Machine Learning, 15, 25-41 (1994)

The Importance of Attribute Selection Measures in Decision Tree Induction

W.Z. LIU (W.Z.LIU@BHAM.AC.UK)
School of Computer Science, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, United Kingdom

A.P. WHITE (A.P.WHITE@BHAM.AC.UK)
School of Computer Science, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, United Kingdom

Editor: J.R. Quinlan

Abstract. Recent work by Mingers and by Buntine and Niblett on the performance of various attribute selection measures has addressed the topic of random selection of attributes in the construction of decision trees. This article is concerned with the mechanisms underlying the relative performance of conventional and random attribute selection measures. The three experiments reported here employed synthetic data sets, constructed so as to have the precise properties required to test specific hypotheses. The principal underlying idea was that the performance decrement typical of random attribute selection is due to two factors. First, there is a greater chance that informative attributes will be omitted from the subset selected for the final tree. Second, there is a greater risk of overfitting, which is caused by attributes of little or no value in discriminating between classes being "*locked in*" to the tree structure, near the root. The first experiment showed that the performance decrement increased with the number of available pure-noise attributes. The second experiment indicated that there was little decrement when all the attributes were of equal importance in discriminating between classes. The third experiment showed that a rather greater performance decrement (than in the second experiment) could be expected if the attributes were all informative, but to different degrees.

Keywords. decision trees, noisy data, induction, attribute selection.

1. Introduction

The induction of decision trees for noisy domains has received fresh attention in the last few years, partly as a result of the somewhat belated recognition of the statistical work carried out by Breiman, Friedman, Olshen, and Stone (1984) on classification trees and partly as a result of work appearing in the machine learning literature by authors such as Quinlan (1986).

More recently, Mingers (1989a) made the surprising claim that *random* selection of attributes, followed by pruning, can achieve the same level of classification accuracy as the use of any of a variety of orthodox measures followed by pruning. He ran experiments with four data sets in which he tested various measures (including information gain, χ^2, the G statistic, the Gini index of diversity, and gain ratio) against a purely random selection method.

W.Z. Liu is now in the School of Mathematics and Statistics at the University of Birmingham (address as above). A.P. White is also an Associate Member of the School of Mathematics and Statistics at the University of Birmingham.

In each case, the resulting tree was pruned using Breiman's error complexity method. The results appeared to show that there was no significant difference in the classification performance, whichever method was used.

Subsequently, Buntine and Niblett (1992) refuted this claim with more carefully constructed experiments and suggested reasons for the disparity between their respective results.

However, their contribution does not exhaust the topic. The remainder of this article seeks to investigate the decrement expected in classification accuracy when random attribute selection is employed and to examine factors that might be expected to influence the magnitude of this decrement.

The following experiments test specific hypotheses in this area and, because of the fact that data sets with particular, precisely defined characteristics were required, synthetic data sets were used, rather than the real data sets employed by the previous investigators quoted.

2. Experimental techniques

The experiments described later in this article use a number of techniques, some of which may not be familiar to researchers in the machine learning field. Brief descriptions of these are given below.

2.1. Attribute selection

A number of measures have been reviewed comprehensively by Mingers (1989a). Breiman et al. (1984) also describe various methods in detail. The method used in this article is sometimes known as *transmitted information* (H_T) and sometimes as *information gain*. This method was chosen more for reasons of tradition (because of the origin of the study of inductive systems in computer science and the use of communication theory and information theory in that discipline) than because of any intrinsic superiority in the measure. The definition is given in Mingers (1989a). (Note that, if logarithms to base 2 are used, then the information gain is actually measured in bits.)

This method is contrasted with purely random attribute selection, as used by Mingers (1989a).

2.2. Binary splitting

One awkward problem with a completely general approach to the construction of classification trees is that there are many possible types of attribute. First, attributes may belong to any of four levels of measurement, as described originally by Stevens (1946) and mentioned briefly by Mingers (1989a). A further problem with ordered variables (i.e., those measured on either an ordinal or an interval scale of measurement) is that they may have tied scores, i.e., the scores may be grouped into categories. As an example, consider a large database of hospital patients. If each patient's age is recorded in years, then there will generally be more than one person at each age level—at least for the more commonly occurring ages in such a population. Statisticians sometimes refer to such variables as *ordered categorical.*

One way of dealing with a wide variety of variables in a classification problem is to reduce them to a common denominator by employing a binary branching technique on *all* the attributes, regardless of type. This approach has been used by a number of researchers, including Breiman et al. (1984), Kononenko, Bratko, and Roskar (1984), Quinlan (1988), White (1987), and White and Liu (1990). All attributes that are not originally binary are converted into "pseudo-binary" attributes by the technique of optimal splitting, as described below. Continuous attributes can be dealt with by splitting the initial attribute between every possible pair of adjacent values (in a sense of numerical order) to yield a number of derived binary variables as candidates to replace the initial non-binary attributes. Considering a pair of adjacent values of a continuous attribute, x_1 and x_2, the average of these two values, x_{12}, is regarded as a splitting point. All other values of this attribute are either less than or equal to ($\leq$) x_{12} or greater than ($>$) x_{12}. In this way, the derived variable is obviously binary. During the construction of the tree, if the original attribute has m distinct values present at the node currently under consideration, this would mean generating $m - 1$ candidate binary variables. The best of these $m - 1$ variables, as judged by some appropriate criterion (which may or may not be the same as that used for attribute selection), then becomes the pseudo-binary attribue that is used in place of the original attribute. In more detail, suppose that H_T is the criterion employed, each of the $m - 1$ derived variables is cross-tabulated against class for all the cases at the node, and a H_T value is calculated for each of them. The variable with the largest H_T is chosen as the pseudo-binary attribute to represent the original attribute.

For categorical attributes, a different variant of the same technique is employed. However, this is not relevant here because ordered attributes are used in all the experiments described.

Perhaps it should also be mentioned that the binary splitting technique allows the possibility that a multi-valued attribute may legitimately be branched on more than once (at different cutting points) in the same path of the decision tree.

2.3. Pruning

Pruning methods are employed to cut back a full-size tree to a smaller one that is likely to give better classification performance. These techniques have been mentioned by Breiman et al. (1984), Niblett and Bratko (1987), White (1985, 1987), White and Liu (1990), and Liu and White (1991). A comprehensive review of pruning methods for decision trees has been given by Mingers (1989b), and for this reason, only a brief mention is made of them here.

The pruning approach is more commonly used in this field and, indeed, there is a good reason to prefer it to a simple stopping rule applied to the growth phase of tree construction. This is because situations can arise in which, at the stage the tree is being grown, it can appear that all significant attributes have been exhausted. However, if growth is allowed to continue, further attributes can show up as important. The simple explanation for this is that the growing tree has uncovered a multiplicative relationship between two (or more) attributes and class.

Pruning methods can be implemented using statistical significance tests. Thus, a significance test is used to determine when to stop "undoing" the branching process. Pruning by significance testing was used in the simulation experiments described later in this article.

2.4. Cross-validation and dynamic path generation

The fair estimation of predictive accuracy is of central importance in the assessment and comparison of classification technique when noise is present. A particularly thorough way of doing this is to employ the technique of cross-validation, in which each case is tested under a model derived from all the *remaining* observations. Cross-validation is described in detail by Breiman et al. (1984). Of course, in the case of decision trees, the model involved is the tree derived from all the observations *except* the one being tested.

One problem with cross-validation is the fact that it is computationally expensive. However, the computing time required can be reduced substantially by combining cross-validation with a technique called dynamic path generation (White, 1987). Briefly, this involves generating just the path required for classifying the case currently under consideration, rather than the entire tree. Thus, in order to cross-validate a data set of N cases, it is only necessary to generate N paths. All the cross-validation results reported in the experiments described in this article were derived using this approach.

2.5. Statistical techniques

The experiments reported in this article employed various experimental designs, each of which is associated with a corresponding analysis of variance (ANOVA). Descriptions of these various designs may be found in Keppel (1973). ANOVA summary tables are reported for all designs with more than one factor, and F tests for single factor designs are reported in the text. By their very nature, F tests are multi-sided in terms of the hypotheses that they test. However, for the applications quoted here, it should be clear from inspection of the corresponding means where the important differences lie. Two-tailed t tests were also employed for parts of the analyses. These test whether or not there is a significant difference in *either* direction between two sets of results. This approach is more statistically conservative than the use of one-tailed tests and is the method generally employed.

3. Experiment I

3.1. Introduction

As stated in the previous section, Mingers (1989a) asserts that the choice of goodness of split measure is unimportant in determining the accuracy of predictive performance of probabilistic classification trees, even to the extent that a purely random attribute selection measure will perform as well as any of the orthodox methods.

In the general case, this cannot be true and, indeed, Buntine and Niblett (1992) showed that it was not, even for the data sets that Mingers himself used. It is instructive to consider why. Provided that a full-sized tree is not being used, i.e., that the tree contains only a proper subset of the available attributes, then it would seem to be important which attributes are selected for membership of this subset. Suppose that some variables contain more information about class membership than others. Clearly, variables that are high in class information are more important to include in the subset of variables used for branching than variables that are low in class information. For the purposes of demonstration, this argument can be taken a stage further. Suppose that only one of the attributes contains information about class and that the remaining attributes are pure noise variables. In this situation, it is obviously of vital importance that the informative variable is included in the branching subset. It is also important (although perhaps not quite as obviously) that as few of the pure noise variables as possible are branched on, because their presence will degrade the true classification accuracy of the induced tree.

Thus, if an effective attribute selection method is being used, then the number of available pure noise attributes should have little effect on classification performance. On the other hand, if a random selection method is employed, then classification performance should decline as the number of available pure noise variables is increased. This is for two reasons. First, the greater the number of pure noise variables available, the smaller the chance of a genuinely informative variable being included in the tree, because of the pressure of competition. Second, as mentioned in the previous paragraph, the more noise variables included in the final tree, the more classification performance will be degraded. The performance with the random method should always be poorer than that obtained from using an orthodox measure, but the difference would be expected to be greater as more pure noise variables are available for inclusion in the tree.

From the foregoing argument, it is obvious that the choice of orthodox selection measure should not be of critical importance. Transmitted information was chosen for this experiment more for traditional reasons than for any more fundamental motive.

3.2. Method

A Monte Carlo simulation experiment was designed in which two different measures of attribute selection were tested. One was transmitted information (H_T), and the other was a purely random selection criterion. Six different conditions were employed. For each condition, 100 different data sets were generated. Each data set consisted of 100 cases. Each case contained a binary class variable and a number of continuous independent variables (attributes). The class variable was generated so as to have 50 cases of each class.

Two different types of independent variable were employed. One type (termed *signal* variables) incorporated information about class membership. The other type (*noise* variables) was produced so as to contain no information about class. All the noise variables were generated as samples from the standard normal distribution (i.e., a normal distribution with zero mean and unit standard deviation), using a random number generator. Signal variables were derived from noise variables by the simple method of adding twice the class variable. (This would result in an expected difference of 2 in the mean scores on a signal variable between the two classes). This ensured that the signal variable contained information about class membership.

Each of the six conditions employed just one signal variable. However, the conditions differed in the number of noise variables. The arrangements were as follows:

1. 1 signal variable plus 1 noise variable;
2. 1 signal variable plus 5 noise variables;

3. 1 signal variable plus 10 noise variables;
4. 1 signal variable plus 20 noise variables;
5. 1 signal variable plus 40 noise variables;
6. 1 signal variable plus 80 noise variables.

Since the independent variables were continuous, the binary splitting approach (described earlier) was used. It was implemented using transmitted information. Thus, during the tree construction phase, at each node, each attribute was split at a value that would maximize the information that it provided about class membership. Attribute selection was then made from among the pseudo-binary attributes thus derived, according to the attribute selection measure employed in that part of the experiment (H_T or random). For reasons of speed, the method of dynamic path generation (described earlier) was employed. In classifying each case, path growth was continued until a pure terminal node was reached, i.e., one with cases from only one class.

This was followed by a pruning phase, which was implemented using the χ^2 statistic and the associated probability from the Chi-square distribution. The threshold value was set at a probability of 0.1. Thus, for any given path, pruning was continued until a significant association between class and the current attribute was uncovered, i.e., one in which the associated χ^2 probability fell below the threshold value. Pruning of this path was then terminated at this point, without undoing the branching at the node at which the probability fell below the threshold value.

A split-plot experimental design (Keppel, 1973, pp. 433–437) was employed, in which the data set was the basic unit of replication. For each of the six conditions (described earlier), both attribute selection methods were applied to the same data sets. Thus, in the language of experimental design, conditions were varied *between* data sets, whereas the attribute selection method was varied *within* data sets. One hundred different data sets were used for each condition, i.e., 100 Monte Carlo trials were carried out for the simulation. (Each trial involved assessing the classification performance on 100 cases, by dynamic path generation.) For each data set, the number of cases correctly classified under cross-validation was recorded for each selection criterion.

3.3. Results and discussion

The results of the experiment are summarized in table 1 and are displayed graphically in figure 1. It can be seen that, whereas the classification performance using H_T was hardly changed as the number of noise variables was increased, performance using the random selection measure declined markedly. As a first step in checking the statistical significance of these findings, a split-plot analysis of variance (ANOVA) was performed on the results. This is summarized in table 2. It can be seen that both main effects and the interaction were highly significant. This step was followed by a two-tailed matched-pairs t test between the results for the two different measures on the first condition (i.e., one noise variable). This showed that, although the results for the different measures were close (77.98% for H_T, as opposed to 77.00% for the random method), there was nevertheless a *significant* difference between them ($t = 2.19$; $df = 99$; $p < 0.05$). Since the differences for the other conditions were far larger, there seemed little point in testing these for statistical significance.

Table 1. Classification performance in experiment I.

	Attribute Selection Method	
Experimental Condition	Random Selection	H_T
1 signal and 1 noise	77.00 (5.11)	77.98 (5.13)
1 signal and 5 noise	68.43 (4.14)	77.59 (5.00)
1 signal and 10 noise	63.10 (5.21)	77.15 (5.71)
1 signal and 20 noise	57.03 (5.04)	76.82 (5.96)
1 signal and 40 noise	53.18 (4.51)	76.05 (6.86)
1 signal and 80 noise	52.66 (4.95)	75.59 (8.06)

Note: Results are expressed in terms of mean percentage of correct classifications, for each experimental condition and attribute selection method. Corresponding standard deviations are parenthesized.

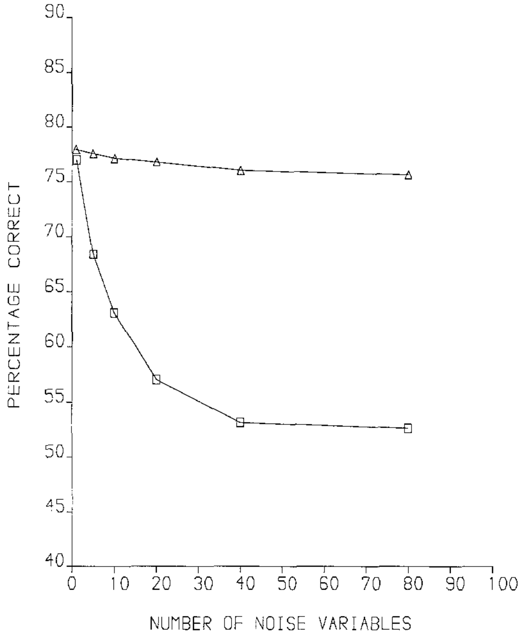

Figure 1. Cross-validated classification performance of the induced trees, expressed as percentage of correct classifications as a function of the number of noise variables. Points marked by triangles represent performance using H_T, and points marked by squares represent performance using the random selection procedure.

Table 2. ANOVA summary table for Experiment I.

Source	SS	df	MS	F	p
Condition	27170.3	5	5434.1	145.5	< 0.001
Dataset (condition)	22181.9	594	37.3		
Measure	67170.4	1	67170.4	2718.6	< 0.001
Condition × Measure	18966.3	5	3793.3	153.5	< 0.001
Measure × Dataset	14676.3	594	24.7		
Total	150165.2	1199			

Turning to the highly significant interaction between selection measure and number of noise variables found in the split-plot ANOVA, it was decided to perform two one-way analyses of variance (one for each selection measure) in order to locate the locus of the effect. Not surprisingly, the result for the random selection method was highly significant ($F = 390$; $df = 5{,}594$; $p < 0.001$). Performance declined from 77.00% to 52.66% over the conditions employed in the experiment. The corresponding results for H_T were quite different. Performance showed a very modest decline (from 77.98% to 75.59%), which did not reach statistical significance ($F = 2.15$; $df = 5{,}594$; $p > 0.05$).

These results confirm absolutely the expectations stated in the introduction to this experiment. If an effective attribute selection method is employed, then the number of available pure noise attributes has little effect on performance. Constrastingly, if a random selection method is used, then classification performance declines steeply as the number of noise variables is increased. From the results, it looks as if performance using the random selection method is tending towards an asymptote of 50%. Even when only a single pure noise variable was used, performance was significantly poorer with random attribute selection.

3.4. Subsidiary analysis on branching order

In order to determine more exactly the loci of these effects, some subsidiary analyses were carried out on the branching order in the classification paths.[1] First of all, the branching behavior was examined to determine whether or not the signal variable was branched on in the various conditions. The results were very clear. For selection by H_T, the signal variable was branched on at the root for every case in every data set, for every experimental condition. By contrast, the random attribute selection method produced a quite different picture. Table 3 shows that, as the number of noise variables increases, the mean number of cases in each data set for which the classification path branches on the signal variable decreases monotonically from 49.78 to 1.67. This result was found to be highly significant on a one-way ANOVA ($F = 6480$; $df = 5{,}594$; $p < 0.001$). Clearly, this factor is of great importance in determining the level of classification performance because, if the signal variable were not branched on in a particular classification path, then the expected classification performance could only be that which would be expected by chance.

Table 3. Summary statistics for the number of cases within each dataset that branch on the signal variable under random attribute selection.

Experimental Condition	Mean	St. dev.	Range
1 signal and 1 noise	49.78	3.06	42–55
1 signal and 5 noise	17.23	3.62	9–25
1 signal and 10 noise	9.71	2.06	5–15
1 signal and 20 noise	5.20	1.69	2–9
1 signal and 40 noise	2.31	0.90	1–4
1 signal and 80 noise	1.67	0.79	0–3

In order to determine whether there was another effect operating due to overfitting, a further analysis was performed on the subset of cases derived by extracting just those cases whose classification paths under random attribute selection branched on the signal variable. The classification performance on this subset of the data for both criteria is shown in table 4. It can be clearly seen that, for random attribute selection, there is a marked performance decrement as the number of noise variables, was increased. In fact, for the condition with one noise variable, classification performance was only slightly less than that for the orthodox measure. With 80 noise variables, classification performance was not much better than chance.

A split-plot ANOVA (similar to that used for the main part of the experiment) was applied to the performance data arising from this subset. Because one of the data sets for the condition with 80 noise variables did not branch on the signal variable for *any* of the cases, subset performance could be assessed for only 99 data sets in this condition. As a result, the design was slightly unbalanced. For this reason, the regression approach for

Table 4. Classification performance for signal branching subset in Experiment I.

	Attribute Selection Method	
Experimental Condition	Random Selection	H_T
1 signal and 1 noise	77.33 (6.69)	77.87 (6.84)
1 signal and 5 noise	68.98 (12.3)	78.00 (10.9)
1 signal and 10 noise	64.22 (16.7)	76.10 (15.3)
1 signal and 20 noise	51.31 (23.3)	79.86 (18.8)
1 signal and 40 noise	53.58 (36.6)	79.83 (29.4)
1 signal and 80 noise	57.24 (42.9)	76.94 (36.9)

Note: Results are expressed in terms of mean percentage of correct classifications for each experimental condition and attribute selection method. Corresponding standard deviations are parenthesized. See text for further explanation.

the attribution of variance components was used. This method attributes to each term only that portion of the variance that is *unique* to that particular source and is the approach generally favored by statisticians for dealing with unbalanced designs. The summary table is shown in table 5. All the effects were found to be highly significant. However, of particular interest is the fact that the interaction term was significant, providing statistical support for the idea that increasing the number of noise variables had little effect on classification performance under H_T but produced a performance decrement on classification by random attribute selection *even for those cases for which the classification path branched on the signal variable.* To be absolutely clear about the source of this effect, one-way ANOVAs were performed on each measure separately. For random attribute selection, the performance decrement was highly significant ($F = 14.2$; $df = 5{,}593$; $p < 0.001$). By contrast, classification of the same cases by H_T showed no significant dependence on the number of noise variables ($F = 0.46$; $df = 5{,}593$; $p > 0.5$).

Thus, there are clearly two quite separate effects in operation for classification by random attribute selection, each of which contributes to the increasing under-performance as the number of noise variables is increased. First, the probability of the signal variable being branched on decreases. Second, even when the signal variable is branched on, increasing the number of noise variables available tends to increase the number of noise variables branched on above the signal variable in the classification path. This degrades classification performance through overfitting.

4. Experiment II

4.1. Introduction

The results from Experiment I suggested that the under-performance of random attribute selection was due to the tree construction algorithm branching on noise variables and producing a suboptimal tree. It was further suggested that the under-performance was due to two distinct factors—first, a reduced likelihood of informative attributes being included in the final tree, and second, a degree of overfitting caused by noise variables being locked into the tree, near the root.

Table 5. ANOVA summary table for subsidiary analysis of performance on signal branching subset of the data. See text for further explanation.

Source	SS	df	MS	F	p
Condition	27141.3	5	4348.3	6.12	< 0.001
Dataset (condition)	421505	593	710.8		
Measure	76563.7	1	76563.7	159	< 0.001
Condition × Measure	29038.2	5	5807.6	12.1	< 0.001
Measure × Dataset	285586	593	481.6		
Total	834377	1197			

Now, if these explanations are correct, then it follows that, if all the variables in the attribute set are equally important in discriminating between classes, then it should not matter which attributes are selected for branching. Neither should the order of branching matter.

Hence, it was hypothesized that, in a probabilistic classification task, if all the attributes contain equal levels of information concerning class membership, then there should be no difference in cross-validated classification performance between H_T and a random method of attribute selection. Experiment II was designed to test this hypothesis.

4.2. Method

Just as in Experiment I, Experiment II involved a Monte Carlo simulation using the same measures of attribute selection—H_T and the random method. Only one experimental condition was employed, using 100 different data sets of 100 cases each, generated in a similar manner to that described for Experiment I.

The only difference was that 11 attributes were used, which were all generated to contain equal levels of information about class membership. Each attribute was derived by simply adding the binary class variable (0 or 1) to a random sample drawn from the standard normal distribution.

Binary splitting and cross-validation by dynamic path generation were employed in the same way as described for Experiment I. As before, the percentage of correct classifications, for each selection method, was recorded for each data set.

4.3. Results and discussion

The mean levels of accuracy for the two measures were 76.61% (random) and 77.84% (H_T). The corresponding standard deviations were 4.08 and 5.27, respectively. A two-tailed matched-pairs t test was performed on the results, which indicated a significant difference in favor of selection using H_T ($t = 2.15$; $df = 99$; $p < 0.05$).

This result was not quite as anticipated, because it was expected that the advantage of using an orthodox measure of attribute selection over a random method would disappear when all the available attributes contained the same level of information about class membership. Although the performance difference found between the measures was small (i.e., of the order of one percentage point), it was nevertheless statistically significant.

A possible explanation for this finding, is that, even though the attributes were arranged to be of equal importance *at the root* of the classification tree, this does not guarantee that they will be of equal importance at every *intermediate node* in the classification path. Random variation will tend to produce some degree of inequality in importance between the attributes at the intermediate nodes in the classification paths, for some of the cases. This means that, if an orthodox method of attribute selection is being used, then the algorithm will operate on these small differences in importance to produce marginally superior classification performance, compared with that given by random attribute selection.

The argument just stated carries the implication that, if two attributes are available for selection at a node then, if they are not of equal importance in discriminating between classes, choosing the less important attribute could be suboptimal (i.e., lead to poorer cross-validated classification performance).

If this explanation is correct, then it should certainly be possible to observe the same effect if the attributes are not of equal importance *to begin with*, i.e., if they all contain different amounts of information concerning class membership.

5. Experiment III

5.1. Introduction

The previous experiment showed that using an orthodox measure of attribute selection will yield only a small improvement in classification performance over that obtainable by using a random selection method in the same situation, when the attributes are equally informative. An argument presented in the discussion for that experiment suggested that this small difference in performance may have been due to inequities between attributes in the information concerning class membership appearing at the intermediate nodes in the tree, as a result of the binary splitting and branching processes. If this argument is correct, then it should also be possible to demonstrate the predictive superiority of an orthodox measure for attribute selection (over that expected with a random method) if the attributes each possess information about class membership to begin with, but to different degrees. It could also be argued that such a situation is more representative of the sort of situation likely to be found in real classification problems in noisy domains. Of course, it is to be expected that the difference in cross-validated classification performance between an orthodox and a random measure of attribute selection would be greater in this situation than in the one simulated in the previous experiment. Experiment III was designed to test this hypothesis.

5.2. Method

The experimental design was very similar to that employed in the previous experiment. Just as before, a single experimental condition was employed, using 100 different data sets of 100 cases each.

Ten attributes were used. They were generated so that they each contained different amounts of information about class membership. Each attribute was derived by adding k times the binary class variable to a standard normal random variable. Ten different values of k were used—one for each attribute. The values of k ranged from 0.2 to 2 in equal steps.

Binary splitting and cross-validation by dynamic path generation were used, just as in the two previous experiments. As before, the percentages of correct classification for both the H_T and random selection methods were recorded for each data set.

5.3. Results and discussion

The mean levels of accuracy for the two measures were 82.36% (random) and 86.62% (H_T). The corresponding standard deviations were 4.04 and 4.63, respectively. A two-tailed matched-pairs t test was performed on the results, indicating a highly significant difference, showing H_T as providing superior performance ($t = 7.62$; $df = 99$; $p < 0.001$).

This result was as expected. Thus, it seems reasonable to say that, unless it can be guaranteed that all the attributes are of exactly equal importance at a node (as regards the information they convey about class membership), then it is a better strategy to employ an orthodox measure of attribute selection, rather than a random one.

5.4. Subsidiary analysis on branching order

In order to check that the more important attributes were indeed branched on closer to the root of the tree when H_T was used for attribute selection than when random attribute selection was employed, a subsidiary analysis was performed on data derived from the branching order.[2] As explained earlier, the technique of dynamic path generation was employed in these experiments, in combination with cross-validation. Thus, for the classification of each case, only the path actually needed for classifying the case concerned was generated. For each case in each data set for each of the two attribute selection methods, the attributes branched on in the classification path were recorded, together with their respective positions in the classification path.

The attributes were labeled with integers from 1 to 10, denoting their level of importance. Thus, the attribute derived by using a value of k of 0.2 was given an importance level of 1, while the attribute produced with $k = 2$ had an importance level of 10. Positions in the classification path were also represented by integers, denoting the number of steps below the root of the node concerned. Thus, the root itself was given a position number of 0, a node one step below the root had a position number of 1, and so on. So, for each case, the branching information was recorded as two sequences of number pairs, one sequence for each method of attribute selection. Each number pair consisted of an importance level and a position number. Each sequence of number pairs traced the final classification path down from the root to the terminal node, after pruning.

Simple examination of the resulting information showed a marked difference in mean path length between the two attribute selection methods, with path length for the random selection method being longer on average. Path length ranged from 1 to 7 steps, with a mean of 2.640, for selection by H_T. On the other hand, the random selection method gave a range from 0 to 14, with a mean of 3.867. (A path length of zero simply means one that has been pruned back to the root.) The difference between the two sets of path lengths was tested with a two-tailed matched-pairs t test. The result was highly significant ($t = 67.2$; $df = 9999$; $p < 0.001$).

In order to test for differences in the branching *order*, the technique used was based on examining differences in importance level as a function of position, between the two selection methods. Only position numbers from 1 to 6 were used, because of the shorter path length given by selection with H_T. This resulted in a two-way factorial experimental design (Keppel, 1973, pp. 195–196). Thus, one factor was the attribute selection method and the other was the position number. The independent variable was the importance number. Replication was provided by the multiple cases and data sets, giving 10,000 classification paths for each condition. However, it should be made clear that the design was necessarily unbalanced, because not all classification paths were of the same length.

The mean importance levels for the various conditions are shown in table 6. This information is also displayed graphically in figure 2. It can be seen quite clearly that, for attribute selection by H_T, mean importance level declines steeply as a function of increasing position number. Thus, under this condition, there is a strong tendency for the more informative attributes to be branched on closer to the root. For random attribute selection, on the other hand, the picture is quite different. Here, position number has little influence on mean importance level, with the latter staying close to the expected importance level of 5.5.

The difference in behavior between the two selection measures was tested with a two-way factorial ANOVA (Keppel, 1973, pp. 195–196). The unbalanced nature of the design was handled by using the regression approach for the attribution of the variance components, as used earlier for the subsidiary analysis in Experiment I. The summary table is given in table 7. Here, the effect of interest is the interaction term, which was found to be highly significant, showing that there was a definite tendency for the graphs to be different for the different selection measures.

A minor point of interest is that, for random attribute selection, mean importance level is slightly greater than the value of 5.5, which would be expected under a simple theory. Furthermore, the results seem to show a slight increase with position number. This would appear to be due to the fact that the method of pruning employed tends to ensure that the final branch in the *pruned* classification path is based on an attribute of high importance, even when random attribute selection has been used to construct the path originally.

Table 6. Results for branching order analysis in Experiment III.

	Attribute Selection Method	
Position	Random Selection	H_T
1	5.552	9.424
	(2.954)	(0.794)
2	5.652	7.650
	(2.961)	(1.872)
3	5.747	5.964
	(2.915)	(2.589)
4	5.785	4.455
	(2.870)	(2.520)
5	5.828	3.336
	(2.818)	(2.605)
6	5.775	2.231
	(2.809)	(1.966)

Note: Results are expressed in terms of mean importance level for the first six positions in the classification path for orthodox and random attribute selection. Corresponding standard deviations are parenthesized.

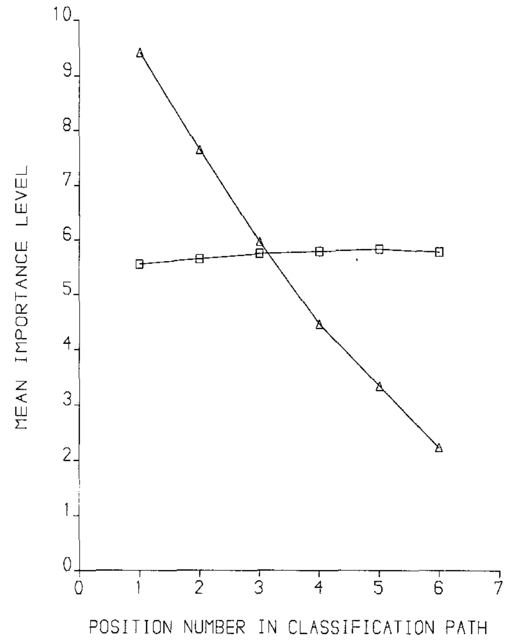

Figure 2. Mean importance level for the first six positions in the classifiction path for orthodox and random attribute selection in Experiment III. Points marked by triangles represent performance using H_T, and points marked by squares represent performance using the random selection procedure.

Table 7. ANOVA summary table for analysis of branching order in experiment III.

Source	SS	df	MS	F	p
Position	35681	5	7442	1174.82	< 0.001
Condition	52482	1	51	8.03	0.005
Position × Condition	45439	5	9088	1434.54	< 0.001
Error	404320	63824	6		
Total	537922	63835			

The results of this subsidiary analysis are entirely as expected. Thus, there is a very strong tendency for the more important attributes to be branched on earlier when an orthodox measure of attribute selection is employed, but not when attributes are selected randomly.

6. Conclusions

From an intuitive point of view, the conclusions drawn from each of the three experiments are hardly surprising. It is only to be expected that an orthodox measure of attribute selection should outperform a random one in most situations. It is instructive to consider why this should be so.

The reason appears to lie in the nature of the tree-building process itself, including the pruning phase. Clearly, it matters a great deal which attributes are selected for branching on, near the root. Furthermore, the closer to the root the branching is taking place, the more it matters. This is because of the problems of overfitting and the practice of pruning used to counteract it. Pruning removes excessive branching (i.e., branching on noise, which causes overfitting). However, the important point is this: *The pruning technique used here tends to remove the lower nodes on the classification path.* To put things another way, pruning starts at the terminal node and proceeds towards the root, until a significant class-attribute association is encountered. At this point, pruning of the current branch is terminated. *This means that if there has been any branching on noise above this point, then it cannot be undone by pruning.* Consequently, any "noisy branching" near the root is locked in and cannot be removed by the pruning process. This, in turn, can produce suboptimal performance because of overfitting.

This line of argument suggests that, with most data sets, random attribute selection would tend to produce decision trees with poorer classification performance than those constructed using orthodox attribute selection measures. This is entirely in accordance with the findings of Buntine and Niblett (1992) on a range of empirical data sets.

The remaining puzzle is how Mingers (1989a) managed to come to different conclusions in his work, particularly since 3 of the 4 data sets that he employed were also used by Buntine and Niblett in their investigations.

Of course, Mingers' results for the unpruned trees do not present a problem because, in such a situation, all the available noise will be included in the tree and, for predictive purposes, it does not matter where in a tree the noisy branching occurs. Mingers' results for the pruned trees are more difficult to explain.

Now, Buntine and Niblett discuss this matter at length. Perhaps the most important point that they make is that Mingers used the test set in the pruning phase to select the best pruned tree. They explain why this is methodologically unsound, and it is unnecessary to repeat their explanation here.

There is also something strange about the statistical analysis that Mingers applied to his data. For 3 out of the 4 data sets that he used, the error rate on the pruned tree for random attribute selection was, in fact, higher than for *any* of the orthodox measures. For example, with the breast cancer data, the conventional measures yielded error rates of between 21.5% and 25.4%, whereas the random measure produced an error rate of 27.5%. The failure of these results to achieve statistical significance might well have been due to the use of an inappropriate model for the ANOVA. Unfortunately, Mingers did not present ANOVA summary tables for his analyses, nor does he quote degrees of freedom when he reports *F* ratios. It is thus somewhat difficult to be sure exactly what he *did* do. However, Buntine and Niblett state that Mingers performed his analysis of variance "*on the matrix of error averages*". Mingers' paper does not actually say this, but the critical *F* values quoted are certainly consistent with this approach having been taken. Such an approach is really not the best way to analyze the data. A more sensitive analysis should have been performed by employing a split-plot model, using the error rates for *each data set* (not the averages) as the basic scores, just as was done for Experiment I reported in this article. This reflects the fact that measures were varied *within* test sets, rather than between them (i.e., *each* test set was tested using *each* of the measures, instead of an independent test set being used each time).

In conclusion, it should be clear that it is of great importance that the method used for attribute selection should branch preferentially on those attributes that convey information about class membership. An extension of the same argument requires that, at every stage of the branching process, an optimal attribute selection measure should select the most informative attribute on which to branch. Any other approach would, in the general case, yield suboptimal results.

Notes

1. The idea of performing an additional analysis on branching order was suggested by an anonymous referee.
2. The idea of performing an additional analysis on branching order was suggested by an anonymous referee.

References

Breiman, L., Friedman, J.H., Olshen, RA., & Stone, C.J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth.

Buntine, W., & Niblett, T. (1992). A further comparison of splitting rules for decision-tree induction. *Machine Learning, 8,* 75–86.

Keppel, G. (1973). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice-Hall.

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

Liu, W.Z., & White, A.P. (1991). A review of inductive learning. In I.M. Graham & R.W. Milne (Eds.), *Research and development in expert systems VIII.* Cambridge: Cambridge University Press.

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

Mingers, J. (1989b). An empirical comparison of pruning methods for decision-tree induction. *Machine Learning, 4,* 227-243.

Niblett, T., & Bratko, I. (1987). Learning decision rules in noisy domains. In M.A. Bramer (Ed.), *Research and development in expert systems III.* Cambridge: Cambridge University Press.

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.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

White, A.P. (1985). PREDICTOR: an alternative approach to uncertain inference in expert systems. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 328–330). Los Altos: Morgan Kaufmann.

White, A.P. (1987). Probabilistic induction by dynamic path generation in virtual trees. In M.A. Bramer (Ed.), *Research and development in expert systems III.* Cambridge: Cambridge University Press.

White, A.P., & Liu, W.Z. (1990). Probabilistic induction by dynamic path generation for continuous variables. In T.R. Addis & R.M. Muir (Eds.), *Research and development in expert systems VII.* Cambridge: Cambridge University Press.

Received May 21, 1992
Accepted November 16, 1992
Final Manuscript April 8, 1993