the accuracy of a P_0-stopped tree compared to the unstopped P_0 tree.

Figures 5a and b show the effects of P_0-stopping as a function of the dataset size. Though the correlation (r^2) is low, it is significant. Stopping tends to be detrimental for very small datasets (though the results are highly variable, and stopping is sometimes beneficial even with a very small dataset). For the largest datasets, P_0-stopping is beneficial and the effect tends to be more certain.

These latter results are consistent with the findings of Schaffer (1993), Fisher (1992), and Fisher and Schlimmer (1988) concerning the effect of dataset size on pre-pruning, and with

Fisher's (1992) suggestion of a more optimistic (non-pruning) strategy for small datasets and a more pessimistic (strong pre-pruning) strategy for large datasets. Figure 5c shows the generally beneficial net effects of the following simple strategy of varying the stopping threshold with the dataset size:

Dataset Size	P_0-stopping Threshold
$N < 50$	do not stop
$50 \leq N < 100$	0.5
$100 \leq N < 200$	0.1
$200 \leq N < 500$	0.05
$N \geq 500$	0.01

Comparing Figure 5c to Figures 5a and b, note that in Figures 5a and b the outcome is very uncertain for the smallest datasets, but stopping is slightly harmful on the average for these small datasets (more uncertain and more harmful on the average the more severely the tree is pruned, i.e., the lower the P_0 threshold). The variable strategy in Figure 5c eliminates both the uncertainty and the slightly harmful average effect by simply not stopping for the very small datasets.

For the largest data sets, the effects of stopping are more certain and more certain not to be harmful to accuracy, even for fairly severe pruning (0.01 P_0 threshold). The variable strategy maintains both this advantage and the reduced complexity the pruning yields.

For the intermediate dataset sizes, the primary effect of the variable threshold strategy is to reduce the uncertainty of the accuracy outcome, as evidenced by the lower scatter of Figure 5c relative to Figures 5a and b in the region $100 \leq N \leq 200$, while maintaining the advantage of reduced complexity.

The high level of uncertainty in Figures 5a and b is largely a consequence of the fact that P_0 and the other metrics are discrete variables, though we treat them as if they were continuous. For small samples, the increments between the discrete values of P_0 are large and small perturbations of the data may cause a large change in P_0, affecting both the choice of the split attribute and whether to stop. The statistics on which our decisions are made are very sensitive to noise and unstable under the perturbations caused by cross-validation re-sampling in small samples, and the variable threshold strategy simply takes this sensitivity and instability into account.

In Fisher's (1992) terminology, the variable threshold strategy is optimistic for the smallest datasets, and more pessimistic as dataset size increases. The strategy can also be viewed from the standpoint of risk management. Defining risk as the likelihood that stopping will result in reduced accuracy and uncertainty as the variance of the y-axis in Figure 5, the strategy is more cautious when the risk and uncertainty of stopping are high (i.e., for very small datasets) and increasingly bold as the risk and uncertainty decrease with increasing dataset size.

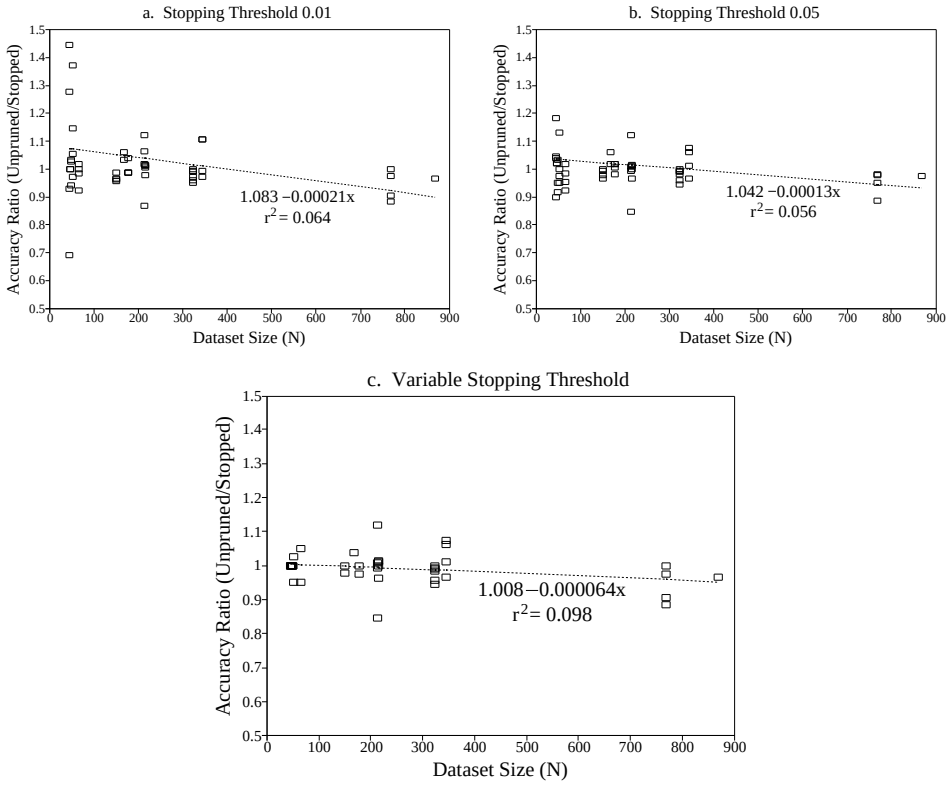

Figure 5. Relative Accuracy vs. Dataset Size

10. Impact of Different Choices of Candidate Split Sets

We have seen in the XOR example that the choice of a candidate set can interact strongly with other factors, particularly with pre-pruning, to preclude or strongly bias against discovering accurate decision trees for some problems. Figure 6 illustrates a different aspect of the choice of candidate sets. Here, there are 2 continuous attributes (x and y) and 2 classes, and the boundary between the two classes is linear (class = 1 if $y > x$, else class = 0). If the candidate splits are restricted to splitting on a single attribute, each of the decision tree leaves covers a rectangular area with sides parallel to the axes. The oblique boundary between the two classes can at best be approximated as a step function, and the accuracy of the tree is directly related to the complexity of the tree and to the sample size (the more leaves and the smaller the area covered by each leaf, the better — the shaded area in Figure 6 is equal to the error rate in the region $0 \leq x \leq 1$ and $0 \leq y \leq 1$). If the splits are further restricted so as to allow only binary splits on continuous attributes, a deeper tree will be required in order to achieve the same accuracy.

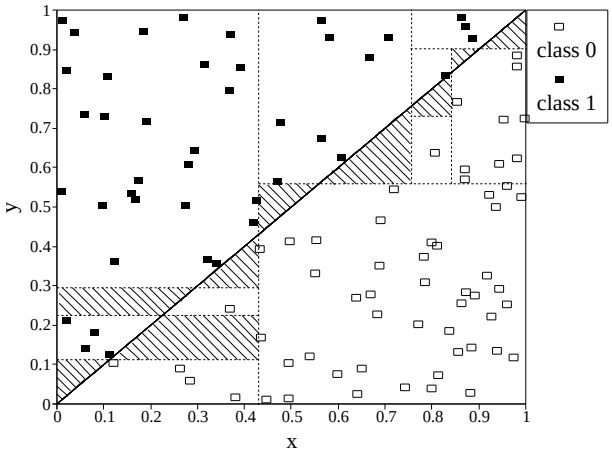

Figure 6. Linear Class Boundaries

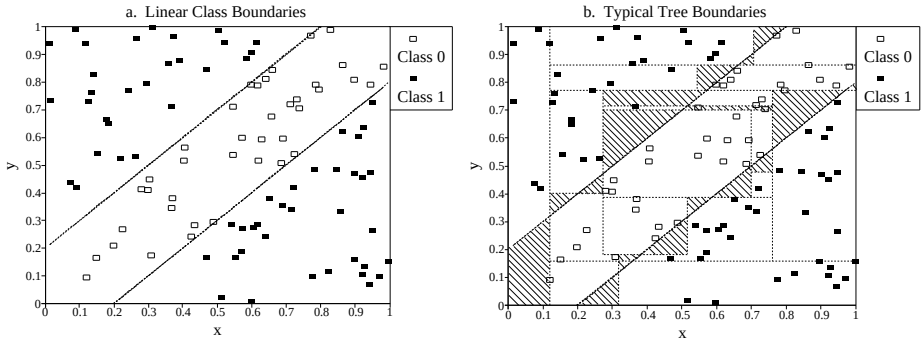

Figure 7. Linear Class Boundaries and XOR

If splits on linear combinations of continuous attributes (e.g., discriminant functions) are allowed, then, for the same sample size, both better accuracy and a simpler tree can be obtained, and trees with accuracy equivalent to single attribute splits can be obtained from smaller samples. See, for instance, Weiss and Indurkhya (1991), Murthy, et al. (1993), Park and Sklansky (1990), and John (1995) for some approaches to handling such linear combinations.

Another feature of this problem is that if only binary splits on a single attribute are allowed, the incremental improvement that could be achieved by any particular split is small. Pre-pruning might preclude splits on continuous attributes in these cases. Again, this is caused by the interaction of pre-pruning with the restrictions on candidate splits, rather than by pre-pruning *per se*.

The choice of a candidate set defines a language for describing the boundaries between classes. If an accurate description of the true class boundaries in this language is very complex (as in XOR or approximating a linear boundary with a step function), then pre-pruning is likely to have a deleterious effect because pre-pruning may prevent discovery of these very complex decision trees.

The point of pruning is to prevent or correct overfitting, the building of trees that are more complex than can be supported by the available data using principles of sound statistical inference. When only very simple candidate splits are allowed, empirical evidence from earlier studies (Breiman, et al., 1984; Quinlan, 1986) indicates that better results are obtained from building overly complex trees and post-pruning than from pre-pruning. The results of our analysis of the XOR and linear boundaries problems indicates that both better accuracy and simpler trees might be obtained by expanding the set of candidate splits. It is not clear whether it is more effective in general to expand the candidate splits and pre-prune, to build more complex trees and post-prune, or to combine the two approaches.

Expanding the set of candidate splits is not a panacea. In the first place, exhaustive search is impractical. Further, when continuous attributes are involved, the set of possible functions combining several attributes is unbounded. It is still necessary to restrict the candidates to relatively simple functions by bounding the number of attributes in a combination and limiting continuous functions to, for instance, linear or quadratic forms.

Expanding the candidate set is not always straightforward. In Figure 7, for instance (class = 1 if $| y - x | > 0.2$, else class = 0), the class boundaries are linear. Linear discriminant analysis (Weiss & Indurkhya, 1991) fails in this case (all of the instances are predicted to be class 1, a 40% error rate) because the simple discriminant analysis assumptions (that each class is adequately described by a single multivariate normal distribution, and that the class means are different) do not hold for these data. This problem (Figure 7) is called the 'parallel oblique lines' problem and is dealt with effectively in the OC1 (Murthy, et al., 1993) algorithm.

In addition to having linear class boundaries, this problem has a trait in common with the XOR problem — diagonally opposite corners of the attribute space have the same class. Ordinary linear discriminant analysis seeks a single line separating two classes, and may fail to find a satisfactory boundary when two lines are required. In this case, the effect of linear discriminant analysis is the same as the effect of pre-pruning in the XOR problem.

In summary, expanding the set of candidate splits is a very powerful tool and can permit discovering decision trees that are both more accurate and less complex. In terms of increasing the number of problems for which reasonably accurate and simple trees can be learned, expanding the set of candidates (within reasonable bounds on the increased search space) is likely to be more effective than is using post-pruning. However, there are no guarantees, and there is no one-size-fits-all strategy for expanding the candidate set. There have also recently been some disquieting results on the effectiveness of lookahead (Murthy & Salzberg, 1995), and on the potentially harmful effects of over-expanding the candidate set, termed 'oversearching' by Quinlan and Cameron-Jones (1995).

11. Conclusions

The following conclusions are drawn from the results and analyses of the experiments performed here:

1. Information gain, gain ratio, distance, orthogonality, chi-squared, and Beta each downplay some part of the influence of the marginal totals of the classes and attribute values. Whenever one or more of the expected values in a split is small, these measures are prone to overestimate the reliability of the split. The divide-and-conquer strategy of building classification trees often leads to very small subtrees where these measures are inadmissible.

2. The P_0 null hypothesis probability measure proposed here overcomes the difficulties encountered when the classes and attribute values are unevenly distributed or the expected frequencies are small. The unpruned trees P_0 builds are typically simpler, more efficient, and no less accurate than those built by the other measures.

3. The ordering of the attribute splits within a tree can profoundly affect the effectiveness and efficiency of pruning (either pre- or post-pruning). This is particularly the case when different heuristics are used for selecting the split and deciding whether to prune or stop splitting.

4. The P_0 measure can be used to stop splitting. This is more practical than post-pruning, particularly C4.5's pessimistic post-pruning routine (Quinlan, 1993), and the resulting trees are typically simpler, more efficient, and no less accurate than unpruned or post-pruned trees. A stopping threshold level which decreases (prunes more severely) as the sample size increases is recommended.

5. The arguments against stopping are equally arguments against use of very sparse (or otherwise ill-conditioned) data, biased heuristics, different inadmissible heuristics for splitting and stopping, and very restricted candidate sets. Assuming a sample size of 50 or more, there is no point in continuing the inductive process when the class distribution is probably independent of the candidate splits ($P_0 > 0.5$), and in most domains there is little point in continuing when $P_0 > 0.05$.

The paper also describes the biases of the various heuristic splitting and stopping metrics and why they are inadmissible. It also largely explains the reasons (other than the well-known XOR problem) for the failure of attempts to formulate a satisfactory stopping strategy using these inadmissible heuristics.

Acknowledgments

The Common Lisp implementation of TDIDT (ID3) used here (substituting various heuristics for information gain and X^2) is due to, and copyright by Dr. Raymond Mooney, University of Texas. Dr. Mooney's copyright notice acknowledges the contributions of others, and gives permission to modify the program provided that the original copyright notice is retained. The comments and suggestions of editor Doug Fisher and several anonymous reviewers are also gratefully acknowledged.

Notes

1. XOR is prototypical of the *myopia* of greedy search, a symptom which may be alleviated by lookahead, as in post-pruning rather than stopping. TDIDT's search strategy is also irrevocable and narrowly focused, problems which may not be alleviated by lookahead and post-pruning.
2. We introduce noise by randomly reversing the class of 1% of the instances, by reversing 1% of the values of A independently of the class noise, and by altering 1% of the values of B independently of the class and attribute A noise, letting $B = 0$ and $B = 2$ change to 1, and $B = 1$ change to either 0 or 2 with equal likelihood.
3. The irrelevant variable is one which is binary and random, and completely independent of the class and of the values of A and B.
4. In this regard, it should be noted that the incomplete Beta function also has a strong relationship to the χ^2, hypergeometric, binomial, Student's t, and F (variance-ratio) distributions. Which is to say that all sensible measures of split utility asymptotically rank attributes in the same order. Hence the repeated empirical findings (e.g., Breiman, et al., 1984; Fayyad & Irani, 1992a) that the various measures are largely interchangeable.
5. If $X_1, \ldots, X_\nu$ are independent random variables, each having a standard (zero mean, unity variance) normal distribution, then $\sum_{i=1}^{\nu} X_i^2$ has a chi-squared (χ^2) distribution with ν degrees of freedom. Here, the $X_i \equiv (f_{cv} - e_{cv})/\sqrt{e_{cv}}$ terms are approximately standard normal *iff* the null hypothesis is true and all of the e_{cv} are large.
6. We say that a statistical procedure is *robust* if the actual significance level is close to the procedure's estimated level, even under deviations from assumptions. An inference procedure is *biased* if its mean deviation from the actual confidence level is not zero. A non-robust biased procedure is *inadmissible*, otherwise, the procedure is admissible.
7. $N = (2, 4, 8, \ldots, 64)$, $n_1 = (1 \ldots N/2)$, $m_1 = (1 \ldots n_1)$, $f_{11} = (0 \ldots m_1)$.
8. See Martin and Hirschberg (1996a) for a thorough analysis of the time complexity of this post-pruning algorithm.

References

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Pacific Grove, CA: Wadsworth & Brooks.

Buntine, W. L. (1990). *A theory of learning classification rules*. PhD thesis. University of Technology, Sydney.

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

Cestnik, B., Kononenko, I. & Bratko, I. (1987). ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In *Progress in Machine Learning, EWSL-87*. Wilmslow: Sigma Press.

Cestnik, B. & Bratko, I. (1991). On estimating probabilities in tree pruning. In *Machine Learning, EWSL-91*. Berlin: Springer-Verlag.

Cochran, W. G. (1952). Some methods of strengthening the common χ^2 tests. *Biometrics, 10*, 417-451.

Elder, J. F. (1995). Heuristic search for model structure. In Fisher, D. & Lenz, H-J. (Eds.) *Learning from Data: Artificial Intelligence and Statistics V, Lecture Notes in Statistics*, v. 112 (pp. 131-142). New York: Springer.

Fayyad, U. M., & Irani, K. B. (1992a). The attribute selection problem in decision tree generation. *Proceedings of the 10th National Conference on Artificial Intelligence, AAAI-92* (pp. 104–110). Cambridge, MA: MIT Press.

Fayyad, U. M., & Irani, K. B. (1992b). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning, 8*, 87-102.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)* (pp. 1022-1027). San Mateo, CA: Morgan Kaufmann.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2*, 139-172.

Fisher, D. H. (1992). Pessimistic and optimistic induction. Technical Report CS-92-22, Department of Computer Science, Vanderbilt University, Nashville, TN.

Fisher, D. H., & Schlimmer, J. C. (1988). Concept simplification and prediction accuracy. *Proceedings of the 5th International Conference on Machine Learning (ML-88)* (pp. 22–28). San Mateo, CA: Morgan-Kaufmann.

Fulton, T., Kasif, S., & Salzberg, S. (1995). Efficient algorithms for finding multi-way splits for decision trees. *Machine Learning: Proceedings of the 12th International Conference (ML-95)* (pp. 244-251). San Francisco: Morgan Kaufmann.

Gluck, M. A., & Corter, J. E. (1985). Information, uncertainty, and the utility of categories. *Proceedings of the 7th Annual Conference of the Cognitive Society* (pp. 283–287). Hillsdale, NJ: Lawrence Erlbaum.

John, G. H. (1995). Robust linear discriminant trees. In Fisher, D. & Lenz, H-J. (Eds.) *Learning from Data: Artificial Intelligence and Statistics V, Lecture Notes in Statistics*, v. 112 (pp. 375-386). New York: Springer.

Kira, K. & Rendell, L. A. (1992). A practical approach to feature selection. *Machine Learning: Proceedings of the 9th International Conference (ML-92)* (pp. 249-256). San Mateo, CA: Morgan Kaufmann.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *Proceedings of the European Conference on Machine Learning (ECML-94)*, (pp. 171-182). Berlin: Springer.

Liu, W. Z., & White, A. P. (1994). The importance of attribute-selection measures in decision tree induction. *Machine Learning, 15*, 25–41.

Lopez de Mantaras, R. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning, 6*, 81–92.

Martin, J. K. (1995). An exact probability metric for decision tree splitting and stopping. Technical Report 95-16, Department of Information & Computer Science, University of California, Irvine, CA.

Martin, J. K. & Hirschberg, D. S. (1996a). On the complexity of learning decision trees. *Proceedings Fourth International Symposium on Artificial Intelligence and Mathematics, AI/MATH-96* (pp. 112-115). Fort Lauderdale, FL.

Martin, J. K. & Hirschberg, D. S. (1996b). Small sample statistics for classification error rates I: Error rate measurements. Technical Report 96-21, Department of Information & Computer Science, University of California, Irvine, CA.

Martin, J. K. & Hirschberg, D. S. (1996c). Small sample statistics for classification error rates II: Confidence intervals and significance tests. Technical Report 96-22, Department of Information & Computer Science, University of California, Irvine, CA.

Mingers, J. (1987). Expert systems — rule induction with statistical data. *Journal of the Operational Research Society, 38*, 39–47.

Mingers, J. (1989a). An empirical comparison of pruning measures for decision tree induction. *Machine Learning, 4*, 227–243.

Mingers, J. (1989b). An empirical comparison of selection measures for decision tree induction. *Machine Learning, 3*, 319–342.

Murphy, P. M., & Aha, D. W. (1995). *UCI Repository of Machine Learning Databases*. (machine-readable data depository). Department of Information & Computer Science, University of California, Irvine, CA.

Murphy P. M. & Pazzani, M. J. (1991). ID2-of-3: Constructive induction of M-of-N concepts for discriminators in decision trees. *Machine Learning: Proceedings of the 8th International Workshop (ML-91)* (pp. 183-187). San Mateo, CA: Morgan Kaufmann.

Murthy, S., Kasif, S., Salzberg, S., & Beigel, R. (1993). OC-1: Randomized induction of oblique decision trees. *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93)* (pp. 322-327). Menlo Park, CA: AAAI Press.

Murthy, S. & Salzberg, S. (1995). Lookahead and pathology in decision tree induction. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)* (pp. 1025-1031). San Mateo, CA: Morgan Kaufmann.

Niblett, T. (1987). Constructing decision trees in noisy domains. In *Progress in Machine Learning, EWSL-87*. Wilmslow: Sigma Press.

Niblett, T., & Bratko, I. (1986). Learning decision rules in noisy domains. In *Proceedings of Expert Systems 86*. Cambridge: Cambridge University Press.

Park, Y. & Sklansky, J. (1990). Automated design of linear tree classifiers. *Pattern Recognition, 23*, 1393-1412.

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

Quinlan, J. R. (1988). Simplifying decision trees. In B. R. Gaines & J. H. Boose (Eds.). *Knowledge Acquisition for Knowledge-Based Systems*. San Diego: Academic Press.

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

Quinlan, J. R. & Cameron-Jones, R.M. (1995). Oversearching and layered search in empirical learning. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)* (pp. 1019-1024). San Mateo, CA: Morgan Kaufmann.

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

Shavlik, J. W., Mooney, R. J., & Towell, G. G. (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine Learning, 6*, 111-143.

Weiss, S. M. & Indurkhya, N. (1991). Reduced complexity rule induction. *Proceedings of the 12th International Joint Conference on Artificial Intelligence, IJCAI-91* (pp. 678–684). San Mateo, CA: Morgan Kaufmann.

Weiss, S. M. & Indurkhya, N. (1994). Small sample decision tree pruning. *Proceedings of the 11th International Conference on Machine Learning (ML-94)*, (pp. 335-342). San Francisco: Morgan-Kaufman.

White, A. P. & Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning, 15*, 321–329.

Received November 8, 1996
Accepted April 28, 1997
Final Manuscript May 7, 1997