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 92717

Technical Report 95-16
May 18, 1995

Abstract

ID3's information gain heuristic is well-known to be biased towards multi-valued attributes. This bias is only partially compensated by the gain ratio used in C4.5. Several other alternatives have been proposed and are examined here (distance, orthogonality, a Beta function, and two chi-squared tests). Gain ratio and orthogonality are strongly correlated, and all of these metrics are biased towards splits with one or more small expected values, under circumstances where the split likely ocurred by chance. Both classical and Bayesian statistics lead to the multiple hypergeometric distribution as the exact posterior probability of the null hypothesis. Both gain and the chi-squared tests are shown to arise in asymptotic approximations to the hypergeometric, revealing similar criteria for admissibility and showing the nature of their biases. Previous failures to find admissible stopping rules in CART and ID3 are traced to coupling these biased approximations with one another or with arbitrary thresholds; problems which are overcome by the hypergeometric. Empirical results show that hypergeometric pre-pruning should be done, as trees pruned in this way are more practical, simpler, more efficient, and generally no less accurate than unpruned or post-pruned trees.

Contents

1 **Introduction and Background** 1

Splitting, Stopping, and Post-Pruning . . . 1

Choosing Candidates/Selecting A Split . . . 1

Inferring Optimal Decision Trees is NP-complete . . . 1

Extending the Candidate Set . . . 1

XOR — Arguments Against Stopping . . . 2

Accuracy/Complexity Tradeoffs . . . 3

2 **Impact of Different Choices of Candidate Split Sets** 3

3 **Impact of Different Choices Among Candidate Splits** 5

4 **Approximate Functions for Selection Among Candidates** 8

5 **Some Troublesome Data Sets** 11

6 **An Exact Test** 12

7 **Stopping Criteria** 15

8 **Empirical Comparisons of the Measures** 17

8.1 Unpruned Decision Trees . . . 18

8.2 Effects of Post-Pruning . . . 19

8.3 Effects of Stopping . . . 20

8.4 Binary *vs.* Multi-way Splits . . . 21

9 **Conclusions** 24

10 **Acknowledgements** 26

A **Data Sets** 27

B **Nominalizing Attributes and Converting to Binary Splits** 28

C **Derivation of the Hypergeometric** 28

D **Approximations to the Hypergeometric** 30

D.1 Gain and Chi-squared . . . 30

D.2 Buntine's Beta . . . 31

E Detailed Experimental Results 33
E.1 Unpruned Trees, Binary Splits 33
E.2 Post-pruned Trees, Binary Splits 37
E.3 Pre-Pruned Trees, Binary Splits 41
E.4 X^2-Stopped Trees, Binary Splits 45
E.5 Comparisons of Stopping and Post-Pruning 49
E.6 Multi-way Splits 51

List of Figures

1 Linear Class Boundaries 4
2 Linear Class Boundaries and XOR 5
3 Alternative Splits 6
4 Comparison of Measures 14
5 Alternative Split/Stop Strategies 17
6 An XOR-like Substructure 21
7 Stopping *vs.* Post-Pruning 24

List of Tables

1 Effects of Split Order 7
2 A Troublesome Data Set 11
3 Unpruned Trees, Binary Splits 19
4 Post-Pruned Trees, Binary Splits 20
5 Effects of Stopping, Binary Splits 22
6 Overall Effects of Pre- and Post-Pruning 23
7 Binary *vs.* Multi-way Splits 25
8 Description of Data Sets 27
9 Accuracy of Unpruned Trees 33
10 No. of Leaves of Unpruned Trees 34
11 Wtd. Avg. Depth of Unpruned Trees 35
12 Run Time of Unpruned Trees 36
13 Accuracy of Post-Pruned Trees 37
14 No. of Leaves of Post-Pruned Trees 38
15 Wtd. Avg. Depth of Post-Pruned Trees 39
16 Run Time of Post-Pruned Trees 40

17 Accuracy of Pre-Pruned Trees, Binary Splits 41
18 Number of Leaves – Pre-Pruned Trees, Binary Splits 42
19 Wtd. Avg. Depth of Pre-Pruned Trees, Binary Splits 43
20 Run Time of Pre-Pruned Trees, Binary Splits 44
21 Accuracy of X^2-Stopped Trees, Binary Splits 45
22 No. of Leaves of X^2-Stopped Trees 46
23 Wtd. Avg. Depth of X^2-Stopped Trees 47
24 Run Time of X^2-Stopped Trees 48
25 Wtd. Avg. Accuracy Comparisons 49
26 Total No. of Leaves Comparisons 49
27 Wtd. Avg. Depth Comparisons 50
28 Total Run Time Comparisons 50
29 Accuracy & Complexity of Multi-way Splits 51
30 Efficiency & Practicality of Multi-way Splits 52

Glossary

C — number of classes, 9

$(1 - d)$ — Lopez de Mantaras' distance, 9

e_{cv} — expected cell frequency, 10

f_{cv} — cell frequency, 9

G^2 — Wilks' χ^2, 10

m_v — subset size, 9

N — sample size, 9

n_c — class size, 9

P_0 — hypergeometric probability, 13

V — number of subsets, 9

$W(\alpha)$ — Buntine's Beta, 9

X^2 — Pearson's χ^2, 10

χ^2 probability distribution 10

***a priori* cut-points** 17

accuracy 3

admissibility 10

Beta 9

biased test 10

CART — Classification and Regression Trees, 5

Cochran's rule admissibility of χ^2, 10

complexity 3

confidence level 12

contingency table 9

dimensionality 19

disparity measures 8

distance $(1-d)$ 9

efficiency 3

ex post **cut-points** 17

Fisher's exact test 12

gain ratio 9

gain 9

hypergeometric 12

ill-conditioned data 15

impurity measures 8

information gain 9

irrelevant attribute 6

likelihood ratio 10

MDL — minimum description length, 10

multiple hypergeometric 13

Mutual Information — information gain, 12

noise 6

NP-complete 1

null hypothesis 10

orthogonality 9

overfitting 4

post-pruning 1

power of a test 12

practicality 3

pre-pruning 1

redundant attribute 6

reliability measures 8

robustness 10

significance level 12

stopping 1

TDIDT — Top-Down Induction of Decision Trees, 1

Type I error 16

Type II error 16

utility measures 8

1 Introduction and Background

Top-Down Induction of Decision Trees (TDIDT [55]) is a family of algorithms for inferring classification rules (in the form of a decision tree) from a set of examples. The goals are varied, and criteria for judging success are equally varied. Motivations for learning intensional descriptions such as decision trees include:

- efficiency or practicality – the population may be infinite or impractically large, with excessive storage or look-up time required
- generalization – if all members of the population are not available for study or if the population is noisy or has continuous attributes, it may be necessary or desirable to interpolate or extrapolate from the available instances
- comprehensibility – to reduce a large set of observations of some phenomenon to a more comprehensible form, such as the phase diagram of a material, which depicts the conditions under which the material is solid, liquid, or gaseous, and the boundary conditions under which two or all three states may co-exist

TDIDT algorithms make a greedy choice of one of a set of candidate splits (decision nodes) for a data set and recursively partition 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 TDIDT algorithms have included additional criteria to stop splitting when the incremental improvement is deemed insignificant. These stopping criteria are collectively referred to as pre-pruning. Other algorithms have added recursive procedures for post-pruning (replacing one or more of the splits with a terminal node). Most procedures described as post-pruning go beyond mere pruning, replacing a split with some other split (typically a child of the replaced node).

At each decision node, split selection is addressed as two separate but interdependent subproblems:

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

Split selection is addressed as two separate subproblems because the problem of inferring optimal decision trees is NP-complete [32]. In more concrete terms, there are more than 10^{13} distinct ways to partition a set containing only 20 items and, thus, more than 10^{13} possible decision trees (for N items, the number of possible distinct partitions, T, is bounded by $(N/2)! \leq T \leq N!$). Practical algorithms can explore only a small portion of such a vast solution 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 operators (possible splits). TDIDT builds complex trees by recursive refinement of simpler trees (*i.e.,* it explores only relatively simple splits at each decision node).

The earliest TDIDT algorithms restricted candidates to splits on the values of a single attribute and only binary splits for continuous attributes. More recent algorithms extend the candidate space in various ways, including

1. multi-way splits for continuous attributes [22]
2. combinations of two discrete attributes [18]

3. combinations of continuous attributes, such as linear discriminant functions [63, 64]
4. m-of-n boolean functions [50] (a function of n boolean attributes that is true if any m or more of the attributes are true)
5. combinations of several trees built by stochastically choosing among the best few candidate splits at each decision node in each of the trees [3]

The process of 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 of the two phases of split selection with one another, with the context created by earlier choices, and with 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 the poor or good performance of an algorithm 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 of 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.

This example is often cited as an argument against pre-pruning. The difficulty actually is the result of the interaction of pre-pruning and allowing only single-attribute splits, and one could equally well 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 exclusive-or. There are, however, at least two approaches which would lead to discovering the 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.

For a given set of candidates, pre-pruning results in shorter learning times but precludes exploring a part of the potential solution space. Post-pruning results in longer learning times but explores all the potential solution space that the set of candidates and greedy search will permit. Since the point of restricting the set of candidates (which typically precludes a larger set of potential trees than does pre-pruning) is to reduce the learning times to a practicable level, it seems hard to justify an argument against pre-pruning solely because it precludes a subset of potential trees.

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 exclusive-or difficulty) of the poor performance of pre-pruning in early empirical studies [9, 55].

In light of the foregoing discussions, the design of experimental studies of different heuristic functions must be guided by two principles:

1. Care must be taken to ensure that only the heuristic functions are being changed, or that changes in other factors are systematic and that the effects of various factors are not confounded in the design. It is particularly important to control the set of candidate splits.

2. For any given combination of a heuristic function and a procedure for choosing a candidate set, one can always find a set of problems on which the combination will perform poorly (or well). Identifying these is important, but evaluation of algorithms or heuristics intended for general use should not be limited to studying only these problems. It is important to make the evaluations on a wide variety of problems. The evaluation data sets should be chosen *a priori*, and should include differences in the following factors:

   (a) data set size

   (b) complexity of the inferred decision trees

   (c) the number, kinds, and arities of attributes

   (d) degree of noise in the data

   (e) application domains

Evaluation criteria involve issues of accuracy, complexity, and accuracy/complexity tradeoffs. There is no single measure which combines these in an appropriate manner for every application (see [43] for a review of these issues). Measures of complexity include the number of leaves and their average depth (weighted according to the fraction of the population covered by each leaf), and the run-time of the learning algorithm. To distinguish between these, the following terminology will be used:

complexity – number of leaves

efficiency – weighted average depth

practicality – tree building, pruning, and cross-validation CPU time

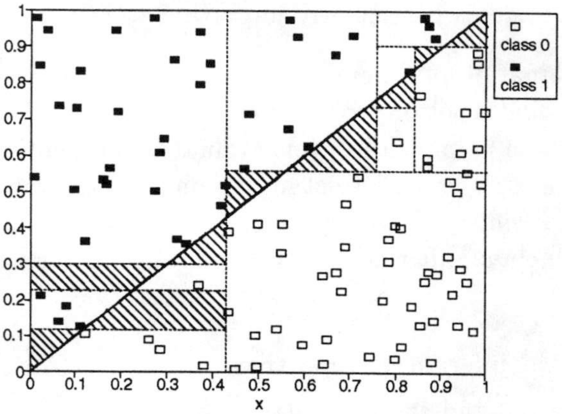
Within rather broad bounds of complexity, the dominant goal is usually to infer trees where the population instances covered by each leaf are, as nearly as possible, members of the same class. If each leaf is labeled with some predicted class, the accuracy of the leaf is defined as the percentage of 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. Either cross-validation or bootstrapping should be used to estimate accuracy and confidence intervals (see [44] for a review of these techniques).

2 Impact of Different Choices of Candidate Split Sets

We have seen in the exclusive-or 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 1 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 classes is linear (class $= 1$ if $y > x$, else class $= 0$). If the candidate splits are restricted to splitting on a single attribute, the leaves of the decision tree each cover a rectangular area with

Figure 1: Linear Class Boundaries



sides parallel to the axes. The boundary between the classes can only 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 1 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.

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.

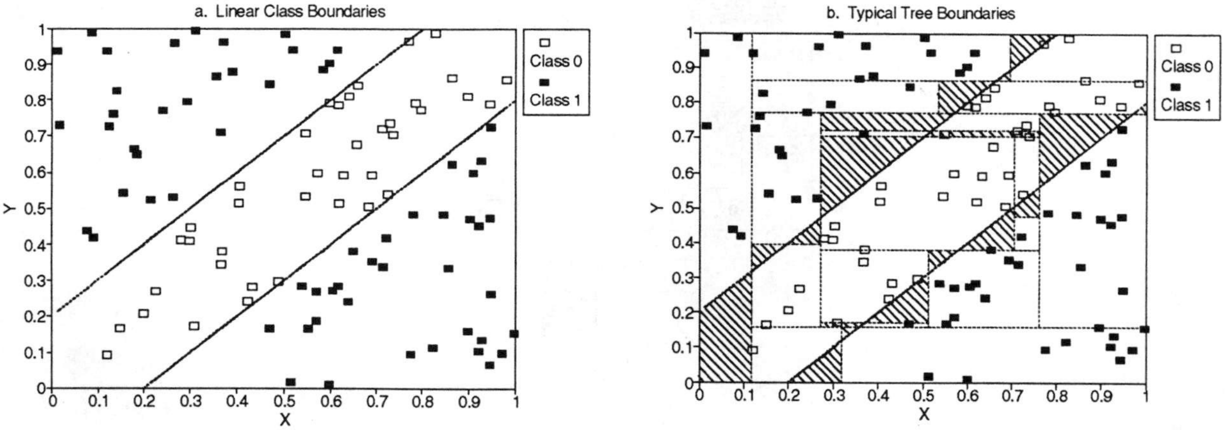
Another feature of this problem is that if only binary splits on a single attribute are allowed, the incremental improvement that could be achieved by any particular split is very small. Pre-pruning might preclude making any binary splits on continuous attributes in these cases. Again, this is caused 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 the very simplest kinds of candidate splits are allowed, the empirical evidence from earlier studies of pre-pruning [9, 55] 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 exclusive-or and linear boundaries problems seem to indicate that both better accuracy and simpler trees can 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, unrestricted search is NP-complete. 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

Figure 2: Linear Class Boundaries and XOR



functions by bounding the number of attributes in a combination and limiting continuous functions to, for instance, linear or quadratic forms.

Expanding the set of candidates is not always straightforward. In Figure 2, for instance, the class boundaries are linear (class = 1 if $| y - x |> 0.2$, else class = 0). Linear discriminant analysis [36, 63, 64] 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 homogeneous and can be adequately described by a single multivariate normal distribution, and that the means of the classes are different) do not hold for these data.

In addition to having linear class boundaries, the problem shown in Figure 2 has a trait in common with the exclusive-or 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 exclusive-or problem.

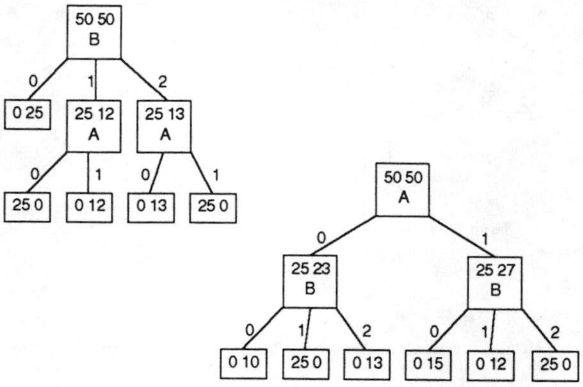
In summary, expanding the set of candidate splits is a very powerful tool and can permit discovering decision trees that are both more accurate and less complex. In terms of increasing the number of problems for which reasonably accurate and simple trees can be learned, expanding the set of candidates (within reasonable bounds on the increased search space) is likely to be more effective than is using post-pruning rather than pre-pruning. However, there are no guarantees, and there is no one-size-fits-all strategy for how to expand the candidate set.

3 Impact of Different Choices Among Candidate Splits

Figure 3 shows two different decision trees for the same data set, choosing a different split at the root of the tree. In this case, the accuracy of the two trees is the same (100%), but one of the trees is more complex and less efficient than the other. This problem has the characteristics that the set of candidate splits is sufficient to fully separate the classes (CART — Classification and Regression Trees terms this a complete data set [9]) and that 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.

Figure 3: Alternative Splits



If we introduce noise into the population[1] then, for 100 samples of size 100 randomly drawn from this noisy population, the average results of splitting on A first versus splitting on B first are shown in Table 1a. Here also, the difference between the alternative split orderings is a matter of complexity and efficiency, not accuracy.

Returning to the noise-free population, if we add an irrelevant attribute[2] X and split on A first then B, or on B first then A, we get the same trees shown in Figure 3 (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, 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 another noisy version of the original 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 1%. The effects of splitting on A, B, and Z in various orders are shown in Table 1c. When attribute Z is more nearly irrelevant, the order of the attribute splits is largely of matter of complexity and efficiency, rather than accuracy. As Z becomes more relevant, but redundant, 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 impacted when redundant attributes are noisy and strongly correlated. Insofar as the accuracy of

[1] We introduce noise by randomly reversing the class of 1% of the instances; by reversing the value of A in 1% of the instances independently of the class noise; and by altering the value of B in 1% of the instances independely 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.

[2] One which is binary and random, completely independent of the class and of the values of A and B.

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%	8.72	5.36	1.84

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	1.75
X, A, B	0	19	12	3
X, B, A	0	17	10	2.75
$X, AB/BA$	0	18	11	2.88
$X, BA/AB$	0	18	11	2.88
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

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 [9, 21] that various heuristic functions for choosing among candidate splits are largely interchangeable.

It is important not to overemphasize differences in complexity and efficiency, in the sense 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 [49], 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 severly restricted the set of candidate splits for the sole purpose of reducing the complexity of an NP-complete problem to manageable proportions. Differences in complexity and efficiency may be greatly magnified as the set of candidate splits is expanded.

Liu and White [41] discuss the importance of discriminating between attributes which are truly 'informative' and those which are not. The examples in Figure 3 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, simple pruning would not correct the overfitting (it would, in fact, be very harmful). The pruning strategy used in the C4.5 [58] 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, 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.

Thus, the following three criteria should be considered in choosing a split selection heuristic:

1. it should prefer splits which most improve the final tree's accuracy and avoid or minimize the impact of those which are harmful to accuracy

2. for splits leading to equivalent accuracies, it should prefer splits which lead to simpler and more efficient trees

3. it should order the splits so as to permit practical pruning

4 Approximate 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 [9, pp. 93-98]), the most telling being that it simply has not worked out well empirically.

Various other measures of a split's utility have been proposed. Virtually all of these utility measures agree as to the extreme points (*i.e.,* that a split in which the class 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 (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 in some function of the class proportions (such as entropy) between the parent and the split subsets. These measures emphasize the purity of the subsets, and CART [9] terms these *impurity* functions.

2. Measures of the difference in some function of the class proportions (typically a distance or an angle) between the split subsets. 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 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 [21] cites several studies showing that various impurity measures are largely interchangeable, *i.e.*, that they result in very similar decision trees, and CART [9] finds that the final (unpruned) tree's properties are largely insensitive to the choice of a splitting rule (utility measure).

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 sample
n_c	is the total no. in class c

Variants of the information gain heuristic used in ID3 [55] have become the *de facto* standard metrics for TDIDT split selection. This heuristic, or various modifications of it, is used (for instance) in FOIL [57], FOCL [53], CART [9], CN2 [15], GID3(*) [20], and C4.5 [58]. The information gain function calculates the difference (decrease) between the entropy of the population and the weighted average entropy of the subpopulations.

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

The gain ratio function used in C4.5 [58] partially compensates for the known bias of gain towards splits having more subsets (larger V).

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

Lopez de Mantaras [42] proposes a different normalization, a distance metric $(1-d)$

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

Fayyad, *et al* [21] give an orthogonality (angular disparity) measure for binary attributes

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

where θ is the angle between the class frequency vectors, f_{c1} and f_{c2}.

Buntine [11] 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 typically either 0.5 or 1.0, and describes the assumed prior distribution for the contingency table cells. Information gain appears as part of an asymptotic approximation to this function. In this regard, it should be noted (see [1, pp. 944-5]) that the incomplete Beta function also has a strong relationship to χ^2, the hypergeometric, the binomial, Student's t, and the F (variance-ratio) distributions. Which is to say that all sensible measures of split utility asymptotically converge (rank attributes in the same order). Hence the repeated empirical findings that the various measures are largely interchangeable.

In addition to the above heuristics from the machine learning literature, the analysis of categorical data has long been studied by statisticians. (See Agresti [2] for a thorough review of this field.)

The Chi-squared statistic [4, pp. 452-462], [35, pp. 320-323], [60, pp. 572-592], [2, pp. 47-48]

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

is distributed *approximately* as χ^2 with $(C-1) \times (V-1)$ degrees of freedom†. 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[3] is a good approximation (is admissible‡) when all of the e_{cv} are greater than 1 and no more than 20% are less than 5 (Cochran's rule [16, 17]).

> † 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 [1, pp. 940-943]. 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.
>
> ‡ A statistical procedure is *robust* if the actual significance level is close to the procedure's estimated level, even under deviations from assumptions [60, p. 321]. An inference procedure is *biased* if its expected (average) deviation from the actual confidence level is not zero. A biased, non-robust procedure is *inadmissible*.

The Likelihood-Ratio Chi-squared statistic [2, pp. 48-49]

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

$$\text{where} \qquad \Lambda = \left(\prod_{c=1}^{C}\prod_{v=1}^{V}(n_c m_v)^{f_{cv}}\right) \Bigg/ \left(N^N \prod_{c=1}^{C}\prod_{v=1}^{V} f_{cv}^{f_{cv}}\right) \qquad \text{is the likelihood ratio}$$

is also distributed *approximately* as χ^2 with $(C-1) \times (V-1)$ degrees of freedom. The asymptotic convergence of the G^2 statistic[4] is slower than that of X^2, and the χ^2 approximation to G^2 is usually poor when $N < 5\ CV$ [37, 38, 40].

Replacing e_{cv} by $(n_c\ m_v/N)$ in Equation 7 and rearranging gives $G^2 = 2\ln(2)N$ gain. In the arguments supporting adoption of information gain, minimum description length (MDL), and general entropy-based heuristics [19, 55, 59], [31, pp. 178-181,216-223], the product of the parent subset 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 a conventional maximum likelihood analysis, and message length compression has a limiting χ^2 distribution that may converge less quickly than the more familiar X^2 test. Mingers [45] discusses the G^2 metric (denoted there as G), and White and Liu [65] recommend that the χ^2 approximation to either G^2 or X^2 be used in preference to gain, gain ratio, *etc.*

[3] Proposed by Karl Pearson in 1900 [54], and clarified by R. A. Fisher in 1922 [23].

[4] Proposed by S. S. Wilks in 1935-1938 [66, 67].

Table 2: A Troublesome Data Set

Cat	Atr A		Atr B		Atr C		Atr D		Atr E		Atr F		Atr G		Atr 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 §	X^2	G^2	W(1) $\div N$	P_0
A	¶ 0.4	.022	.193	.021	.502	1.69	1.99	.687	.375
B	4.9	.020	.028	.012	.078	1.86	1.81	.693	.101
C	3.8	.009	.014	.006	.041	.79	.77	.700	.185
D	4.9	.017	.024	.010	.051	1.45	1.53	.697	.131
E	10.5	.019	.019	.010	.045	1.69	1.69	.697	.090
F	4.1	.018	.027	.011	.079	1.65	1.60	.694	.119
G	6.8	.019	.022	.010	.056	1.67	1.64	.696	.099
H	11.3	.015	.015	.008	.035	1.36	1.37	.700	.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 §	X^2	G^2	W(1)	P_0
A	1	1	1	1	2.1	1	1	8
B	2.0	7.4	5.0	7.3	1	2.0	4.4	1.3
C	8	8	8	7.9	8	8	7.8	3.3
D	3.7	7.6	5.8	7.7	3.7	3.6	6.5	2.0
E	2.7	7.8	6.1	7.9	2.1	2.7	6.6	1
F	3.2	7.5	5.4	7.3	2.4	3.2	4.8	1.7
G	3.0	7.7	5.9	7.7	2.2	3.0	5.9	1.2
H	4.6	7.9	6.9	8	4.3	4.6	8	1.4

§ Note the strong correlation (0.998) of gain ratio and ORT

5 Some Troublesome Data Sets

Consider a data set which produces the trial splits shown in Table 2. Information gain chooses attribute A for the first split. 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 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, evidence that these measures also suffer (even more) from this hyper-sensitivity. Mingers [47] 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. There are two arguments for postponing use of this attribute until late in building the tree:

1. it is inefficient to place it at the root — every new instance to be classified must be subjected to this test, which is irrelevant for most instances
2. it may not be necessary to use the attribute test at all, if the information it conveys is completely subsumed by some other split

Church, *et al* [14] use an information gain statistic, *Mutual Information*, to select which of many co-occurring words are likely to be useful features for disambiguating word senses in natural language. In this context the bias of entropy-based statistics is often catastrophic, being expressed as a strong preference for one-in-a-million chance co-occurrences. (A word which occurs only once in a million-word text happens to occur near one instance of the target word and is chosen in preference to more pertinent words which co-occur frequently. In particular, this leads to a predilection for choosing proper nouns to disambiguate the senses of a verb — the fact that some proper noun occurs might be pertinent, but the particular name certainly is not.)

6 An Exact Test

> "...the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data." — *R. A. Fisher* (1925)
>
> "To-day exact tests of significance need no apology. ...In most cases the new methods actually simplify the handling of the data. The conservatism of some university courses in elementary statistics, in stereotyping unnecessary approximations and inappropriate conventions, still hinders many students in the use of exact methods. ...departures from tradition have not been made capriciously, but only when they have been found to be definitely helpful." — *R. A. Fisher* (1954) [25]

Fisher's Exact Test for 2×2 contingency tables (sometimes called the Fisher-Irwin Test)[5] [2, pp. 59-62], [35, pp. 332-337], [60, pp. 586-592] 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).

$$P_0 \equiv \binom{n_1}{f_{11}}\binom{n_2}{f_{21}} \Big/ \binom{N}{m_1} \tag{8}$$

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. Tocher [62] shows that Fisher's test is uniformly most powerful unbiased, *i.e.*, that 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 [60, p. 290] — see Hodges [35, pp. 393-400] on the role of uniformly most powerful tests in choosing a test statistic).

[5] Proposed independently by R. A. Fisher [24, 25] and J. O. Irwin [33] in 1935.

White and Liu [65] note that, for small e_{cv}, Fisher's exact test should be used in place of the χ^2 approximation, and suggest that a similar test for larger tables could be developed. The extension of Fisher's exact test for contingency tables larger than 2×2 is given by the multiple hypergeometric distribution [2, pp. 62-64] [26]

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

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 [2, pp. 62-64], or from Bayesian statistics, as the probability that the null hypothesis is true given the observed data (see Appendix C).

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 a given split or any more extreme split, given that the null hypothesis is true. (This is consistent with Minger's [45] suggested use of G^2).

The following approximate relationships can be derived (see Appendix D):

$$\begin{aligned} 2\ln(2)\,N\text{ gain} &\approx -2\ln(P_0) - (C-1)(V-1)\ln(2\pi N) \\ &\quad + (\text{terms increasing as the interaction weighted sum of squares}) \qquad (10) \\ X^2 &\approx -2\ln(P_0) - (C-1)(V-1)\ln(2\pi N) \\ &\quad + (\text{terms increasing as the main-effects sum of squares}) \qquad (11) \end{aligned}$$

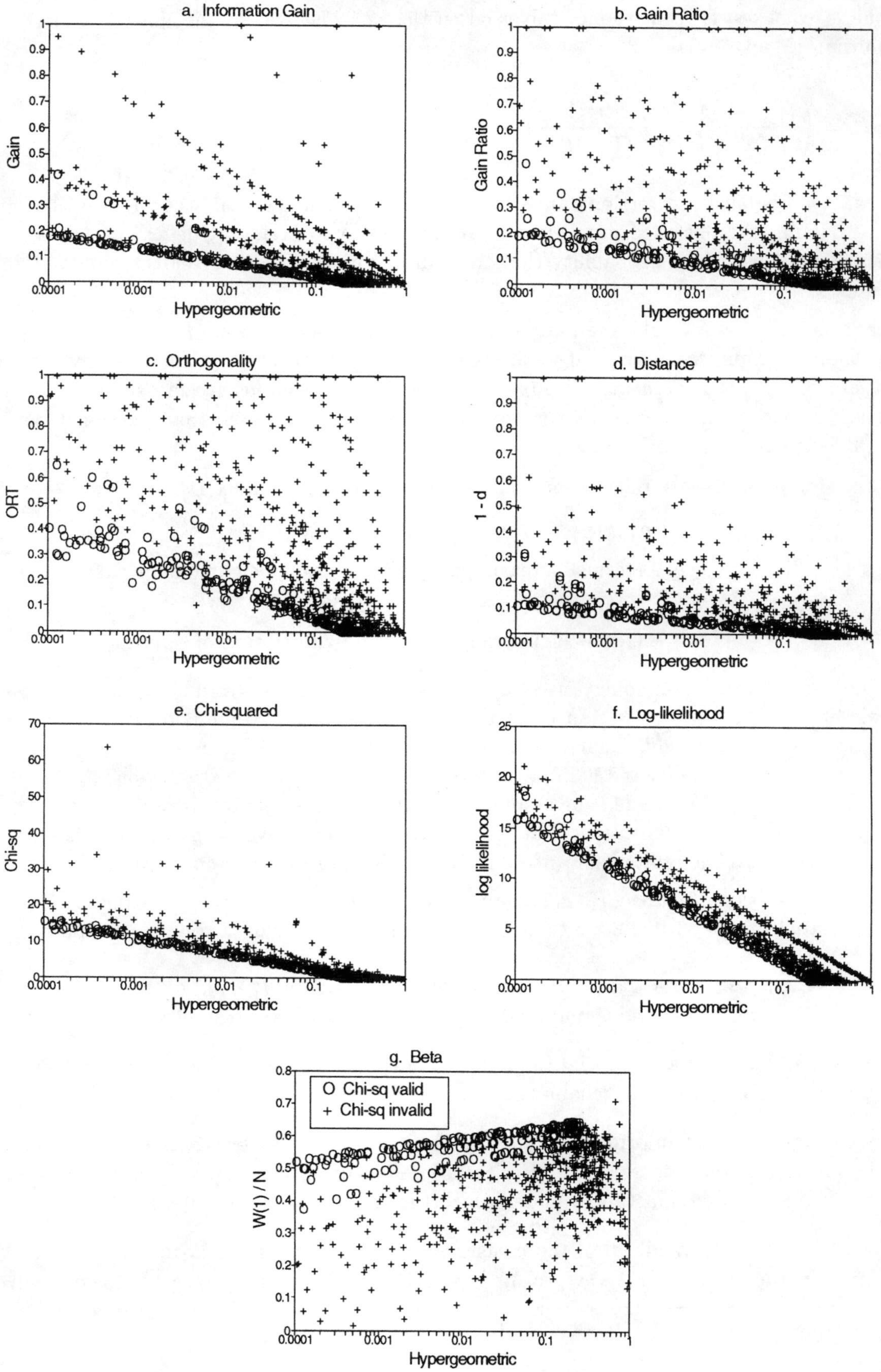
(The sum of squares terminology used here arises in analysis of variance (ANOVA) (see [48]) — main-effects refers to the variances of the marginal row and column totals, m_v and n_c, and interaction refers to the additional variance of the f_{cv} terms over that imposed by the m_v and n_c totals). Thus, both X^2, G^2, and gain arise as terms in alternative approximations to the reliability of a split. In neither case should it be assumed that all the remaining 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 relationship of P_0 to Buntine's Beta is discussed in Appendix D.

Values of each of the measures (P_0, gain, gain ratio, distance, orthogonality, chi-squared, and Beta) were calculated for over 1,000 2×2 tables[6] (see Figure 4). These data confirm the analyses above:

- when Cochran's criteria are satisfied, $X^2 \approx -2.927 - 2\ \ln(P_0)$ and $G^2 \approx X^2$; when they are not, X^2 and G^2 tend to be spuriously high, and overestimate reiability
- a similar linear relation to $\ln(P_0)$ is found for the other measures, with an even stronger tendency to overestimate 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
- when $X^2 \approx \chi^2$ 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

[6] $N = (2, 4, 8, \ldots, 64)$, $n_1 = (1 \ldots N/2)$, $m_1 = (1 \ldots n_1)$, $f_{11} = (0 \ldots m_1)$.

Figure 4: Comparison of Measures



a. Information Gain

b. Gain Ratio

c. Orthogonality

d. Distance

e. Chi-squared

f. Log-likelihood

g. Beta

- applying logarithmic or trigonometric transformation functions to discrete data can give very misleading results, particularly when any of the integer quantities involved is small

Observation 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.)

Observation 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:

$$\begin{array}{rcl} \textit{class} &=& (\textit{pos, neg}) \qquad \textit{color} \;=\; (\textit{blue, red}) \qquad \textit{size} \;=\; (\textit{large, small}) \\ \textit{Let}\;\; \alpha(i,j) &=& \textit{Prob}\{\textit{class} = i \mid \textit{color} = j\} \qquad \gamma(j) = \textit{Prob}\{\textit{color} = j\} \\ \beta(i,k) &=& \textit{Prob}\{\textit{class} = i \mid \textit{size} = k\} \qquad \delta(k) = \textit{Prob}\{\textit{size} = k\} \\ \textit{and}\;\; \theta(i) &=& \textit{Prob}\{\textit{class} = i\} \\ \textit{Now,}\;\; \theta(i) &=& \alpha(i, \textit{blue}) \times \gamma(\textit{blue}) + \alpha(i, \textit{red}) \times \gamma(\textit{red}) \\ &=& \beta(i, \textit{small}) \times \delta(\textit{small}) + \beta(i, \textit{large}) \times \delta(\textit{large}) \end{array} \tag{12}$$

Because of Equation 12, the statements

$$\alpha(i, \textit{blue}) \textit{ is independent of } \alpha(i, \textit{red}) \qquad \textit{and} \qquad \beta(i, \textit{small}) \textit{ is independent of } \beta(i, \textit{large})$$

cannot both be true of the same population.

Observation 3 *The null hypothesis probability function P_0 appears to be a measure which properly incorporates all these factors, and may be a more suitable split selection metric than gain, gain ratio, distance, orthogonality, Beta, or chi-squared.*

7 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* [9] initially searched for a minimum gain threshold to prevent overfitting. Since ($N \times$ gain) has approximately a χ^2 distribution (which has very complex thresholds), attempting to find a simple threshold for gain was foreordained to fail.

Quinlan [55] 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 exclusive-or problem, there are two reasons that this strategy does not work well:

1. for splits with small e_{cv} components the χ^2 approximation to X^2 is not valid, and should not be used (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 (where 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 efficient pruning by X^2

Both of these approaches were abandoned in favor of some form of post-pruning (either a cost-complexity [9, pp 65-81], reduced-error [56], or pessimistic pruning [58] approach). There have been a number of studies in this area [10, 12, 46, 47, 51, 52, 61]. 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 [55]
- 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). There are no certainties in statistical inference, at best there is a balancing of the risks and costs of various inferential errors.

Observation 4 *The previous negative results concerning pre-pruning 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.*

Consider, for example, the following potential splits:

Attribute A

	$A = 0$	$A = 1$	Total
Class N	27	473	500
Class P	43	457	500
Total	70	930	1000

info. gain=.00286 gain ratio=.00782
$1 - d$=.00209 ORT=.02912
G^2=2.123 X^2=3.932
$W(1)/N$=.69629 P_0=.0139

Attribute B

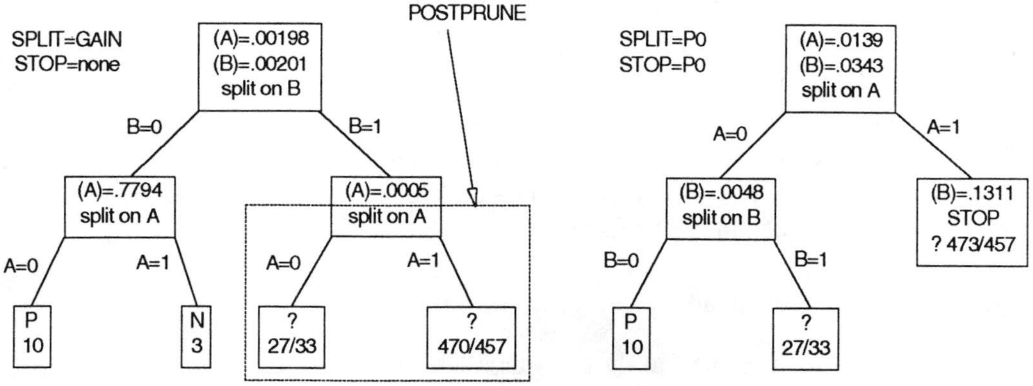
	$B = 0$	$B = 1$	Total
Class N	3	497	500
Class P	10	490	500
Total	13	987	1000

info. gain=.00290 gain ratio=.02902
$1 - d$=.00265 ORT=.12291
G^2=2.386 X^2=3.819
$W(1)/N$=.69563 P_0=.0343

Attribute B has the larger gain (and gain ratio, *etc.*) X^2 for this split is slightly below the 95% cut-off for χ^2. Both splits are, in fact, significant at the 95% level, and A is the better choice.

Splitting on gain and stopping on X^2 stops without generating any tree. Splitting on gain and post-pruning leads to (see Figure 5)

Figure 5: Alternative Split/Stop Strategies



$$[(B = 0) \wedge (A = 0) \Rightarrow (\text{Class } P)] \ \wedge \ [(B = 0) \wedge (A = 1) \Rightarrow (\text{Class } N)]$$

Splitting and stopping using P_0 leads to the more general rule $[(A = 0) \wedge (B = 0) \Rightarrow (\text{Class } P)]$ directly, without generating and later pruning a subtree under $(B = 1)$.

8 Empirical Comparisons of the Measures

Sixteen data sets were used to evaluate the generality of these results. The particular data sets are described briefly in Appendix A, and were chosen to give a good variety of application domains, sample sizes, and attribute properties. None of the data sets chosen has any missing values. Two issues arise with respect to handling the attributes:

- Numeric attributes must be nominalized (made discrete). 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* (also considering the classification)
  4. multi-valued *vs.* binary cuts
  5. the function used to evaluate potential cut-points

  The particular method used has important consequences for both efficiency and predictive accuracy, and (especially for the on-line re-evaluation approaches) 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 [58] describes a procedure for iteratively merging branches of a split using gain or gain ratio; this procedure could either be pursued until only a 'best' binary split remained, or stopped at some threshold and the resulting Q-way split converted to Q binary attributes. Other procedures are given in [9, 13].

The hypergeometric function (and, strictly, the other measures, as well) applies only when the cut-points for continuous attributes and the binarization of discrete attributes are defined *a priori*. Defining these *ex post*, as in C4.5 [58, pp. 25-26] and CART [9, p. 108], *etc.* directly contradicts the null hypothesis (that the class is *a priori* independent of the subset membership). The modifications to the expression for P_0 necessary to accomodate *ex post* cut-points and binarization, and full consideration of the efficacy of various strategies for handling numeric and multivalued attributes are planned topics for a future paper. 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 for every numeric attribute in every data set.

1. 'natural' cut-points determined by the procedure in Appendix A
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 average using the 'natural' cut-points was approximately 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. For binarization, all attributes having $V > 2$ values were replaced with V binary attributes. In each experiment, a tree was grown using all of the instances, and the complexity and efficiency of this tree were determined. Accuracy was then estimated by 10-fold cross-validation.

8.1 Unpruned Decision Trees

The results for the unpruned trees built using the various metrics are summarized in Table 3. (Only the data sets where there were noticeable differences in accuracy are shown in detail, full results are shown in Appendix E.1). Two different values are shown for the Beta metric's α parameter; 1, corresponding to a uniform prior distribution, and 0.5, the Jeffreys prior (see [30, pp. 48-50,79]). The G^2 and X^2 trees were built without regard to significance or admissibility.

None of the differences in accuracy between the split metrics is statistically significant. The differences between the arbitrary and 'natural' nominalizations is generally very small (but see the Glass and WAIS data sets), and sometimes positive, sometimes negative. The average accuracy for orthogonality is slightly lower than the accuracies of the other metrics (the difference is *not* significant at the 95% level).

The trees have about the same number of leaves on the average. The G^2 and X^2 trees have the fewest leaves (12% fewer than P_0 on the average), and orthogonality, gain ratio, gain, and the Beta function have the most leaves (6% more than P_0 on the average).

The gain ratio and orthogonality trees are 60% deeper (less efficient), and the P_0 trees 25% more efficient on the average. With arbitrary subsets, all of the measures build shallower trees with more leaves, though the change in depth for P_0 is small. The quartile trees are all about the same depth. These results reflect the fact that a classifier must be more complex to deal with arbitrary division into subsets, and the tendency for all the metrics except P_0 to be 'fooled' into using the very small splits present in the natural subsets data.

P_0 is more practical in virtually every case, reducing training time by 30% on the average over the nearest competitor (X^2) and by 60% over the least practical (gain ratio). With a single exception

Table 3: Unpruned Trees, Binary Splits

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
Cross-Validation Accuracy, %									
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

(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.

These data support the conjecture that in virtually every case unpruned trees grown using P_0 are less complex, more efficient and practical, 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.

8.2 Effects of Post-Pruning

Quinlan's pessimistic post-pruning method was used (C4.5 [58, pp. 35-43,159-163,278-279]), at the default 0.25 confidence factor level. The results are summarized in Table 4 (full details are given in Appendix E.2). Some of the noteworthy features of these data are:

1. There are *no* significant differences in accuracy between the unpruned and post-pruned trees, nor between the various metrics. That is to say, the choice of a splitting heuristic and the decision whether or not to post-prune are largely matters of complexity, efficiency, and practicality, not of accuracy.

2. The differences between metrics in the number of leaves and depth of the trees, though still present after post-pruning, are much smaller. That is, the metrics that overfit most (notably gain ratio and orthogonality) benefit most from post-pruning, though some overfitting remains after post-pruning.

Table 4: Post-Pruned Trees, Binary Splits

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
Cross-Validation Accuracy, %									
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

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 is largely wasted effort.

4. Post-pruning is very expensive, and not cost-effective. Simpler, more efficient, and equally accurate trees 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.

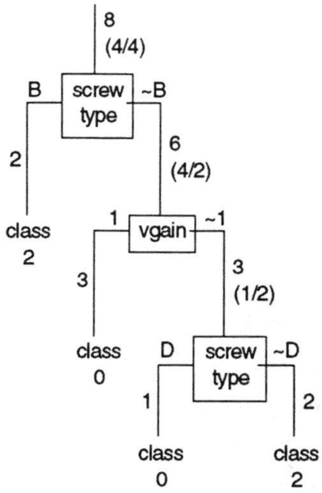
The run-times for tree building and post-pruning are roughly proportional to the data set size and to the square of the unpruned tree depth. The extremely long run times for gain ratio and orthogonality are largely due to the word sense and Pima (natural cut points) data sets — both are large samples with many attributes and several very small split subsets.

8.3 Effects of Stopping

The effects of stopping based on P_0 are summarized in Table 5. (Full details are given in Appendix E.3). Though accuracy for the Servo and Obesity problems is reduced by pruning at the 0.05 level, the differences are not statistically significant (at the 95% level). The improved accuracy of the pre-pruned quartiles Pima data is significant at the 99.5% confidence level.

The decreased accuracy for the Servo data is largely due to pruning the XOR-like subtree shown in Figure 6. For the Obesity data, linear discriminant analysis fails, suggesting that the classes are not homogeneous (see Figure 2). The Obesity attributes are very noisy and correlated, and the data are very sparse relative to the concept being studied (see [48, pp. 224-229]).

Figure 6: An XOR-like Substructure



screw type B?	vgain =1?	screw type D?	class =0?
1	---	0	0
0	1	---	1
0	0	1	1
0	0	0	0

Overall, accuracy is mildly concave, peaking at around the $P_0 = 0.05$ level. These results support the conjecture that 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.

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 trees built using information gain, 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 Table 6 and Figure 7. For comparison, the results of pre-pruning using the χ^2 criterion (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) are shown for the other metrics (see Appendix E.4 for full details). This rule rarely resulted in a tree different from the unpruned tree. Splitting and stopping using P_0 is more practical, and results in trees which are simpler, more efficient, and generally no less accurate than splitting and post-pruning using any of the metrics.

8.4 Binary *vs.* Multi-way Splits

An additional set of experiments was conducted to determine the effects of using binary as opposed to multi-way splits. These data are summarized in Table 7. The most striking features of these data are that the multi-way trees have 2 or 3 times as many leaves as the binary split trees, are only one-half to one-third as deep, and reduce training/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 number of attribute-value pairs.

A very substantial time penalty is incurred when V-ary attributes are forced into V binary splits. Overall, learning time increases at least quadratically in the dimensionality of the data set. Approaches such as those suggested by Weiss and Indurkhya [63] to reduce dimensionality and optimum binarization techniques such as those used in C4.5 [58] and ASSISTANT [13] shoud 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.

Table 5: Effects of Stopping, Binary Splits

Data Set		Unpruned		Pruning Threshold Level					
			conf. limits	0.5	0.1	0.05	0.01	0.005	0.001

Cross-Validation Accuracy, %

Finance 1	Quartiles	75	57-90	79	79	79	71	64	¶ 44
Obesity	Natural	58	37-77	47	44	49	40	¶ 33	¶ 13
Obesity	Quartiles	51	31-71	42	49	49	40	¶ 29	¶ 36
Pima	Quartiles	65	60-70	68	§ 73	§ 73	§ 74	§ 74	§ 75
Servo Motors		95	89-98	93	91	89	89	90	¶ 81
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

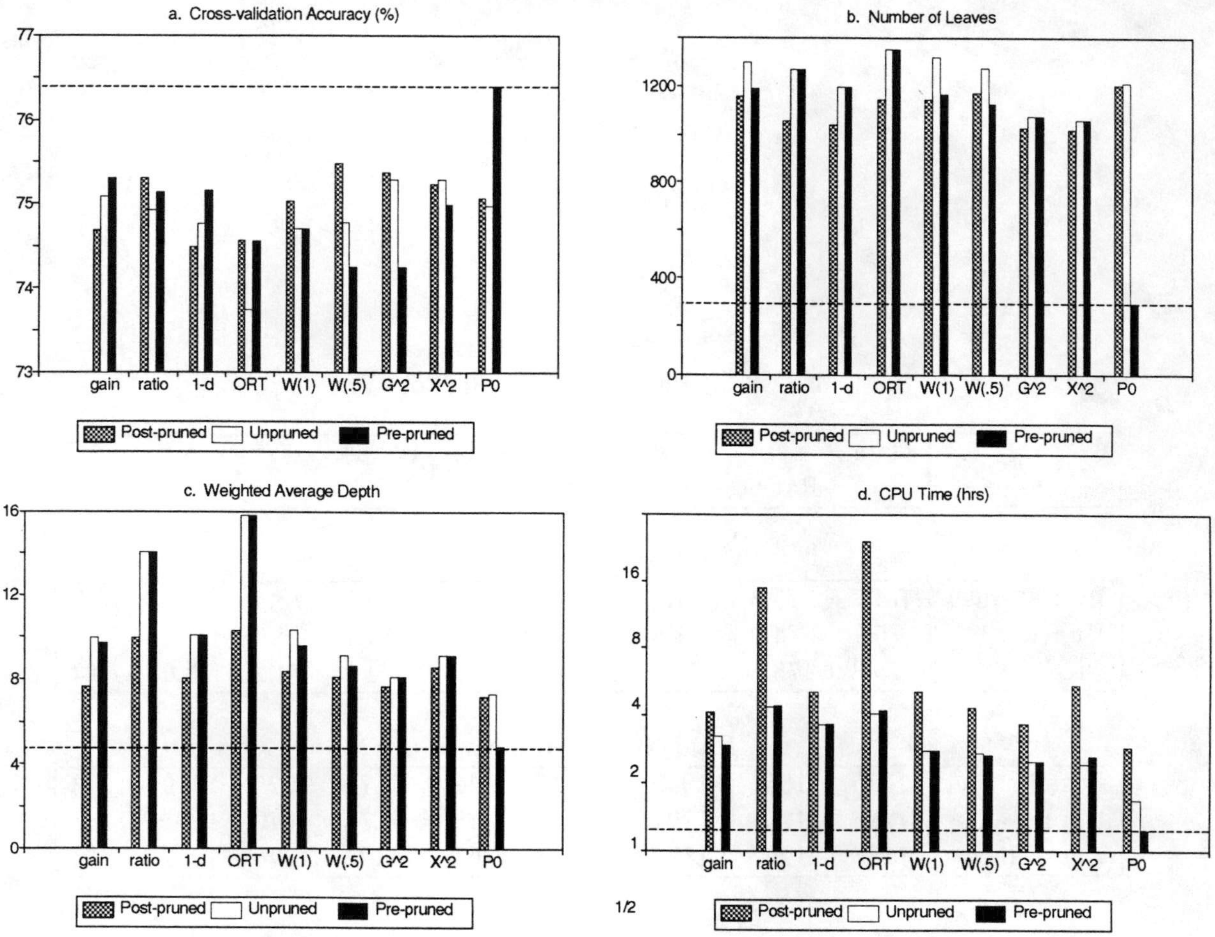
Total Run Time (sec)

Overall		3028		2799	2387	2242	1960	1889	1733
Natural		746		684	569	520	461	437	386
Quartiles		428		409	335	307	259	252	221

Table 6: Overall Effects of Pre- and Post-Pruning

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
Wtd. Avg. Accuracy, %									
Post-Pruned	74.7	75.3	74.5	74.6	75.0	75.5	75.4	75.2	75.1
Unpruned	75.1	74.9	74.8	73.7	74.7	74.8	75.3	75.3	75.0
Pre-pruned	75.3	75.1	75.2	74.5	74.7	74.2	74.2	75.0	76.4
Total Number of Leaves									
Post-Pruned	1155	1051	1039	1142	1141	1171	1025	1019	1203
Unpruned	1295	1267	1191	1351	1318	1272	1070	1061	1213
Pre-pruned	1186	1267	1191	1351	1164	1123	1070	1061	295
Wtd. Avg. Depth									
Post-Pruned	7.6	10.0	8.1	10.3	8.4	8.1	7.7	8.6	7.2
Unpruned	9.9	14.0	10.1	15.8	10.3	9.1	8.1	9.1	7.3
Pre-Pruned	9.7	14.0	10.1	15.8	9.6	8.6	8.1	9.1	4.8
Total Run Time (hh:mm)									
Post-Pruned	2:04	7:28	2:33	11:53	2:35	2:10	1:51	2:43	1:28
Unpruned	1:38	2:11	1:50	2:03	1:23	1:23	1:15	1:14	:50
Pre-Pruned	1:29	2:13	1:51	2:08	1:24	1:20	1:15	1:19	:37

Figure 7: Stopping *vs.* Post-Pruning



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. The reversal at the most severe pruning levels is a consequence of the binary split trees being over-pruned at those levels. When attributes are converted to binary splits in this way P_0 is larger (more apt to be pruned) for each of the new binary features than for the multi-valued feature (see Appendix A), and the cut-off level of P_0 should be adjusted upward accordingly.

9 Conclusions

1. Information gain, gain ratio, distance, orthogonality, chi-squared, and Beta each downplay some part of the influence of the number of partitions or 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 almost inevitably leads to very small subtrees where these measures are inadmissible.

2. The P_0 null hypothesis probability measure proposed here overcomes the difficulties encoun-

Table 7: 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 %

Finance 1	Nat	77	75	¶ 77	65	69	67	65	¶ 58	67	54
	Qua	¶ 72	75	79	¶ 71	64	¶ 52	71	71	¶ 52	56
Glass	Nat	51	¶ 53	52	52	44	47	¶ 44	51	50	50
Obesity	Nat	56	58	49	§ 40	§ 33	51	58	56	§ 62	§ 56
	Qua	40	51	49	40	§ 29	49	40	44	58	§ 63
Wine	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

tered when the classes and attribute values are unevenly distributed or the number of partitions is large. The unpruned trees it builds are simpler, more efficient, and generally no less accurate than those built by the other measures.

3. The P_0 measure can be used to stop splitting. This is more practical, and the resulting trees are simpler, more efficient, and generally no less accurate than unpruned or post-pruned trees. A stopping threshold level of $P_0 > 0.05$ is recommended.

4. 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. There is no point in continuing the inductive process when the null hypothesis is probably true ($P_0 > 0.5$), and in most domains little point in continuing when $P_0 > 0.05$.

10 Acknowledgements

The base-line Common Lisp implementation of ID3 used here is due to, and copyright by, Dr. Raymond Mooney (University of Texas). Dr. Mooney's copyright notice gives permission to copy and modify the program provided that his original copyright notice is included, and he acknowledges various sources for different parts of the code. The code was modified by substituting the various heuristic functions for information gain and chi-squared, and by optionally calling a Common Lisp implementation of Dr. J. Ross Quinlan's (University of Sydney) pessimistic post-pruning routine [58], adapted to Dr. Mooney's data structures. The copyright notice accompanying Dr. Quinlan's C source code gives permission to copy the code for personal use, provided that neither the code nor any modified code is distributed.

Appendices

A Data Sets

Table 8: Description of Data Sets

Data Set	Source	Description
BUPA	UCI	liver disorder data, 345 instances in 2 classes 6 numeric attributes ⟼ 35 binary, very nonuniform
Finance 1	Morrison	insurer financial ratios, 52 instances in 2 classes 6 numeric attributes ⟼ 29 binary, very nonuniform
Finance 2	Morrison	bankruptcy financial ratios, 66 instances in 2 classes 5 numeric attributes ⟼ 26 binary, very nonuniform
Solar Flare C Solar Flare M Solar Flare X	UCI	solar flare data type C has 3 classes type M has 4 classes type X has 2 classes 10 nominal attributes ⟼ 25 binary, fairly uniform
Glass	UCI	forensic data on glass, 214 instances in 7 classes 9 numeric attributes ⟼ 58 binary, very nonuniform
Iris	either	Iris species data, 150 instances in 3 classes 4 numeric attributes ⟼ 17 binary, fairly uniform
Obesity	Morrison	obesity characteristics, 45 instances in 4 classes 13 numeric attributes ⟼ 80 binary, some uniform, some not
Pima	UCI	diagnosis of diabetes, 768 instances in 2 classes 8 numeric attributes ⟼ 36 binary, some uniform, some not
Servo Motors	UCI	servo motor rise time, 167 instances continuous response ⟼ 5 nominal classes 4 nominal attributes ⟼ 19 binary, highly uniform
Soybean	UCI	soybean diseases, 47 instances in 4 classes 21 nominal attributes ⟼ 46 binary attributes, highly uniform
Thyroid	UCI	thyroid disorder, 215 instances in 3 classes 5 numeric attributes ⟼ 32 binary, very nonuniform
WAIS	Morrison	senility factor data, 49 instances in 2 classes 4 numeric attributes ⟼ 4 binary, some uniform, some not
Wine	UCI	wine type data, 178 instances in 3 classes 13 numeric attributes ⟼ 75 binary, fairly uniform
Word Sense	author	senses of lie/lay/lying, 869 instances in 16 classes 100 binary attributes, extremely non-uniform

The data sets are described in Table 8. Except for the word sense data, raw data and more detailed descriptions with full citations can be found in one or the other of two convenient sources:

UCI University of California, Irvine machine learning data depository [49].

Morrison *Multivariate Statistical Methods*, by Donald F. Morrison [48, App. B, pp 466-480].

B Nominalizing Attributes and Converting to Binary Splits

To nominalize a numeric attribute, all the attribute's values were lumped without regard to class, and a histogram measured for $(x_{i-1} < x \leq x_i)$ where the x_i and Δx are taken at the nearest "nice" (0.1, 0.2, 0.5, *etc.*) value to range$(x)/40$. This histogram was smoothed by applying a linear filter (3-point moving average of the 3-point moving averages). The cut points were taken as the local minima and 'shoulders' in this smoothed histogram. With q cut points in ascending order, nominal values were assigned, $(0, \text{ if } x \leq \text{cut}_1)$, $(1, \text{ if } \text{cut}_1 < x \leq \text{cut}_2)$, $\cdots$ $(q, \text{ if } \text{cut}_q < x)$.

To make a V-valued attribute binary, V binary variables were created and associated with the attribute values in order of their appearance in the source documentation or in ascending numeric order (1 if the attribute has that value, 0 if it has any other value).

One consequence of this conversion to binary splits is that each of the binary splits has a higher P_0 value (hence, increased risk of *Type II error*) than the multi-way split — consider a tertiary attribute for binary classes:

$$\begin{aligned}
P_0(\text{multi}) &= \left(\frac{n_1!\; n_2!}{N!}\right)\left(\frac{m_1!}{f_{11}!\; f_{21}!}\;\frac{m_2!}{f_{12}!\; f_{22}!}\;\frac{m_3!}{f_{13}!\; f_{23}!}\right)\\
P_0(\text{binary-1}) &= \left(\frac{n_1!\; n_2!}{N!}\right)\left(\frac{m_1!}{f_{11}!\; f_{21}!}\;\frac{(m_2+m_3)!}{(f_{12}+f_{13})!\;(f_{22}+f_{23})!}\right)\\
P_0(\text{multi}) &= \text{constant} \times \binom{m_2}{f_{12}}\binom{m_3}{f_{13}}\\
P_0(\text{binary-1}) &= \text{constant} \times \binom{m_2+m_3}{f_{12}+f_{13}}\\
&= \text{constant} \times \sum_{i=0}^{m_2}\binom{m_2}{i}\binom{m_3}{f_{12}+f_{13}-i}, \qquad \text{(See [1, p.822])}\\
&= P_0(\text{multi}) + \text{constant} \times \sum_{i=0, i\neq f_{12}}^{m_2}\binom{m_2}{i}\binom{m_3}{f_{12}+f_{13}-i}\\
&> P_0(\text{multi}) \qquad \text{likewise for binary-2 and binary-3}
\end{aligned}$$

C Derivation of the Hypergeometric

The posterior likelihood of a hypothesis, H, is given by Bayes' Theorem

$$P(H \mid \text{data}) = P(\text{data} \mid H)\; P(H) \;/\; P(\text{data}) \tag{13}$$

Typically, $P(H) = 1$ and $P(\text{data}) = 1$ (we believe *a priori* that our hypothesis is correct and our data unbiased). This results in $P(H \mid \text{data}) = P(\text{data} \mid H)$, the general principle (see Iversen [34, p. 62]) that Bayesian methods using non-informative priors lead to the same numeric results as classical statistics (the interpretation is different). (*Note*—For random sampling data, $P(\text{data}) = 1$ applies to the data without regard to order, not to the particular permutation.)

Given models $P(\text{data} \mid H, \hbar)$ and $P(\text{data} \mid \Re)$ in terms of some sets of parameters, $\hbar$ and $\Re$, then

$$P(\text{data} \mid H) = \int \cdots \int P(\text{data} \mid H, \hbar)\; P(\hbar)\; (d\hbar_1 d\hbar_2 \cdots) \tag{14}$$

$$P(\text{data}) = \int \cdots \int P(\text{data} \mid \Re)\; P(\Re)\; (d\Re_1 d\Re_2 \cdots) \tag{15}$$

Absent prior knowledge of the parameters, their priors, $P(\hbar)$ and $P(\Re)$, should be non-informative (see Box & Tiao [8, pp. 25-60]); and it is helpful if they are conjugate (easily integrated).

For N independent random instances $(i = 1 \cdots N)$ of an experiment having $C \times V$ distinct possible outcomes $[(y_i, x_i)$, for $y_i = 1 \ldots C$ and $x_i = 1 \ldots V]$ and constant likelihood for each outcome $[p_{cv} \equiv P(y = c, x = v)]$, the likelihoods of the sequence $\{(y_1, x_1) \ldots (y_N, x_N)\}$ and the unordered data $f_{cv} \equiv \text{freq}(y = c, x = v)$ are given by the multinomial distribution ([30, pp. 96-101])

$$P(\{y_i, x_i\} \mid \{p_{cv}\}) = \prod\prod p_{cv}^{f_{cv}} \tag{16}$$

$$P(\{f_{cv}\} \mid \{p_{cv}\}) = \frac{N!}{\prod\prod f_{cv}!} \prod\prod p_{cv}^{f_{cv}} \tag{17}$$

$$\begin{aligned}
\text{where}\quad & n_c = \textstyle\sum_v f_{cv} \qquad m_v = \sum_c f_{cv} \qquad \sum_c n_c = \sum_v m_v = \sum_c \sum_v f_{cv} = N \\
& p_{cv} = \phi_{cv} \varphi_v \qquad \phi_{cv} \equiv P(y_i = c \mid x_i = v) \\
& \varphi_v \equiv P(x_i = v) = \textstyle\sum_c p_{cv} = \sum_c \phi_{cv} \varphi_v \\
& \theta_c \equiv P(y_i = c) = \textstyle\sum_v p_{cv} = \sum_v \phi_{cv} \varphi_v \\
& \textstyle\sum_c \phi_{cv} = \sum_v \varphi_v = \sum_c \theta_c = \sum_c \sum_v p_{cv} = 1
\end{aligned} \tag{18}$$

For application of Bayes theorem to the multinomial, Dirichlet's integral

$$\mathcal{D}(\{r_i\}, q) \equiv \int \cdots \int \left(\prod_{i=1}^{q} \psi_i^{r_i - 1} \right) \; d\psi_1 d\psi_2 \cdots d\psi_{q-1} = \frac{\prod \Gamma(r_i)}{\Gamma(M)} \; S^{M-1} \tag{19}$$

is helpful, where the integration is taken over the region $\psi_i \geq 0$, $S = \sum \psi_i$, $M = \sum r_i$. A conjugate prior for the multinomial (Equation 16 or 17) takes the form of a Dirichlet density

$$P(\{p_{cv}\} \mid \alpha) = \frac{\Gamma(CV\alpha)}{\Gamma(\alpha)^{CV}} \prod\prod p_{cv}^{\alpha - 1} \tag{20}$$

which is non-informative for $\alpha \approx 1$(see Box & Tiao [8, pp. 25-60]).

Multiplying Equations 17 and 20, and applying the Dirichlet integral (Equation 19) gives

$$P(\{f_{cv}\} \mid \alpha) = \frac{N!}{\prod\prod f_{cv}!} \; \frac{\Gamma(CV\alpha)}{\Gamma(\alpha)^{CV}} \; \frac{\prod\prod \Gamma(f_{cv} + \alpha)}{\Gamma(N + CV\alpha)}$$

See Hartigan [30, pp 100-101] on the difficulties of selecting the parameter α for a Dirichlet prior — rather than seeking a particular α, a pure Bayesian approach chooses a non-informative prior for α near unity $((1 - \delta) \leq \alpha \leq (1 - \delta + \Delta))$ where δ and Δ are small positive values. Choosing

$$P(\alpha) = \left(\frac{\Gamma(\alpha)^{CV}}{\Gamma(CV\alpha)} \right) \left(\frac{\prod\prod \Gamma(f_{cv} + 1)}{\Gamma(N + 1)} \right) \left(\frac{\Gamma(N + CV\alpha)}{\prod\prod \Gamma(f_{cv} + \alpha)} \right) \left(\frac{1}{\Delta} \right) \tag{21}$$

gives $P(\{f_{cv}\}) = P(\text{unordered data}) = 1$, which is consistent with our *a priori* assumption that the observed data are unbiased. Applying this same process (Equations 19, 20, and 21) to Equation 16 gives

$$P(\text{data}) = P(\{y_i, x_i\}) = \frac{\prod\prod f_{cv}!}{N!} \tag{22}$$

Under the null hypothesis, H_0, $(\forall v : \phi_{cv} = \theta_c$, and $p_{cv} = \theta_c\ \varphi_v)$, and the multinomial of Equation 16 becomes

$$P(\{y_i, x_i\} \mid H_0, \{\theta_c\}, \{\varphi_v\}) = \prod\prod (\theta_c\varphi_v)^{f_{cv}} = \left(\prod \varphi_v^{m_v}\right)\left(\prod \theta_c^{n_c}\right) \tag{23}$$

Taking $P(H_0) = 1$, independent priors for $\{\theta_c\}$ and $\{\varphi_v\}$

$$P(\{\theta_c\} \mid H_0, \alpha_y) = \frac{\Gamma(C\alpha_y)}{\Gamma(\alpha_y)^C}\prod \theta_c^{\alpha_y - 1}, \qquad P(\{\varphi_v\} \mid H_0, \alpha_x) = \frac{\Gamma(V\alpha_x)}{\Gamma(\alpha_x)^V}\prod \varphi_v^{\alpha_x - 1}$$

and appropriate priors for $P(\alpha_y)$ and $P(\alpha_x)$ (analogous to Equation 21) leads to

$$P(\text{data} \mid H_0) = P(\{y_i, x_i\} \mid H_0) = \left(\frac{\prod n_c!}{N!}\right)\left(\frac{\prod m_v!}{N!}\right) \tag{24}$$

From the expression for $P(\text{data})$ (Equation 22) and $P(\text{data} \mid H_0)$ (Equation 24) and Bayes' theorem (Equation 13) we get the multiple hypergeometric distribution:

$$P_0 \equiv P(H_0 \mid \text{data}) = \left(\frac{\prod n_c!}{N!}\right)\left(\frac{\prod m_v!}{N!}\right) \Big/ \left(\frac{\prod\prod f_{cv}!}{N!}\right) \tag{25}$$

D Approximations to the Hypergeometric

D.1 Gain and Chi-squared

For $n \neq 0$, Stirling's approximation, $\ln(n!) \approx 0.5\ln(2\pi n) + n\ln(n) - n$, leads to

$$-\ln\left(\frac{\prod_{j=1}^{k} M_j!}{N!}\right) \approx N\ln(2)\ I_k(M_j/N) - \frac{1}{2}\left((k-1)\ln(2\pi N) + \sum_{j=1}^{k}\ln(M_j/N)\right) \tag{26}$$

where $I_k(M_j/N) \equiv \sum[-(M_j/N)\log_2(M_j/N)]$ is the entropy (information) function. Taking second order terms of a Taylor series for $\ln(M_j/N)$ about $1/k = \text{avg}(M_j/N)$,

$$\begin{aligned} -\ln\left(\frac{\prod_{j=1}^{k} M_j!}{N!}\right) \ &\approx\ N\ln(2)\ I_k(M_j/N)\ +\ \frac{k^2(k-1)}{4}\text{Var}(M_j/N) \\ &\quad + \frac{1}{2}\left(k\ln(k) - (k-1)\ln(2\pi N)\right) \end{aligned} \tag{27}$$

Applying the approximation from Equation 27 to the hypergeometric of Equation 25, and noting that

$$\text{gain} \ \equiv\ I_C(n_c/N) + I_V(m_v/N) - I_{CV}(f_{cv}/N)$$

gives

$$\begin{aligned}
-2\ln(P_0) \;\approx\;& 2N\ln(2)\text{ gain} \\
&+ (C-1)(V-1)\ln(2\pi N) \;-\; [C(V-1)\ln(C) + V(C-1)\ln(V)] \\
&-\frac{1}{2}\Big[C^2V^2(CV-1)\ \mathrm{Var}(f_{cv}/N) - \\
&C^2(C-1)\ \mathrm{Var}(n_c/N) \;-\; V^2(V-1)\ \mathrm{Var}(m_v/N)\Big]
\end{aligned} \qquad (28)$$

Or, if $\ln(f_{cv})$ is replaced by its Taylor series around $e_{cv} = n_c m_v / N$ (rather than around the overall mean, N/CV)

$$\begin{aligned}
-2\ln(P_0) \;\approx\;& X^2 \\
&+ (C-1)(V-1)\ln(2\pi N) \;+\; [C(V-1)\ln(C) + V(C-1)\ln(V)] \\
&-\frac{(C-1)(V-1)}{2}\Big[C^2\ \mathrm{Var}(n_c/N) + V^2\ \mathrm{Var}(m_v/N)\Big] \\
&-\frac{1}{2}\sum\sum\Big[(f_{cv}-e_{cv})(f_{cv}-3e_{cv})\ /\ e_{cv}^2\Big]
\end{aligned} \qquad (29)$$

D.2 Buntine's Beta

In Buntine's derivation [11]

$$\begin{aligned}
p_{cv} \;\equiv\;& P(y_i = c, x_i = v \mid H, \beta) = \phi_{cv}(H)\cdot\varphi_v(\beta) \\
\text{where}\quad & \phi_{cv}(H) \equiv P(y_i = c \mid x_i = v, H) \qquad \varphi_v(\beta) \equiv P(x_i = v \mid \beta) \\
& \beta \text{ is a set of (unknown) population characteristics} \\
\text{and}\quad & H \text{ is some hypothesis about the classification and partition}
\end{aligned}$$

Assuming that the priors of the hypothesis and the population characteristics are *a prior* independent, $P(H,\beta) = P(H)\cdot P(\beta)$, then

$$\begin{aligned}
P(H \mid y_i = c, x_i = v) \;&\propto\; \phi_{cv}\cdot P(H) \\
P(H \mid \{f_{cv}\}) \;&\propto\; P(H)\cdot\prod\prod\phi_{cv}^{f_{cv}}
\end{aligned}$$

The hypothesis H is represented as the conjunction of the structure of the partition, τ, and the class probabilities $\{\phi_{cv}\}$

$$\begin{aligned}
P(H) \;&=\; P(\tau,\{\phi_{cv}\}) = P(\tau)\cdot P(\{\phi_{cv}\}\mid\tau) \\
P(H \mid \{\phi_{cv}\}) \;&\propto\; P(\tau)\cdot P(\{\phi_{cv}\}\mid\tau)\cdot\prod\prod\phi_{cv}^{f_{cv}}
\end{aligned}$$

$P(\{\phi_{cv}\}\mid\tau)$ is modeled as the product of V independent Dirichlet densities,

$$P(\{\phi_{cv}\}\mid\tau) = \prod_{v=1}^{V}\left[\left(\frac{\Gamma(C\alpha)}{\Gamma(\alpha)^C}\right)\prod_{c=1}^{C}\phi_{cv}^{\alpha-1}\right] = \frac{\Gamma(C\alpha)^V}{\Gamma(\alpha)^{CV}}\prod\prod\phi_{cv}^{\alpha-1} \qquad (30)$$

leading to Buntine's Beta,

$$P(\tau \mid \{y_i, x_i\}) \propto P(\tau)\cdot\frac{\Gamma(C\alpha)^V}{\Gamma(\alpha)^{CV}}\cdot\frac{\prod\prod\Gamma(f_{cv}+\alpha)}{\prod\Gamma(m_v + C\alpha)} = e^{-W(\tau,\alpha)} \qquad (31)$$

Applying Stirling's approximation and a Taylor series as in Section D.1 leads to $I_{CV}(f_{cv}/N)$ as an approximation to $W(\tau, \alpha)$.

There are two notable differences between Buntine's Beta derivation and that given for the hypergeometric in Appendix C:

1. Buntine gives his expression in parametric form (the Dirichlet parameter, α), while the hypergeometric removes α via a non-informative prior. Though this difference is obvious in the formulas, its impact on split rankings may be subtle for α near unity.
2. The priors (Equations 20 and 30) differ in two respects:
   - $\Gamma(CV\alpha)$ *versus* $\Gamma(C\alpha)^V$ — since these are only normalizing factors for the integration, this is probably of little consequence.
   - $\prod\prod p_{cv}^{\alpha-1} = \left(\prod \varphi_v^{C(\alpha-1)}\right) \prod\prod \phi_{cv}^{\alpha-1}$ *versus* $\prod\prod \phi_{cv}^{\alpha-1}$ — in applying the Dirichlet integral, the hypergeometric derivation incorporates the linear constraints $\sum_v \varphi_v \phi_{cv} = \theta_c$ in addition to $\sum_c \phi_{cv} = 1$, while Buntine does not. This difference may have a large impact when some of the φ_v or θ_c terms are small.

Factor 2 above arises from Buntine's assumption of V independent Dirichlet priors for $\{\phi_{cv}\}$, where the hypergeometric derivation assumes that only $V-1$ are independent. This difference is equivalent to conditioning on both sets of marginal counts in the contingency table (leading to the hypergeometric) *vs.* treating only one of the marginal distributions as fixed (Beta) — the difference between conditional and unconditional tests. Agresti [2, pp. 65-66] cites and summarizes various arguments [5, 6, 7, 27, 28, 29, 39, 68] as to whether these analyses should be conditional.

E Detailed Experimental Results

E.1 Unpruned Trees, Binary Splits

Table 9: Accuracy of Unpruned Trees

Data Set		Gain	Gain	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
			Ratio							
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

Weighted Averages

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

Table 10: No. of Leaves of Unpruned Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
BUPA	Nat	44	41	43	41	55	56	32	32	48
	Qua	127	56	50	136	115	117	52	51	116
Finance 1	Nat	17	20	17	20	21	19	17	17	17
	Qua	13	13	13	13	13	13	13	13	13
Finance 2	Nat	8	8	8	8	14	8	8	8	8
	Qua	8	8	8	8	8	8	8	8	8
Solar Flare C		64	62	61	69	80	80	64	60	67
Solar Flare M		57	80	62	82	68	58	57	57	58
Solar Flare X		21	28	21	25	33	27	21	22	22
Glass	Nat	58	58	56	56	54	53	58	51	52
	Qua	64	68	64	69	65	66	64	63	63
Iris	Nat	11	11	11	11	11	11	11	11	11
	Qua	15	16	15	15	15	16	15	15	16
Obesity	Nat	13	16	13	18	13	13	13	13	16
	Qua	14	16	15	16	13	13	14	13	13
Pima	Nat	180	167	182	169	106	102	83	84	104
	Qua	234	253	202	244	250	237	199	200	217
Servo Motors		14	14	14	14	14	14	8	8	14
Soybean		4	4	4	4	4	4	4	4	4
Thyroid	Nat	20	22	21	22	28	27	20	22	19
	Qua	23	24	24	24	26	24	23	24	24
WAIS	Nat	5	5	5	5	5	5	5	5	6
	Qua	20	20	20	18	18	20	20	20	19
Wine	Nat	15	17	15	19	17	17	15	16	17
	Qua	13	14	13	19	13	13	13	14	13
Word Sense		233	226	234	226	259	251	233	230	248

Totals

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

Table 11: Wtd. Avg. Depth of Unpruned Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
BUPA	Nat	19.6	21.8	19.6	21.8	19.9	16.8	14.0	14.0	7.6
	Qua	7.8	4.3	4.1	8.5	7.6	7.8	4.1	4.1	7.5
Finance 1	Nat	5.1	9.2	5.3	6.2	6.5	5.9	5.1	5.1	5.1
	Qua	4.2	4.1	4.0	4.1	4.0	4.0	4.2	4.0	4.0
Finance 2	Nat	2.4	2.4	2.4	2.4	4.9	2.4	2.4	2.4	2.4
	Qua	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
Solar Flare C		7.5	9.4	8.0	10.3	10.3	8.1	7.5	7.4	6.8
Solar Flare M		6.3	10.0	8.3	10.6	10.3	7.8	6.3	7.9	6.8
Solar Flare X		3.6	4.5	3.6	4.5	6.1	4.5	3.6	3.7	3.7
Glass	Nat	8.8	18.0	9.0	17.0	9.7	9.5	8.8	7.6	5.7
	Qua	6.6	7.7	6.7	7.9	6.9	6.9	6.6	7.3	6.5
Iris	Nat	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1
	Qua	4.1	4.7	4.1	4.6	4.1	4.0	4.1	4.1	4.0
Obesity	Nat	5.3	7.4	5.3	8.0	5.3	5.3	5.3	5.3	4.6
	Qua	4.1	4.7	4.2	4.7	3.8	3.8	4.1	3.9	3.8
Pima	Nat	21.5	24.0	21.6	25.1	12.4	11.5	11.2	11.1	8.7
	Qua	8.4	10.1	8.6	10.0	8.9	8.5	8.2	8.0	8.0
Servo Motors		2.9	2.9	2.9	2.9	2.2	2.2	1.5	1.5	2.2
Soybean		2.0	2.4	2.0	2.4	2.0	2.0	2.0	2.0	2.0
Thyroid	Nat	5.7	11.0	7.8	11.1	6.8	6.6	5.7	7.8	5.6
	Qua	3.8	3.8	3.8	3.8	4.0	3.8	3.8	3.8	3.8
WAIS	Nat	3.3	3.3	3.3	3.4	3.4	3.3	3.3	3.3	3.1
	Qua	4.9	5.7	4.9	4.8	4.8	4.5	4.9	4.8	4.2
Wine	Nat	4.0	5.1	4.0	7.1	4.0	4.0	4.0	4.4	4.0
	Qua	3.8	4.0	3.9	5.3	3.8	3.8	3.8	4.2	3.8
Word Sense		15.8	34.8	16.5	44.4	21.9	18.3	15.8	22.3	14.6

Weighted Averages

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

Table 12: Run Time of Unpruned Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
BUPA	Nat	204	278	282	191	209	203	196	195	122
	Qua	141	94	97	91	93	99	65	56	66
Finance 1	Nat	23	32	29	19	18	17	16	14	12
	Qua	12	18	21	9	10	11	13	10	7
Finance 2	Nat	11	12	13	8	13	8	9	7	7
	Qua	6	10	11	5	6	7	7	6	4
Solar Flare C		101	125	129	94	99	102	85	77	69
Solar Flare M		93	133	134	97	102	91	79	80	66
Solar Flare X		42	58	59	42	57	46	39	38	36
Glass	Nat	251	372	257	272	187	188	188	161	119
	Qua	95	153	159	80	85	91	95	82	54
Iris	Nat	13	15	17	10	11	13	12	11	10
	Qua	13	21	20	12	13	15	14	13	9
Obesity	Nat	76	97	89	62	47	50	53	43	34
	Qua	30	53	54	25	24	26	29	22	14
Pima	Nat	497	654	689	481	410	389	371	352	312
	Qua	398	457	459	348	304	305	290	263	215
Servo Motors		19	20	23	13	13	15	14	12	10
Soybean		13	15	16	10	9	9	10	10	7
Thyroid	Nat	54	97	87	68	45	73	55	64	48
	Qua	24	35	38	19	24	28	25	22	17
WAIS	Nat	1.6	1.6	1.7	1.2	0.9	1.3	1.3	0.9	0.8
	Qua	8	15	17	7	7	7	9	7	4
Wine	Nat	105	152	141	130	95	107	97	95	81
	Qua	52	80	83	60	53	62	57	56	38
Word Sense		3570	4885	3660	5241	3067	2992	2661	2716	1666

Totals

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

E.2 Post-pruned Trees, Binary Splits

Table 13: Accuracy of Post-Pruned Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
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

Weighted Averages

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

Table 14: No. of Leaves of Post-Pruned Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
BUPA	Nat	37	32	34	34	39	40	27	27	46
	Qua	117	51	50	122	115	114	52	51	114
Finance 1	Nat	17	16	17	20	16	17	17	17	17
	Qua	13	13	13	13	13	13	13	13	13
Finance 2	Nat	8	8	8	8	8	8	8	8	8
	Qua	8	8	8	8	8	8	8	8	8
Solar Flare C		60	59	59	63	67	66	58	60	67
Solar Flare M		57	52	56	53	51	54	57	53	58
Solar Flare X		20	26	20	25	25	25	20	22	22
Glass	Nat	54	48	53	48	49	49	55	50	52
	Qua	62	67	62	68	65	66	62	61	63
Iris	Nat	11	11	11	11	11	11	11	11	11
	Qua	15	15	15	14	15	16	15	15	16
Obesity	Nat	13	15	13	17	13	13	13	13	16
	Qua	13	15	15	16	13	13	13	13	13
Pima	Nat	88	72	76	86	81	84	73	75	101
	Qua	221	236	191	231	237	234	188	191	216
Servo Motors		14	8	8	14	14	14	8	8	14
Soybean		4	4	4	4	4	4	4	4	4
Thyroid	Nat	19	18	18	18	18	18	19	22	18
	Qua	23	24	24	24	24	24	23	24	24
WAIS	Nat	5	5	5	5	5	5	5	5	6
	Qua	19	19	19	16	16	20	19	19	19
Wine	Nat	15	15	15	16	17	17	15	16	17
	Qua	13	14	13	17	13	13	13	14	13
Word Sense		229	200	229	191	204	225	229	219	247

Totals

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

Table 15: Wtd. Avg. Depth of Post-Pruned Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
BUPA	Nat	8.4	15.2	13.9	11.2	14.9	10.7	11.7	11.7	7.4
	Qua	7.7	4.1	4.0	8.1	7.6	7.7	4.1	4.1	7.5
Finance 1	Nat	5.1	6.7	5.3	6.2	5.0	5.2	5.1	5.1	5.1
	Qua	4.2	4.1	4.0	4.1	4.0	4.0	4.2	4.0	4.0
Finance 2	Nat	2.4	2.4	2.4	2.4	2.5	2.4	2.4	2.4	2.4
	Qua	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
Solar Flare C		7.0	8.8	7.8	9.2	8.1	7.1	7.0	7.4	6.8
Solar Flare M		6.3	7.4	7.2	8.7	7.1	7.0	6.3	7.1	6.3
Solar Flare X		3.7	4.4	3.7	4.5	4.4	4.4	3.7	3.7	3.7
Glass	Nat	7.8	12.4	7.8	12.6	7.7	7.6	8.1	7.5	5.7
	Qua	6.6	7.6	6.7	7.8	6.9	6.9	6.6	7.1	6.5
Iris	Nat	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1
	Qua	4.1	4.1	4.1	4.0	4.1	4.0	4.1	4.1	4.0
Obesity	Nat	5.3	7.1	5.3	7.6	5.3	5.3	5.3	5.3	4.6
	Qua	3.8	4.4	4.2	4.7	3.8	3.8	3.8	3.9	3.8
Pima	Nat	9.0	11.4	9.8	11.9	9.6	9.3	9.8	9.8	8.5
	Qua	8.4	9.1	8.1	9.3	8.7	8.5	8.0	8.0	8.0
Servo Motors		2.2	1.5	1.5	2.2	2.2	2.2	1.5	1.5	2.2
Soybean		2.0	2.4	2.0	2.4	2.0	2.0	2.0	2.0	2.0
Thyroid	Nat	5.1	7.8	7.8	7.7	5.4	5.4	5.1	7.8	4.9
	Qua	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
WAIS	Nat	3.3	3.3	3.3	3.4	3.4	3.3	3.3	3.3	3.1
	Qua	4.5	4.8	4.5	4.2	4.2	4.5	4.5	4.5	4.2
Wine	Nat	4.0	4.5	4.0	5.7	4.0	4.0	4.0	4.4	4.0
	Qua	3.8	4.0	3.9	4.8	3.8	3.8	3.8	4.2	3.8
Word Sense		15.7	24.5	16.2	25.1	18.0	17.0	15.7	21.0	14.6

Weighted Averages

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

Table 16: Run Time of Post-Pruned Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
BUPA	Nat	403	911	589	1467	735	611	476	500	185
	Qua	209	155	138	180	158	159	100	94	143
Finance 1	Nat	26	49	34	27	24	22	20	18	17
	Qua	19	22	23	14	13	14	15	12	11
Finance 2	Nat	11	13	14	9	16	9	10	8	7
	Qua	10	11	12	9	8	8	8	7	7
Solar Flare C		163	236	202	224	220	160	141	130	113
Solar Flare M		145	278	223	247	214	156	119	166	106
Solar Flare X		66	77	77	68	96	60	58	52	49
Glass	Nat	306	850	382	929	282	265	248	211	155
	Qua	149	191	195	140	116	121	122	90	93
Iris	Nat	19	20	22	16	16	18	16	15	15
	Qua	24	36	29	29	21	22	21	21	19
Obesity	Nat	74	107	95	77	50	54	59	49	39
	Qua	41	54	56	36	26	28	31	25	21
Pima	Nat	723	2371	1134	2893	860	754	713	701	500
	Qua	662	720	640	589	569	546	589	508	474
Servo Motors		20	23	26	17	16	19	17	15	14
Soybean		13	16	17	12	10	10	9	10	8
Thyroid	Nat	83	185	117	172	100	105	74	100	70
	Qua	38	37	45	34	32	35	33	30	29
WAIS	Nat	2.6	3.0	3.1	2.5	2.5	2.5	2.6	2.4	2.4
	Qua	17	22	22	12	11	10	13	11	10
Wine	Nat	121	161	149	191	104	120	108	112	95
	Qua	75	92	94	92	64	73	68	67	60
Word Sense		4028	20320	4826	36020	5540	4406	3606	6839	3020

Totals

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

E.3 Pre-Pruned Trees, Binary Splits

Table 17: Accuracy of Pre-Pruned Trees, Binary Splits

		Unpruned		Pruning Threshold Level					
data set	Nominalize		conf. limits	0.5	0.1	0.05	0.01	0.005	0.001
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

Weighted Averages

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

Table 18: Number of Leaves – Pre-Pruned Trees, Binary Splits

		Unpruned	Pruning Threshold Level					
data set	Nominalize		0.5	0.1	0.05	0.01	0.005	0.001
BUPA	Natural	48	48	32	6	3	3	2
	Quartiles	116	93	27	18	8	5	2
Finance 1	Natural	17	12	5	4	2	2	2
	Quartiles	13	12	5	5	4	4	2
Finance 2	Natural	8	7	4	4	2	2	2
	Quartiles	8	6	3	3	3	3	3
Solar Flare C		67	38	14	9	3	3	2
Solar Flare M		58	38	9	8	5	4	3
Solar Flare X		22	12	4	2	2	2	1
Glass	Natural	52	37	21	18	12	10	7
	Quartiles	63	53	32	21	15	15	10
Iris	Natural	11	17	11	9	7	5	5
	Quartiles	16	12	8	8	6	6	5
Obesity	Natural	16	15	10	8	6	4	3
	Quartiles	13	12	10	8	5	3	3
Pima	Natural	104	61	16	12	8	7	5
	Quartiles	217	164	44	28	11	9	6
Servo Motors		14	12	8	8	6	5	3
Soybean		4	4	4	4	4	4	4
Thyroid	Natural	19	13	9	8	8	6	6
	Quartiles	24	19	9	6	6	6	5
WAIS	Natural	6	5	2	2	2	2	1
	Quartiles	19	11	3	2	2	1	1
Wine	Natural	17	16	12	11	8	7	6
	Quartiles	13	12	10	10	8	8	7
Word Sense		248	166	94	73	46	38	29

Totals

Overall	1213	895	406	295	192	164	125
Natural	298	231	122	82	58	48	39
Quartiles	502	394	151	109	68	60	44

Table 19: Wtd. Avg. Depth of Pre-Pruned Trees, Binary Splits

		Unpruned	Pruning Threshold Level					
data set	Nominalize		0.5	0.1	0.05	0.01	0.005	0.001
BUPA	Natural	7.6	6.8	4.1	4.1	1.9	1.9	1.0
	Quartiles	7.5	7.3	5.5	4.8	3.3	2.0	1.0
Finance 1	Natural	5.1	4.2	2.1	1.7	1.0	1.2	1.0
	Quartiles	4.0	4.0	2.5	2.5	2.4	2.4	1.0
Finance 2	Natural	2.4	2.4	1.7	1.7	1.0	1.0	1.0
	Quartiles	2.5	2.4	1.7	1.7	1.7	1.7	1.7
Solar Flare C		6.8	6.2	4.4	3.9	1.9	1.9	1.0
Solar Flare M		6.8	6.0	3.9	3.8	3.0	2.8	1.9
Solar Flare X		3.7	3.5	2.3	1.0	1.0	1.0	0.0
Glass	Natural	5.7	5.4	4.7	4.6	3.5	3.3	2.5
	Quartiles	6.5	6.4	5.8	4.8	4.4	4.4	3.7
Iris	Natural	4.5	6.1	2.7	2.4	2.0	1.7	1.7
	Quartiles	4.0	3.8	3.4	3.4	3.1	3.1	2.7
Obesity	Natural	4.6	3.6	4.0	3.6	3.2	2.6	1.9
	Quartiles	3.8	4.7	3.4	3.2	2.4	1.7	1.7
Pima	Natural	8.7	7.2	5.2	4.9	4.1	3.9	3.2
	Quartiles	8.0	7.8	6.0	5.1	3.7	3.5	2.9
Servo Motors		2.2	2.1	2.0	2.0	1.9	1.7	1.3
Soybean		2.0	2.0	2.0	2.0	2.0	2.0	2.0
Thyroid	Natural	5.6	5.2	4.7	4.4	4.4	3.6	3.6
	Quartiles	3.8	3.7	2.8	2.4	2.4	2.4	2.3
WAIS	Natural	3.1	2.8	1.0	1.0	1.0	1.0	0.0
	Quartiles	4.2	3.7	1.3	1.0	1.0	0.0	0.0
Wine	Natural	4.0	4.0	3.6	3.9	3.0	2.9	2.6
	Quartiles	3.8	3.8	3.7	3.7	3.1	3.1	3.0
Word Sense		14.6	13.4	11.3	10.7	9.3	8.9	8.2

Weighted Averages

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

Table 20: Run Time of Pre-Pruned Trees, Binary Splits

		Unpruned	Pruning Threshold Level					
data set	Nominalize		0.5	0.1	0.05	0.01	0.005	0.001
BUPA	Natural	122	109	78	66	47	43	34
	Quartiles	66	62	48	43	31	28	19
Finance 1	Natural	12	10	6	6	5	4	4
	Quartiles	7	6	5	5	4	3	2
Finance 2	Natural	7	6	6	5	5	5	5
	Quartiles	4	4	4	4	3	6	3
Solar Flare C		69	65	51	45	29	24	21
Solar Flare M		66	62	48	43	34	33	28
Solar Flare X		36	33	26	20	18	16	11
Glass	Natural	119	112	101	94	83	73	63
	Quartiles	54	52	47	42	38	38	34
Iris	Natural	10	10	10	9	8	8	8
	Quartiles	9	9	9	8	8	8	8
Obesity	Natural	34	32	29	28	23	22	17
	Quartiles	14	14	13	12	10	9	8
Pima	Natural	312	278	218	193	173	169	150
	Quartiles	215	204	157	140	113	110	97
Servo Motors		10	10	10	10	9	9	8
Soybean		7	7	7	7	7	7	7
Thyroid	Natural	48	46	44	42	42	40	38
	Quartiles	17	16	15	14	14	14	14
WAIS	Natural	0.8	0.8	0.7	0.6	0.5	0.6	0.4
	Quartiles	4.3	3.7	2.1	1.8	1.5	1.1	0.9
Wine	Natural	81	80	77	75	74	72	68
	Quartiles	38	38	37	37	36	36	36
Word Sense		1666	1530	1342	1291	1143	1110	1050

Totals

Overall	3028	2799	2387	2242	1960	1889	1733
Natural	746	684	569	520	461	437	386
Quartiles	428	409	335	307	259	252	221

E.4 X^2-Stopped Trees, Binary Splits

Table 21: Accuracy of X^2-Stopped Trees, Binary Splits

Data Set		Gain	Gain	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
			Ratio							¶
BUPA	Nat	52	57	53	56	53	50	52	51	57
	Qua	61	62	63	56	56	60	55	68	57
Finance 1	Nat	79	75	75	77	73	73	73	71	77
	Qua	77	75	71	77	69	67	69	73	79
Finance 2	Nat	94	95	94	94	95	94	94	94	94
	Qua	92	91	89	94	91	92	92	91	97
Solar Flare C		86	87	87	87	87	86	87	87	88
Solar Flare M		84	84	85	84	85	85	84	85	90
Solar Flare X		96	97	98	96	97	97	97	97	98
Glass	Nat	49	52	50	47	48	51	46	52	52
	Qua	69	66	70	67	71	68	66	66	70
Iris	Nat	95	95	95	93	95	93	95	95	95
	Qua	92	91	93	91	89	91	90	90	92
Obesity	Nat	58	49	58	64	53	51	56	60	49
	Qua	49	42	53	56	51	56	49	53	49
Pima	Nat	73	72	72	72	72	72	72	73	72
	Qua	69	65	66	65	67	65	67	66	73
Servo Motors		95	95	94	96	95	96	94	95	89
Soybean		98	98	98	98	98	98	98	96	96
Thyroid	Nat	90	89	90	91	91	91	91	90	93
	Qua	94	93	92	93	95	92	93	93	92
WAIS	Nat	80	84	84	84	82	82	84	84	84
	Qua	67	67	65	69	63	63	63	63	63
Wine	Nat	92	90	91	90	93	90	94	92	92
	Qua	92	94	93	88	92	90	89	90	88
Word Sense		64	65	64	64	64	64	65	63	66

¶ Stopped using P_0

Weighted Averages

Overall	75.3	75.1	75.2	74.5	74.7	74.2	74.2	75.0	76.4
Natural	72.5	73.2	72.6	72.8	72.4	71.7	72.0	72.7	73.4
Quartiles	74.5	72.6	73.3	71.7	72.5	71.9	71.5	73.4	75.1

Table 22: No. of Leaves of X^2-Stopped Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0 ¶
BUPA	Nat	44	41	43	41	35	35	32	32	6
	Qua	52	56	50	136	51	51	52	51	18
Finance 1	Nat	17	20	17	20	21	19	17	17	4
	Qua	13	13	13	13	13	13	13	13	5
Finance 2	Nat	8	8	8	8	14	8	8	8	4
	Qua	8	8	8	8	8	8	8	8	3
Solar Flare C		64	62	61	69	80	74	64	60	9
Solar Flare M		57	80	62	82	68	58	57	57	8
Solar Flare X		21	28	21	25	33	27	21	22	2
Glass	Nat	58	58	56	56	54	53	58	51	18
	Qua	64	68	64	69	65	66	64	63	21
Iris	Nat	11	11	11	11	11	11	11	11	9
	Qua	15	16	15	15	15	16	15	15	8
Obesity	Nat	13	16	13	18	13	13	13	13	8
	Qua	14	16	15	16	13	13	14	13	8
Pima	Nat	180	167	182	169	91	89	83	84	12
	Qua	200	253	202	244	201	200	199	200	28
Servo Motors		14	14	14	14	8	8	8	8	8
Soybean		4	4	4	4	4	4	4	4	4
Thyroid	Nat	20	22	21	22	28	27	20	22	8
	Qua	23	24	24	24	26	24	23	24	6
WAIS	Nat	5	5	5	5	5	5	5	5	2
	Qua	20	20	20	18	18	20	20	20	2
Wine	Nat	15	17	15	19	17	17	15	16	11
	Qua	13	14	13	19	13	13	13	14	10
Word Sense		233	226	234	226	259	251	233	230	73

¶ Stopped using P_0

Totals

Overall	1186	1267	1191	1351	1164	1123	1070	1061	295
Natural	371	365	371	369	289	277	262	259	82
Quartiles	422	488	424	562	423	424	421	421	109

Table 23: Wtd. Avg. Depth of X^2-Stopped Trees

Data Set		Gain	Gain	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
			Ratio							¶
BUPA	Nat	19.6	21.8	19.6	21.8	14.3	14.2	14.0	14.0	4.1
	Qua	4.1	4.3	4.1	8.5	4.1	4.1	4.1	4.1	4.8
Finance 1	Nat	5.1	9.2	5.3	6.2	6.5	5.9	5.1	5.1	1.7
	Qua	4.2	4.1	4.0	4.1	4.0	4.0	4.2	4.0	2.5
Finance 2	Nat	2.4	2.4	2.4	2.4	4.9	2.4	2.4	2.4	1.7
	Qua	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	1.7
Solar Flare C		7.5	9.4	8.0	10.3	10.3	7.8	7.5	7.4	3.9
Solar Flare M		6.3	10.0	8.3	10.6	10.3	7.8	6.3	7.9	3.8
Solar Flare X		3.6	4.5	3.6	4.5	6.1	4.5	3.6	3.7	1.0
Glass	Nat	8.8	18.0	9.0	17.0	9.7	9.5	8.8	7.6	4.6
	Qua	6.6	7.7	6.7	7.9	6.9	6.9	6.6	7.3	4.8
Iris	Nat	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1	2.4
	Qua	4.1	4.7	4.1	4.6	4.1	4.0	4.1	4.1	3.4
Obesity	Nat	5.3	7.1	5.3	8.0	5.3	5.3	5.3	5.3	3.6
	Qua	4.1	4.7	4.2	4.7	3.8	3.8	4.1	3.9	3.2
Pima	Nat	21.5	24.0	21.6	25.1	11.5	11.3	11.2	11.1	4.9
	Qua	8.3	10.1	8.6	10.0	8.3	8.1	8.2	8.0	5.1
Servo Motors		2.9	2.9	2.9	2.9	1.5	1.5	1.5	1.5	2.0
Soybean		2.0	2.4	2.0	2.4	2.0	2.0	2.0	2.0	2.0
Thyroid	Nat	5.7	11.0	7.8	11.1	6.8	6.6	5.7	7.8	4.4
	Qua	3.8	3.8	3.8	3.8	4.0	3.8	3.8	3.8	2.4
WAIS	Nat	3.3	3.3	3.3	3.4	3.4	3.3	3.3	3.3	1.0
	Qua	4.9	5.7	4.9	4.8	4.8	4.5	4.9	4.8	1.0
Wine	Nat	4.0	5.1	4.0	7.1	4.0	4.0	4.0	4.4	3.9
	Qua	3.8	4.0	3.9	5.3	3.8	3.8	3.8	4.2	3.7
Word Sense		15.8	34.8	16.5	44.4	21.9	18.3	15.8	22.3	10.7

¶ Stopped using P_0

Weighted Averages

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
Overall	9.7	14.0	10.1	15.8	9.6	8.6	8.1	9.1	4.8
Natural	13.6	16.6	13.9	17.1	9.4	9.2	8.9	9.0	4.1
Quartiles	5.8	6.7	5.9	7.5	5.9	5.7	5.8	5.8	4.2

Table 24: Run Time of X^2-Stopped Trees

Data Set		Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0 ¶
BUPA	Nat	248	276	280	249	212	213	205	194	66
	Qua	82	99	105	71	62	62	77	62	43
Finance 1	Nat	22	31	29	20	19	17	17	14	6
	Qua	16	19	21	12	10	11	12	10	5
Finance 2	Nat	10	12	13	8	13	8	9	7	5
	Qua	9	10	11	7	6	7	7	6	4
Solar Flare C		106	126	133	103	101	98	86	78	45
Solar Flare M		98	136	135	105	103	92	81	79	43
Solar Flare X		45	57	56	48	60	45	39	39	20
Glass	Nat	243	379	291	311	188	194	191	153	94
	Qua	123	155	163	104	87	91	94	79	42
Iris	Nat	14	16	17	12	12	13	12	11	9
	Qua	17	20	20	15	13	15	14	13	8
Obesity	Nat	69	96	89	68	46	49	53	46	28
	Qua	40	53	53	34	24	25	29	23	12
Pima	Nat	608	661	694	603	395	383	372	354	193
	Qua	354	449	452	322	285	272	292	254	140
Servo Motors		18	20	23	14	13	15	14	12	10
Soybean		12	16	16	11	9	9	10	10	7
Thyroid	Nat	62	98	83	83	64	71	54	64	42
	Qua	30	35	38	25	24	28	25	23	14
WAIS	Nat	1.5	1.6	1.7	1.3	0.8	1.3	1.3	1.2	0.6
	Qua	13	15	16	9	8	8	9	7	1.8
Wine	Nat	109	152	133	147	92	107	97	100	75
	Qua	67	79	82	76	54	62	57	55	37
Word Sense		2947	4939	3703	5213	3133	2924	2619	3076	1291

¶ Stopped using P_0

Totals

Overall	5364	7952	6658	7671	5034	4820	4478	4769	2242
Natural	1388	1724	1631	1502	1041	1057	1010	945	520
Quartiles	750	935	961	676	573	580	618	531	307

E.5 Comparisons of Stopping and Post-Pruning

Table 25: Wtd. Avg. Accuracy Comparisons

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
				All 26 Data Sets					
Post-Pruned	74.7	75.3	74.5	74.6	75.0	75.5	75.4	75.2	75.1
Unpruned	75.1	74.9	74.8	73.7	74.7	74.8	75.3	75.3	75.0
Pre-pruned	75.3	75.1	75.2	74.5	74.7	74.2	74.2	75.0	76.4
				Natural Cut-Points					
Post-Pruned	72.5	73.4	72.7	72.2	73.1	73.3	72.4	72.9	73.2
Unpruned	72.8	72.6	72.1	70.9	72.5	72.4	72.5	72.9	73.0
Pre-pruned	72.5	73.2	72.6	72.8	72.4	71.7	72.0	72.7	73.4
				Quartiles Cut-Points					
Post-Pruned	72.2	73.3	72.1	72.3	72.5	73.8	74.0	73.8	73.0
Unpruned	73.2	73.1	73.1	71.9	73.0	73.5	74.3	73.8	72.6
Pre-pruned	74.5	72.6	73.3	71.7	72.5	71.9	71.5	73.4	75.1

Table 26: Total No. of Leaves Comparisons

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
				All 26 Data Sets					
Post-Pruned	1155	1051	1039	1142	1141	1171	1025	1019	1203
Unpruned	1295	1267	1191	1351	1318	1272	1070	1061	1213
Pre-pruned	1186	1267	1191	1351	1164	1123	1070	1061	295
				Natural Cut-Points					
Post-Pruned	267	240	253	263	257	262	243	244	292
Unpruned	371	365	371	369	324	311	262	259	298
Pre-pruned	371	365	371	369	289	277	262	259	82
				Quartiles Cut-Points					
Post-Pruned	504	462	410	529	519	521	406	409	499
Unpruned	531	488	424	562	536	527	421	421	502
Pre-pruned	422	488	424	562	423	424	421	421	109

Table 27: Wtd. Avg. Depth Comparisons

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
				All 26 Data Sets					
Post-Pruned	7.6	10.0	8.1	10.3	8.4	8.1	7.7	8.6	7.2
Unpruned	9.9	14.0	10.1	15.8	10.3	9.1	8.1	9.1	7.3
Pre-Pruned	9.7	14.0	10.1	15.8	9.6	8.6	8.1	9.1	4.8
				Natural Cut-Points					
Post-Pruned	7.0	9.9	8.5	9.5	7.7	7.5	7.9	8.1	6.4
Unpruned	13.6	16.6	13.9	17.1	10.6	9.6	8.9	9.0	6.6
Pre-Pruned	13.6	16.6	13.9	17.1	9.4	9.2	8.9	9.0	4.1
				Quartiles Cut-Points					
Post-Pruned	6.4	6.2	5.7	7.0	6.5	6.5	5.7	5.8	6.2
Unpruned	6.5	6.7	5.9	7.5	6.7	6.5	5.8	5.8	6.2
Pre-Pruned	5.8	6.7	5.9	7.5	5.9	5.7	5.8	5.8	4.2

Table 28: Total Run Time Comparisons

	Gain	Gain Ratio	$1-d$	Ort	$W(1)$	$W(.5)$	G^2	X^2	P_0
				All 26 Data Sets					
Post-Pruned	7448	26868	9165	43506	9301	7787	6676	9794	5261
Unpruned	5852	7883	6585	7394	5004	4956	4490	4410	3028
Pre-Pruned	5364	7952	6658	7671	5034	4820	4478	4769	2242
				Natural Cut-Points					
Post-Pruned	1769	4669	2539	5783	2189	1960	1727	1717	1086
Unpruned	1236	1711	1606	1242	1036	1049	998	943	746
Pre-Pruned	1388	1724	1631	1502	1041	1057	1010	945	520
				Quartiles Cut-Points					
Post-Pruned	1244	1340	1255	1135	1017	1016	1000	866	867
Unpruned	778	936	959	655	621	652	606	536	428
Pre-Pruned	750	935	961	676	573	580	618	531	307

E.6 Multi-way Splits

Table 29: Accuracy & Complexity of Multi-way Splits

Data Set		Accuracy %					No. of Leaves				
		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
BUPA	Natural	55	57	59	51	57	137	142	26	14	10
	Quartiles	58	58	57	58	57	100	253	70	43	40
Finance 1	Natural	67	65	58	67	54	53	53	11	11	11
	Quartiles	52	71	71	52	56	37	37	13	7	4
Finance 2	Natural	95	92	91	91	92	29	29	16	6	6
	Quartiles	89	89	97	97	97	19	19	4	4	4
Solar Flare C		85	87	88	88	88	108	120	27	14	14
Solar Flare M		85	85	89	89	90	114	124	22	14	14
Solar Flare X		97	97	97	97	97	38	53	6	4	4
Glass	Natural	47	44	51	50	50	194	212	79	55	42
	Quartiles	66	69	62	62	57	141	152	52	38	35
Iris	Natural	95	93	97	97	97	26	26	11	11	11
	Quartiles	92	90	90	93	93	19	19	16	13	10
Obesity	Natural	51	58	56	62	56	49	49	33	33	33
	Quartiles	49	40	44	58	63	28	28	13	10	7
Pima	Natural	70	70	71	72	72	260	268	59	37	30
	Quartiles	57	67	71	74	74	502	496	112	61	52
Servo Motors		96	95	93	92	93	40	40	24	8	8
Soybean		98	98	96	98	98	6	6	6	6	6
Thyroid	Natural	91	91	91	91	91	56	56	32	32	18
	Quartiles	92	93	92	92	92	58	58	16	13	13
WAIS †	Quartiles	65	67	73	71	71	37	37	9	5	5
Wine	Natural	89	90	88	91	90	76	76	44	38	38
	Quartiles	90	92	90	93	91	34	37	22	16	16

	Weighted Averages					Totals				
Overall	75.5	76.1	77.1	77.5	77.4	2161	2390	723	493	431
Natural	71.0	71.1	72.6	71.9	72.2	880	911	311	237	199
Quartiles	71.9	72.7	73.1	75.0	74.1	938	1099	318	205	181

Table 30: Efficiency & Practicality of Multi-way Splits

Data Set		Wtd. Avg. Depth					Run Time (sec)				
		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
BUPA	Natural	4.4	4.3	2.1	1.5	1.0	15	11	8	7	7
	Quartiles	2.1	3.9	3.0	2.7	2.7	15	11	9	8	8
Finance 1	Natural	3.5	3.5	1.6	1.6	1.6	3.8	1.9	1.3	1.3	1.3
	Quartiles	2.5	2.5	1.8	1.3	1.0	3.0	1.5	1.3	1.1	1.0
Finance 2	Natural	2.0	2.0	1.7	1.0	1.0	2.4	1.3	1.3	1.1	1.0
	Quartiles	1.6	1.6	1.0	1.0	1.0	1.6	0.9	0.8	0.8	0.8
Solar Flare C		3.5	3.2	2.0	1.5	1.5	26	17	12	11	10
Solar Flare M		3.5	3.6	2.1	1.7	1.7	25	16	12	12	11
Solar Flare X		2.0	2.4	1.2	1.0	1.0	13	11	8	8	8
Glass	Natural	4.1	3.9	2.9	2.8	2.5	23	12	11	10	10
	Quartiles	3.7	3.6	2.9	2.6	2.5	21	10	9	8	8
Iris	Natural	1.4	1.4	1.2	1.2	1.2	2.2	1.7	1.5	1.4	1.7
	Quartiles	1.8	1.8	1.8	1.7	1.5	2.3	1.6	1.6	1.5	1.5
Obesity	Natural	2.0	2.0	1.7	1.7	1.7	9	3	3	3	3
	Quartiles	2.5	2.4	1.8	1.6	1.3	7	3	3	2	2
Pima	Natural	6.3	6.2	4.5	4.0	3.9	49	39	33	32	32
	Quartiles	4.2	4.1	3.2	2.8	2.6	57	33	29	26	25
Servo Motors		1.6	1.6	1.5	1.3	1.3	4	2	2	2	2
Soybean		1.3	1.3	1.3	1.3	1.3	5	2	2	2	2
Thyroid	Natural	2.4	2.4	2.2	2.2	1.8	6	4	4	4	4
	Quartiles	2.4	2.4	1.9	1.6	1.6	6	3	3	3	3
WAIS †	Quartiles	2.4	2.4	1.3	1.0	1.0	1.9	0.9	0.7	0.6	0.6
Wine	Natural	2.2	2.2	2.0	2.0	2.0	17	9	9	9	8
	Quartiles	2.3	2.3	2.1	2.0	2.0	12	7	7	6	6

	Weighted Averages					Totals				
Overall	3.5	3.6	2.6	2.3	2.1	327	203	173	162	159
Natural	4.3	4.2	3.0	2.7	2.5	127	83	72	68	67
Quartiles	3.1	3.3	2.7	2.4	2.2	126	71	63	58	57

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover Publications, Inc., New York, 1972.

[2] A. Agresti. *Categorical Data Analysis.* John Wiley & Sons, New York, 1990.

[3] K. Ali, C. Brunk, and M. Pazzani. Learning multiple relational rule-based models. In *Proceedings Fifth International Workshop on Artificial Intelligence and Statistics*, pages 8–14, Fort Lauderdale, FL, 1995.

[4] T. W. Anderson and S. L. Sclove. *The Statistical Analysis of Data.* The Scientific Press, Palo Alto, 2nd edition, 1986.

[5] G. A. Barnard. A new test for 2 × 2 tables. *Nature*, 156:177, 1945.

[6] G. A. Barnard. Significance tests for 2 × 2 tables. *Biometrika*, 34:123–138, 1947.

[7] G. A. Barnard. Statistical inference. *Journal of the Royal Statistical Society*, B11:115–149, 1949.

[8] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, MA, 1973.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1984.

[10] W. Buntine and T. Niblett. A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8:75–85, 1992.

[11] W. L. Buntine. *A Theory of Learning Classification Rules.* PhD thesis, University of Technology, Sydney, 1990.

[12] B. Cestnik and I. Bratko. On estimating probabilities in tree pruning. In *Proceedings of the European Working Session on Learning (EWSL-91)*, pages 138–150, Berlin, 1991. Springer-Verlag.

[13] B. Cestnik, I. Kononenko, and I. Bratko. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In *Progress in Machine Learning — Proceedings of EWSL 87: 2nd European Working Session on Learning*, pages 31–45, Wilmslow, 1987. Sigma Press.

[14] K. Church, W. Gale, P. Hanks, and D. Hindle. Using statistics in lexical acquisition. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build A Lexicon*, pages 115–164. Lawrence Erlbaum & Assoc., Hillsdale, NY, 1991.

[15] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3:261–284, 1989.

[16] W. G. Cochran. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23:315–345, 1950.

[17] W. G. Cochran. Some methods of strengthening the common χ^2 tests. *Biometrics*, 10:417–451, 1952.

[18] J. F. Elder. Heuristic search for model structure. In *Proceedings Fifth International Workshop on Artificial Intelligence and Statistics*, pages 199–210, Fort Lauderdale, FL, 1995.

[19] U. M. Fayyad. *On the Induction of Decision Trees for Multiple Concept Learning.* PhD thesis, University of Michigan, 1991.

[20] U. M. Fayyad, J. Cheng, K. B. Irani, and Z. Qian. Improved decision trees: A generalized version of ID3. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 100–108, San Mateo, CA, 1988. Morgan Kaufmann.

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

[22] U. M. Fayyad and K. B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8:87–102, 1992.

[23] R. A. Fisher. On the interpretation of chi-square from contingency tables. *Journal of the Royal Statistical Society*, 85:87–94, 1922.

[24] R. A. Fisher. The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, 98:39–82, 1935.

[25] R. A. Fisher. *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh, 14th edition, 1970. (the quotations are from the prefaces to the first (1925) and twelfth (1954) editions).

[26] G. H. Freeman and J. H. Halton. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38:141–149, 1951.

[27] M. Haber. An exact unconditional test for the 2 × 2 comparative trial. *Psychological Bulletin*, 99:129–132, 1986.

[28] M. Haber. A comparison of some conditional and unconditional tests for 2 × 2 contingency tables. *Communications in Statistics*, B16:999–1013, 1987.

[29] M. Haber. Do the marginal totals of a 2 × 2 contingency table contain information regarding the table proportions? *Communications in Statistics*, A18:147–156, 1989.

[30] J. A. Hartigan. *Bayes Theory.* Springer-Verlag, New York, 1983.

[31] A. Hutchinson. *Algorithmic Learning.* Graduate Texts in Computer Science, series no. 2. Clarendon Press, Oxford, 1994.

[32] L. Hyafil and R. L. Rivest. Constructing optimal binary decision trees is NP-Complete. *Information Processing Letters*, 5:15–17, 1976.

[33] J. O. Irwin. Tests of significance for differences between percentages based on small numbers. *Metron*, 12:83–94, 1935.

[34] G. R. Iversen. *Bayesian Statistical Inference.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-043. Sage Publications, Beverly Hills, 1984.

[35] J. J. L. Hodges and E. L. Lehman. *Basic Concepts of Probability and Statistics.* Holden-Day, Inc., Oakland, CA, 1970.

[36] M. James. *Classification Algorithms.* W. M. Collins & Sons, London, 1985.

[37] K. Koehler. Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, 81:481–493, 1986.

[38] K. Koehler and K. Larntz. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, 75:336–344, 1980.

[39] H. O. Lancaster. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56:223–234, 1961.

[40] K. Larntz. Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, 73:253–263, 1978.

[41] W. Z. Liu and A. P. White. The importance of attribute-selection measures in decision tree induction. *Machine Learning*, 15:25–41, 1994.

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

[43] J. K. Martin. Evaluating and comparing classifiers: Complexity measures. In *Proceedings Fifth International Workshop on Artificial Intelligence and Statistics*, pages 372–378, Fort Lauderdale, FL, 1995.

[44] J. K. Martin and D. S. Hirschberg. Small sample statistics for classification error rates. Technical Report 95-1, University of California, Irvine, Irvine, CA, 1995.

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

[46] J. Mingers. An empirical comparison of pruning measures for decision tree induction. *Machine Learning*, 4:227–243, 1989.

[47] J. Mingers. An empirical comparison of selection measures for decision tree induction. *Machine Learning*, 3:319–342, 1989.

[48] D. F. Morrison. *Multivariate Statistical Methods.* McGraw-Hill, New York, 3rd edition, 1980.

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

[50] P. M. Murphy and M. J. Pazzani. ID2-of-3: Constructive induction of m-of-n concepts for discriminators in decision trees. In L. A. Bernbaum and G. C. Collins, editors, *Machine Learning: Proceedings of the 8th International Workshop (ML91)*, pages 183–187, San Mateo, CA, 1991. Morgan Kaufmann.

[51] T. Niblett. Constructing decision trees in noisy domains. In I. Bratko and N. Lavrac, editors, *Progress in Machine Learning: Proceedings of the European Working Session on Learning (EWSL-87).* Sigma Press, Wilmslow, 1987.

[52] T. Niblett and I. Bratko. Learning decision rules in noisy domains. In *Proceedings of Expert Systems 86*, Cambridge, 1986. Cambridge U. Press.

[53] M. J. Pazzani. *Creating a Memory of Causal Relationships: An Integration of Empirical and Explanation-Based Learning Methods.* Lawrence Erlbaum, Hillsdale, NJ, 1990.

[54] K. Pearson. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In E. S. Pearson, editor, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge, 1948.

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

[56] J. R. Quinlan. Simplifying decision trees. In B. R. Gaines and J. H. Boose, editors, *Knowledge Acquisition for Knowledge-Based Systems.* Academic Press, San Diego, 1988.

[57] J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5:239–266, 1990.

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

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

[60] S. Rasmussen. *An Introduction to Statistics with Data Analysis.* Brooks/Cole Publishing Co., Pacific Grove, CA, 1992.

[61] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10:153–178, 1993.

[62] K. D. Tocher. Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika*, 37:130–144, 1950.

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

[64] S. M. Weiss and N. Indurkhya. Optimized rule induction. *IEEE Expert*, 8:61–69, 1993.

[65] A. P. White and W. Z. Liu. Bias in information-based measures in decision tree induction. *Machine Learning*, 15:321–329, 1994.

[66] S. S. Wilks. The likelihood test of independence in contingency tables. *Annals of Mathematical Statistics*, 6:190–196, 1935.

[67] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1938.

[68] F. Yates. Tests of significance for 2 × 2 contingency tables. *Journal of the Royal Statistical Society*, A147:426–463, 1984.