Machine Learning, 15, 321–329 (1994)

Technical Note

Bias in Information-Based Measures in Decision Tree Induction

ALLAN P. WHITE A.P.WHITE@BHAM.AC.UK
Computer Centre, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, United Kingdom[1]

WEI ZHONG LIU W.Z.LIU@BHAM.AC.UK
School of Mathematics and Statistics, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, United Kingdom

Editor: J.R. Quinlan

Abstract. A fresh look is taken at the problem of bias in information-based attribute selection measures, used in the induction of decision trees. The approach uses statistical simulation techniques to demonstrate that the usual measures such as information gain, gain ratio, and a new measure recently proposed by Lopez de Mantaras (1991) are all biased in favour of attributes with large numbers of values. It is concluded that approaches which utilise the chi-square distribution are preferable because they compensate automatically for differences between attributes in the number of levels they take.

Keywords: Decision trees, noise, induction, unbiased attribute selection, information-based measures

1. Introduction

The task of inducing a decision tree is typically handled by a recursive partitioning algorithm which, at each non-terminal node in the tree, branches on that attribute which discriminates best between the cases filtered down to that node.

Traditionally, the measure called 'information gain' by Quinlan (1986), has been using for the purpose of deciding which of the available attributes should be branched on. However, in recent years, researchers such as Kononenko et al. (1984) have become aware that this measure is liable to favour unfairly attributes with large numbers of values at the expense of those with few.

As a preferred alternative, Quinlan (1986) suggested the use of another information-based measure which he termed the 'gain ratio'. This was derived from information gain by dividing by attribute information. This acted as a sort of normalising factor by virtue of the fact that attribute information tends to increase as the number of possible values increases. At the time, it was thought that this would eliminate the bias. However, it was acknowledged that attributes with very *low* information values (i.e. low *attribute* information) then appeared to gain an unfair advantage.

More recently still, Lopez de Mantaras (1991) proposed another information-based criterion which, it was claimed, was free of this latter problem also. The purpose of this paper is to take a fresh look at the problem of bias in measures which should preferably be bias-free in deciding between contending attributes with different numbers of values.

2. Definitions of the measures

Although the various measures have been defined elsewhere, it is felt that a simpler approach to the definitions than is given in other papers would be beneficial.

Suppose that we are dealing with a problem with k classes and that an attribute, A, with m distinct values is under consideration at a particular node. The following contingency table (Table 1) represents the cross-classification of classes and attribute values:

Table 1. A general contingency table.

	a_1	a_2		a_m	
C_1	n_{11}	n_{12}	...	n_{1m}	$n_{1.}$
C_2	n_{21}	n_{22}	...	n_{2m}	$n_{2.}$
⋮	⋮	⋮	...	⋮	⋮
C_k	n_{k1}	n_{k2}	...	n_{km}	$n_{k.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.m}$	$n_{..}$

where C_i $(i = 1, k)$ and a_j $(j = 1, m)$ represent class and attribute values respectively; n_{ij} $(i = 1, k;\ j = 1, m)$ represent the frequency counts of cases with attribute value a_j and class C_i; and:

$$n_{i.} = \sum_{j=1}^{m} n_{ij}$$

$$n_{.j} = \sum_{i=1}^{k} n_{ij}$$

$$n_{..} = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} = N$$

Various probabilities can be defined, as follows:

$$p_{ij} = \frac{n_{ij}}{n_{..}}$$

$$p_{i.} = \frac{n_{i.}}{n_{..}}$$

$$p_{.j} = \frac{n_{.j}}{n_{..}}$$

Given an event with l possible outcomes, each of probability p_i (i = 1, 2, ..., l), the information associated with this event is (Edwards, 1964):

$$-\sum_{i=1}^{l} p_i \log_2 p_i$$

Applying this to the contingency table above, we can define the information associated with each possible cell, class and attribute, respectively, as:

$$H_{cell} = -\sum_{i=1}^{k}\sum_{j=1}^{m} p_{ij} \log_2 p_{ij} \tag{1}$$

$$H_C = -\sum_{i=1}^{k} p_{i.} \log_2 p_{i.} \tag{2}$$

$$H_A = -\sum_{j=1}^{m} p_{.j} \log_2 p_{.j} \tag{3}$$

From these quantities, transmitted information may be defined as:

$$H_T = H_C + H_A - H_{cell} \tag{4}$$

The concept of transmitted information is very useful. In the current context, it means *the information about class membership which is conveyed by attribute value.*

Each of the information-theoretic measures can now be expressed in terms of the quantities defined in Equations 1 to 4. Firstly, it should be noted that Quinlan's 'information gain' measure is *identical* to *transmitted information*, H_T. The 'gain ratio', G_R, is simply transmitted information 'normalised' by attribute information:

$$G_R = \frac{H_T}{H_A} \tag{5}$$

Perhaps it should be mentioned that Quinlan (1986) favours an application of the gain ratio which, in his words:

> ... selects, from among those attributes with an average–or–better gain, the attribute that maximises the above ratio.

The distance measure proposed by Lopez de Mantaras (1991), d_N, is:

$$d_N = 1 - \frac{H_T}{H_{cell}} \tag{6}$$

This needs to be *minimised* over attributes, rather than maximised like the other measures, so it seems preferable to discuss the *complement* of this measure, $1 - d_N$, which is simply transmitted information 'normalised' by cell information.

$$1 - d_N = \frac{H_T}{H_{cell}} \tag{7}$$

Another measure of interest is the G statistic, described by Mingers (1987, 1989), as:

$$G = 2NH_T$$

Unfortunately, Mingers is in error.[2] The problem arises from the fact that Kullback (1959, pp. 158-159) was working with logarithms to base e, whereas all the other information-based measures use logarithms of base 2, in order that all the quantities concerned are expressed in bits. Therefore, the correct definition should be:

$$G = 2NH_T \log_e 2 \tag{8}$$

We also need to define a statistical measure, χ^2, from the same table:

$$\chi^2 = \sum_i \sum_j \frac{(E_{ij} - O_{ij})^2}{E_{ij}} \tag{9}$$

where O_{ij} is the observed number of cases with value a_j in class C_i, i.e. $O_{ij} = n_{ij}$, and E_{ij} is the expected number of cases which should be in cell (C_i, a_j) in the contingency table, if the null hypothesis (of no association between attribute and class) is true:

$$E_{ij} = \frac{n_{.j} n_{i.}}{n_{..}}$$

Both the G statistic and χ^2 are well approximated by the chi-square distribution with ν degrees of freedom, where:

$$\nu = (k-1)(m-1)$$

However, it should be remembered that this will not be true for the G statistic if logarithms of the wrong base are used. It should also be mentioned that *both* approximations become poor with small expected frequencies. This fact is well documented in the case of χ^2. For example, Siegel (1956) recommends that the χ^2 test should not be used if more than 20% of the expected frequencies are less than 5, or any are less than 1. If this warning is not heeded then, in this type of situation, the probability derived from the chi-square distribution will be smaller than the true probability of getting a value of χ^2as large as that obtained. This means that the χ^2test becomes over-optimistic in detecting informative attributes under these circumstances.

3. The problem of bias

The whole problem of bias arising from differences in the number of levels of attributes has not been adequately addressed in the area of machine learning. Until now, arguments for or against the existence of this type of bias have been based on a particular type of

argument which relies on randomly partitioning an attribute A to produce a derived attribute A' which has a larger number of values.

This line of argument was initiated by Quinlan (1986, 1988) who showed that:

$$H_T(A') \geq H_T(A)$$

This means that, in general, the derived attribute will transmit more information about class membership than the original one. However, as the additional partitioning required to derive A' from A was random, A' cannot be reasonably be preferred to A as a candidate for branching. Thus, the information measure, H_T, is not comparable between attributes which have different numbers of values.

More recently, Lopez de Mantaras (1991) proved that:

$$d_N(A') \geq d_N(A)$$

This result is then taken to imply that this distance measure does not favour attributes with large numbers of values. In fact, this type of proof is inappropriate. What is needed is a *statistical* approach which takes into account the *distributional* properties of the measures. More precisely, for any attribute selection measure, f, to be fair requires that, under the null hypothesis of no association between class and attribute:

$$p(f(x') \geq f(A')) = p(f(x) \geq f(A))$$

where x' and x represent general attribute variables with the same number of attribute values as A' and A respectively. This means that, under the null hypothesis, the probabilities of getting values for f greater than or equal to those actually obtained must be equal for A and A'. Anything other than equality in this equation would mean that attributes with larger numbers of values would be favoured at the expense of those with fewer, or vice versa.

The reason that a statistical approach is preferable is concerned with the risk of including in the tree attributes which do not provide genuine discrimination between the classes. Previous work by Liu and White (1994) has shown the importance of the attribute selection measure discriminating between attributes which are genuinely informative concerning class membership and those that are not. If the attribute selection measure is biased towards variables with large numbers of values, then noise variables with large numbers of values could be in contention for selection with genuinely informative attributes with fewer values. In general, this would lead to poorer predictive performance from the induced tree. For optimal predictive performance, the attribute selection process should avoid the selection of noise variables, because of their degrading effect on performance. Similar remarks can be made concerning the suboptimality of selecting attributes with large numbers of values which discriminate only weakly between the classes, when more powerful discriminators are available among those attributes with fewer distinct values.

4. Demonstration of bias by simulation

4.1. Introduction

The issue raised in the previous section is really a matter of the distribution of the test statistic differing according to the number of levels of the attribute under consideration. The proposal here, is to use Monte Carlo simulation techniques to explore these differences and thereby expose the bias (if any) in the use of various information-based measures. The point is that each of the information-based measures is affected by the number of cells in the class $\times$ attribute contingency table. Consequently, changing either the number of attribute values or the number of classes would be expected to affect the magnitude of the measure. Of course, for any given application, the number of classes is fixed. However, we would expect the effect of number of attribute values to be present whatever the number of classes. On the other hand, the probability-based measures are appropriately parameterised for the number of cells in the contingency table and hence would not be affected by either of these factors.

The intention here is to use simulation techniques to derive approximations to the theoretical central distributions for the various information-based test statistics, in order that the estimated parameters derived from these distributions may be compared for attributes with different numbers of values, m. The *central* distribution of a test statistic is the distribution of that statistic when no effect is operating in the population from which the samples are drawn. In the current context, this means that we are concerned with the distribution of a particular test statistic *when there is no actual association between class and attribute in the population* from which the samples are drawn. For a given measure, if the distributions differ significantly according to m, then bias is present in that measure.

The demonstration described below was designed to illustrate the presence of such a bias and also to show the effects of class probability and number of classes on this bias.

4.2. Method

The basis of the demonstration involved simulating attributes with different numbers of values, drawn from populations that had no association with class.

Three different conditions were employed, as follows:

1. two equiprobable classes
2. two classes with an odds ratio of 4:1
3. five equiprobable classes

A sample size of 600 cases was used, with the number of cases fixed as belonging to each of the classes, according to the condition just described. Class membership was cross-tabulated against three attributes, having respectively two, five and ten values. For each attribute, the values were generated independently for each class, from the appropriate discrete uniform distribution.

Table 2. Means for the various measures, for each condition.

Condition		Attribute Selection Measure				
		H_T	G_R	$1-d_N$	$p(\chi^2)$	$p(G)$
1	m=2	0.0012	0.0012	0.0006	0.5005	0.5023
	m=5	0.0048	0.0021	0.0015	0.5006	0.5019
	m=10	0.0107	0.0032	0.0025	0.4869	0.4910
2	m=2	0.0012	0.0012	0.0007	0.4989	0.5002
	m=5	0.0049	0.0021	0.0016	0.5026	0.5049
	m=10	0.0109	0.0033	0.0027	0.4960	0.5029
3	m=2	0.0049	0.0049	0.0015	0.5052	0.5064
	m=5	0.0196	0.0084	0.0042	0.5058	0.5132
	m=10	0.0441	0.0133	0.0079	0.4980	0.5189

1000 Monte Carlo trials were employed. On each trial, each of the information-based measures, H_T, G_R and $1-d_N$ were calculated for each attribute. In addition, χ^2 and G were also calculated and the probabilities for getting values as extreme as those obtained (denoted by $p(\chi^2)$ and $p(G)$, respectively) were derived from the cumulative chi-square distribution with $(k-1)(m-1)$ degrees of freedom.

4.3. Results and discussion

Means for each of the five measures, for each value of m are displayed in Table 2. The following points should be noted:

1. The results show clearly that, for each condition, the mean value for each of the first three measures increases as m is increased. Conversely, the means for the two probability-based measures show no tendency to vary systematically with m. The significance of these findings was checked by performing F tests for the application of each measure to each condition. The results of these were so clear that it was felt unnecessary to quote each one separately. Briefly, for the three information-based measures, the F ratios for the three conditions ranged from 443 to 6727, with $2, 2997$ degrees of freedom. Even the smallest of these was significant beyond the 0.001 level. By contrast, the corresponding F ratios for the two probability-based measures ranged from 0.07 to 0.75, giving p values all greater than 0.4.

2. The results also show clearly that, for any particular value of m, the means for the first three measures do not really differ between the first two conditions but are substantially higher in the third condition. By contrast, the means for the two probability-based measures do not really differ between *any* of the conditions.

The reasons underlying these findings are as follows. Both χ^2 and H_T (and the other information-based measures derived from it) are quantities whose distributions are parameterised by the number of degrees of freedom of the contingency tables from which

they have been derived. For a contingency table with k rows and m columns, the number of degrees of freedom is given by $(k-1)(m-1)$. The practical consequence of this fact is that measures derived from tables with different numbers of cells are not directly comparable, because they have different probability distributions. This is why the means for the first three measures in Table 2 increase as m is increased. It also explains why the means for these measures are higher for the third condition than for the other two, i.e., because the number of classes has been increased. On the other hand, changing the class probabilities while keeping the number of classes constant does *not* produce changes in these measures because this operation does not change the number of cells in the contingency table.

By contrast, the two probability-based measures do not suffer from these problems because the manner in which they are calculated takes into account the number of degrees of freedom. This means that probabilities derived in this way from tables with different numbers of cells *are* directly comparable.

For comparison purposes, a brief test was also made of the behaviour of Quinlan's variant of the gain ratio on the first experimental condition, i.e. with two equiprobable classes. (As Quinlan did not define what he meant by 'average', this was calculated using the median value for H_T as the first stage in the computation). The results showed the same tendency as the other information-based measures, with means of 0.00206, 0.00314 and 0.00439 for 2, 5 and 10 attribute values, respectively.

5. Conclusions

The simulation demonstration, just described, shows convincingly that H_T, G_R and $1 - d_N$ each favour attributes with larger numbers of values. The results suggest that H_T (transmitted information) is the worst of the three measures in this respect and also that G_R is the least biased. Furthermore, the results also show that the magnitude of this bias is not affected by class probability but is strongly dependent on the number of classes, increasing as k is increased.

The nature of the problem is this. The central distribution for H_T is really dependent on m and so is that of the 'normalised' measures derived from it. These distributions are also dependent on k. In fact, *each of the information-based measures is dependent on the number of cells in the contingency table.* Of course, for attribute selection purposes in real inductive applications, the dependence on k is not important because the number of classes is fixed for the application concerned. However, the dependence on m remains and failure to take this fact into account in the proper way means that bias is operating in situations where various attributes differ in the number of values that they take.

The simple demonstration provided in this paper also shows a way out of the problem. Either χ^2 probabilities should be used or, if information-theoretic measures are preferred, then the G statistic (as defined in Equation 8), could be used instead. In either case, the statistic is distributed approximately as the chi-square distribution with $(m-1)(k-1)$ degrees of freedom. As the results show, comparing the resulting probabilities offers a simple and fair way of evaluating the relative importance of discrete attributes having different numbers of levels. The problem of small expected frequencies that was men-

tioned earlier can be handled in one of two ways. In the case of problems with just two classes, Fisher's exact probability test (Siegel, 1956) can be used in place of the χ^2test. For more than two classes, a similar approach could be developed.

Notes

1. A.P. White is also an Associate Memober of the School of Mathematics and Statistics at the University of Birmingham.
2. In fact, the arithmetic examples in Mingers (1987) are correct because he uses natural logarithms for his computation of information gain.

References

Edwards, E. (1964). *Information Transmission, An Introductory Guide to the Application of the Theory of Information to the Human Sciences*. London: Chapman and Hall.

Keppel, G. (1973). *Design and Analysis: a Researcher's Handbook*. Englewood Cliffs, N.J: Prentice-Hall Inc.

Kononenko, I., Bratko, I., & Roskar, E. (1984). *Experiments in automatic learning of medical diagnostic rules*. (Technical Report). Ljubljana, Yugoslavia: Jozef Stefan Institute.

Kullback, S. (1959). *Information Theory and Statistics*. New York: John Wiley and Sons.

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

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

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

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

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

Quinlan, J. R. (1988). Decision trees and multi-valued attributes. *Machine Intelligence 11*, 305–318.

Siegel, S. (1956). *Nonparametric Statistics*. New York: McGraw-Hill.

Received March 31, 1993
Accepted August 17, 1993
Final Manuscript October 28, 1993