An Exact Probability Metric for Decision Tree Splitting and Stopping

J. Kent Martin
jmartin@ics.uci.edu

October 26, 1994

Abstract

ID3's information gain heuristic [16] is well-known to be biased towards multi-valued attributes. This bias is only partially compensated by the gain ratio used in C4.5 [20]. Several alternatives have been proposed, notably orthogonality [9], and Beta [5]. Gain ratio and orthogonality are strongly correlated, and all of the metrics share a common bias 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 posterior probability of the null hypothesis. Both gain and the chi-squared significance test 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 [3, pp 59-66] and ID3 [20, pp 36-37] 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 pre-pruning should be done, as trees pruned in this way are simpler, more efficient, and no less accurate than unpruned trees. Average training time is reduced by up to 30%, and expensive post-pruning avoided.

1 Introduction

Variants of the information gain heuristic used for the ID3 algorithm [16] have become the *de facto* standard metrics for attribute selection in top-down decision tree learning. This heuristic, or various modifications of it, is used (for instance) in FOIL [19], FOCL [15], CART [3], CN2 [7], GID3(*) [8], and C4.5 [20]. Fayyad [9] terms these impurity measures, and cites several studies [3, 12, 11] showing that the various members of this class are interchangeable (*i.e.*, they result in very similar decision trees).

Information gain calculates the difference (decrease) between the entropy of the population and the weighted average entropy of the subpopulations. The candidate split showing the largest gain is selected.

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

where

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
$f_{c,v}$	is the no. of those which are in class c
N	is the total no. in the population
n_c	is the total no. in class c

The gain ratio function used in C4.5 [20] partially compensates for the bias of gain towards splits having larger V.

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

Fayyad, *et al* [9] give an orthogonality metric for binary splits

$$ORT = 1 - \left(\sum_{c=1}^{C} f_{c,1} \cdot f_{c,2} \right) \Big/ \left[\left(\sum_{c=1}^{C} f_{c,1}^2 \right) \left(\sum_{c=1}^{C} f_{c,2}^2 \right) \right]^{1/2} \tag{3}$$

The Chi-squared statistic (see [2, pp. 452-462], [10, pp. 320-323], [21, 572-592])

$$X^2 = \sum_{c=1}^{C} \sum_{v=1}^{V} \frac{(f_{cv} - e_{cv})^2}{e_{cv}}, \qquad \text{where } e_{cv} = (n_c \, m_v / N) \tag{4}$$

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—i.e.*, assuming that the class distribution is independent of the split. This significance test is a good approximation when all of the e_{cv} are greater than 1 and no more than 20% of the e_{cv} are less than 5 (Cochran's rule).

Buntine [5] derives a Beta splitting rule

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

In which information gain appears as part of an asymptotic approximation to $W(t, \alpha)$. 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 attribute relevance asymptotically converge (rank attributes in the same order). Hence the repeated empirical findings that the various measures are largely interchangeable.

Any advantage that one metric might have over another is not to be found in the asymptotic behavior, but rather in the results obtained from small samples and samples with non-uniform class and attribute distributions. Such small or non-uniform samples are more likely to be found than are samples for which the asymptotic conditions hold, since the divide-and-conquer tree building (splitting) process itself tends to destroy the asymptotic conditions even when they do hold for the entire sample data set. (In Equation 4, for instance, $E(e_{cv})$ decreases exponentially with the number of splits—Cochran's rule requires minimally that $E(e_{cv}) \geq 5$, which *must* fail for some subsets after $\log_2(\mathcal{N}/5)$ splits (where $\mathcal{N}$ is the size of the entire data set).)

2 An Exact Significance Test

From a detailed Bayesian analysis, the posterior probability of the null hypothesis (H_0) is the multiple hypergeometric distribution

$$P_0 \equiv \text{Prob}(H_0 \mid \text{data}) = \left(\frac{\prod_{c=1}^{C} n_c!}{N!}\right) \prod_{v=1}^{V} \left(\frac{m_v!}{\prod_{c=1}^{C} f_{cv}!}\right) \tag{6}$$

Note that the second term here is very similar to Buntine's Beta function (inverted); and that, if $(\forall c, v: \ f_{cv} \neq 0 \text{ and } e_{cv} \gg 0)$, then

$$\begin{aligned} -2\ln(P_0) \ \approx \ & 2\ln(2)\, N \textbf{ gain} + (C-1)(V-1)\ln(2\pi N) \\ & - [C(V-1)\ln(C) + V(C-1)\ln(V)] \\ & -\frac{1}{2}\Big[C^2V^2(CV-1)\ \text{Var}(f_{cv}/N) \\ & - C^2(C-1)\ \text{Var}(n_c/N) \ - V^2(V-1)\ \text{Var}(m_v/N)\Big] \end{aligned} \tag{7}$$

alternatively,

$$\begin{aligned} -2\ln(P_0) \ \approx \ & \boldsymbol{X^2} + (C-1)(V-1)\ln(2\pi N) \\ & + [C(V-1)\ln(C) + V(C-1)\ln(V)] \\ & -\frac{(C-1)(V-1)}{2}\Big[C^2\ \text{Var}(n_c/N) + V^2\ \text{Var}(m_v/N)\Big] \\ & -\frac{1}{2}\sum_{c=1}^{C}\sum_{v=1}^{V}\Big[(f_{cv}-e_{cv})(f_{cv}-3e_{cv})\ /\ e_{cv}^2\Big] \end{aligned} \tag{8}$$

Thus, both *chi-squared and gain* arise as terms in *alternative approximations to the significance of a split.* In neither case should it be assumed that all the remaining terms vanish, even as $N \to \infty$. The crucial condition in Equation 7 dictates that the interaction weighted sum-of-squares should be small. In Equation 8 the similar condition is that the main sums-of-squares are small and e_{cv} largely dominate $\mid f_{cv} - e_{cv} \mid$. These terms are negative, indicating a *tendency for both measures to overestimate the significance of very non-uniform splits.*

Values of each of the measures (gain, gain ratio, orthogonality, X^2, Beta, and P_0) were calculated for 1,067 (binary classes, binary attribute) cases. These data confirm the analyses above:

- When X^2 is valid, $X^2 \approx -2.927 - 2\ \ln(P_0)$. When X^2 is not valid, it tends to be spuriously high, overestimating the significance of the split.
- A similar linear relation to $\ln(P_0)$ is found for the other measures when X^2 is valid, with an even stronger tendency to overestimate the split's significance when X^2 is not valid.
- Very high values of information gain and the other measures occur with high frequency when the null hypothesis cannot be rejected ($P_0 \geq 0.5$). Occurrence of these high values is very strongly correlated with circumstances under which the X^2 approximation is invalid.
- When X^2 is valid, all of the measures converge (tend to rank splits in roughly the same order, though differing in detail). When X^2 is invalid, the split rankings can be quite divergent.

Information gain, gain ratio, X^2, Beta, and orthogonality all tend to downplay the contribution of either the n_c (priors of the classes) or m_v terms (priors imposed by the split), or the contribution due simply to the number of partitions, or may downplay all three. By downplaying these priors the calculations do not fully take into account the possibility that the data might have been predisposed to be highly orthogonal (or of low averaged entropy, etc.) For such ill-conditioned data, these metrics entail a high likelihood of Type I error. The null hypothesis probability function P_0 appears to be a measure which properly incorporates all these factors, and may be a more suitable attribute selection metric than gain, gain ratio, X^2, Beta, or orthogonality.

3 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). Quinlan [16] originally proposed that the χ^2 significance test (Equation 4) be used to prevent this in ID3 by stopping the process of splitting a branch if the split so produced were not statistically significant; and Breiman, *et al* [3] initially searched for a stopping rule in the form of a minimum gain threshold. Both of these approaches were abandoned in favor of some form of post-pruning (either a cost-complexity [3, pp 65-81] or reduced-error [17] approach). There have been a number of studies in this area [4, 6, 11, 12, 13, 14, 22].

Section 2, above, proposes the P_0 function for attribute selection. This same measure might be used for pre-pruning (when it is deemed desirable to do so), and is a valid statistic even in cases when the χ^2 statistic is not. The previous negative results concerning pre-pruning appear to be due to use of different inadmissable approximate statistics for attribute selection and stopping, rather than to any inherent fault of pre-pruning. Use of the P_0 function for both selection and stopping might permit more efficient construction of decision trees without loss of predictive accuracy.

4 Empirical Comparisons of the Measures

Sixteen data sets were used, chosen to give a good variety of application domains, and a good mix of attribute properties (numeric *vs.* nominal, many attributes *vs.* few), sample sizes, hard *vs.* easy classification problems, and balanced *vs.* unbalanced priors. 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, and the particular method used has important consequences for both efficiency and predictive accuracy, and can interact with selection and stopping criteria in unpredictable ways.
- Orthogonality is defined (see Equation 3) strictly for binary splits, and each attribute having $V > 2$ distinct values must be converted to V binary attributes for this measure.

The hypergeometric function (and the other measures, as well) applies only when the cut-points are defined *a priori* (knowing only the attribute value). Defining the cut-points *ex post*, as in C4.5 [20, pp. 25-26] and CART [3, p. 108], directly contradicts the null hypothesis (that the class distribution is *a priori* independent of the subset membership). The modifications to the expression for P_0 necessary to accomodate *ex post* cut-points and full consideration of the efficacy of various

Table 1: Unpruned Trees, Binary Splits

Data Set	Accuracy			No. Leaves			Wtd Avg Depth			Train/Val (sec)		
	Gain	Ort	P_0	Gain	Ort	P_0	Gain	Ort	P_0	Gain	Ort	P_0
BUPA	63	58	62	52	53	116	4.1	4.2	7.5	63	52	66
Fin 1	72	77	75	13	13	13	4.2	4.1	4.0	12	9	7
Fin 2	86	91	92	8	8	8	2.5	2.5	2.5	6	5	4
Flare C	87	86	86	64	69	67	7.5	10.3	6.8	101	94	69
Flare M	86	82	85	57	82	58	6.3	10.6	6.8	93	97	66
Flare X	97	97	97	21	25	22	3.6	4.5	3.7	42	42	36
Glass	72	72	70	64	69	63	6.6	7.9	6.5	95	80	54
Iris	91	92	90	15	15	16	4.1	4.6	4.0	13	12	9
Obesity	47	51	42	14	16	13	4.1	4.7	3.8	30	25	15
Pima	68	67	65	200	209	217	8.3	9.4	8.0	282	262	215
Servo	95	96	95	14	14	14	2.9	2.9	2.2	19	13	10
Soybean	98	98	98	4	4	4	2.0	2.4	2.0	13	10	7
Thyroid	93	92	93	23	24	24	3.8	3.8	3.8	24	19	17
WAIS	61	65	65	20	18	19	4.9	4.8	4.2	8	7	4
Wine	93	89	89	13	19	13	3.8	5.3	3.8	52	60	38
Word	64	63	64	233	226	248	15.8	44.4	14.6	3570	5241	1666
	75.3	74.0	75.0	1186	1233	1213	9.7	15.5	7.3	5658	7269	3028

strategies for handling numeric attributes are planned topics for a future paper. In order to avoid bias in comparing the selection metrics, arbitrary cut-points at approximately the quartiles were used (approximate because the cut-points are not allowed to separate instances with equal values). To avoid confounding the present evaluation with questions of the relative efficacy of binary *vs.* multi-way splits, the data sets were all also converted to binary forms.

Only the three most different split metrics (gain, orthogonality, and P_0) were evaluated. In each experiment, a tree was grown using all of the instances. The accuracy of this tree was then estimated by 10-fold cross-validation. (Split the data set into 10 test sets. For each test set, build a tree using the other 90% of the data and determine its accuracy on the test set. Average accuracy over all 10 test sets.)

The results for the unpruned trees are summarized in Table 1. None of the small differences in accuracy between split metrics is significant. These data support the conjecture that *in every case* trees grown using the null hypothesis probability P_0 are more efficient, and *no less accurate* than the gain and orthogonality trees.

For gain and orthogonality the χ^2 stopping rule *never* stopped splitting for any of the data sets, even for $p = 0.999$. The effects of stopping based on P_0 are summarized in Table 2. The accuracy data are mildly concave, peaking at around the 95% confidence level. Only a summary of the complexity data is given, the results for individual data sets are entirely consistent with the overall results. These results strongly support the conjecture that growing and stopping decision trees using P_0 at the 95% confidence level *does no harm* and may, in fact, be mildly beneficial to accuracy. Training and validation time is reduced by 25-30% from the unpruned trees, and by 60% from the unpruned trees built using information gain (not including the time required to post-prune those trees).

Table 2: Stopping Effects on Accuracy

data set	Pruning Confidence Level						
	0	50	90	95	99	99.5	99.9
BUPA	62	65	59	57	64	62	54
Finance 1	75	79	79	79	71	64	¶ 44
Finance 2	92	88	97	97	92	97	94
Flare C	86	89	88	88	89	89	89
Flare M	85	85	89	90	90	90	90
Flare X	97	98	98	98	98	98	98
Glass	70	68	67	70	65	61	63
Iris	90	91	92	92	94	94	92
Obesity	51	42	49	49	40	¶ 29	¶ 36
Pima	65	68	73	73	74	74	75
Servo	95	93	91	89	89	90	¶ 81
Soybean	98	98	98	96	98	98	98
Thyroid	93	94	93	92	92	91	91
WAIS	65	67	65	63	63	74	76
Wine	89	90	89	88	86	89	85
Word Sense	64	65	66	66	67	65	64
	75.0	75.3	76.4	76.4	75.9	75.5	74.1
No. of Leaves	1213	895	406	295	192	164	125
Wtd Avg Depth	7.30	6.71	5.21	4.78	3.83	3.57	2.96
Train/Val Time (sec)	3028	2799	2387	2242	1960	1889	1733

¶ below the 95% confidence limit of unpruned accuracy

5 Conclusions

1. Information gain, gain ratio, orthogonality, and Beta each downplay some part of the influence of the number of partitions or the prior distributions. Whenever one or more of the expected values in a split is small, these measures (in common with χ^2) are prone to overestimate the significance of the split. The divide-and-conquer strategy of building decision 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 encountered when the classes and attribute values are unevenly distributed or the number of partitions large. The unpruned trees it builds are much more efficient, and *no less accurate*, than those built by the other measures.

3. The P_0 measure should also be used to stop splitting subtrees. The resulting trees are simpler and no less accurate than the unpruned trees. A stopping confidence level of 95 or 99% is recommended. Training times are reduced by about 30%, and expensive post-pruning steps avoided entirely.

References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions.* Dover, New York, 1972. (Originally published by the U.S. National Bureau of Standards in 1964 as *No. 55, Applied Mathematics Series.* Corrected edition of 10th (1972) U.S. Government Printing Office printing.).

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

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Monterey, CA, 1984.

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] J. L. Hodges, Jr. and E. L. Lehmann. *Basic Concepts of Probability and Statistics.* Holden-Day, Oakland, CA, second edition, 1970.

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

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

[13] 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.

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

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

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

[17] J. R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234, 1987. (reprinted in [18]).

[18] 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.

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

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

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

[22] C. Schaffer. When does overfitting decrease prediction accuracy in induced decision trees and rule sets? In *Proceedings of the European Working Session on Learning (EWSL-91)*, pages 192–205, Berlin, 1991. Springer-Verlag.