Machine Learning 3: 319–342, 1989

An Empirical Comparison of Selection Measures for Decision-Tree Induction

JOHN MINGERS (BSRCD@CU.WARWICK.AC.UK)
School of Industrial and Business Studies, University of Warwick, Coventry CV4 7AL, U.K.

(Received: May 6, 1987)
(Revised: October 19, 1988)

Keywords: Decision trees, knowledge acquisition, induction, noisy data

Abstract. One approach to induction is to develop a decision tree from a set of examples. When used with noisy rather than deterministic data, the method involves three main stages – creating a complete tree able to classify all the examples, pruning this tree to give statistical reliability, and processing the pruned tree to improve understandability. This paper is concerned with the first stage – tree creation – which relies on a measure for "goodness of split," that is, how well the attributes discriminate between classes. Some problems encountered at this stage are missing data and multi-valued attributes. The paper considers a number of different measures and experimentally examines their behavior in four domains. The results show that the choice of measure affects the size of a tree but not its accuracy, which remains the same even when attributes are selected randomly.

1. Introduction

There are a number of approaches to inductive learning (Michalski, Carbonell, & Mitchell, 1983, 1986; Bratko & Lavrac, 1987), one of which involves the construction of decision trees. This method was developed initially by Hunt, Marin, and Stone (1966) and later modified by Quinlan (1979, 1983), who applied his ID3 algorithm to deterministic domains such as chess end games. Breiman, Friedman, Olshen, and Stone (1984) independently developed a similar approach to classification. Recent research has focused on induction in domains that are uncertain and noisy rather than deterministic (Quinlan, 1986a). Broadly speaking, there are three phases to induction with non-deterministic data – the creation of an initial tree from examples; pruning this tree to remove branches with little statistical validity; and processing the pruned tree to improve its understandability.

This paper provides an empirical comparison of a number of methods and strategies for the creation phase. A further paper (Mingers, 1988) addresses the pruning phase. Section 2 outlines Quinlan's basic algorithm and the various measures and approaches covered in this paper. Section 3 describes the data and experimental procedure, and Section 4 summarizes the results of the study.

Table 1. Values of two attributes from the breast cancer domain.

RADIATION	MENOPAUSE	CLASS
NO	< 60	RECUR
NO	≥ 60	RECUR
NO	< 60	RECUR
NO	NOT	RECUR
YES	≥ 60	NOT RECUR
YES	< 60	NOT RECUR
YES	≥ 60	NOT RECUR
NO	NOT	NOT RECUR
NO	< 60	NOT RECUR
NO	< 60	RECUR

2. Strategies for decision-tree induction

The ID3 algorithm has been extensively detailed elsewhere (Quinlan 1979, 1983, 1986b) and so only a brief outline will be given here. It begins with a set of examples already divided into classes. Each example is described in terms of a set of attributes, which can be numeric or symbolic. The overall approach is to choose the attribute that best divides the examples into their classes and then partition the data according to the values of that attribute. This process is recursively applied to each partitioned subset, with the procedure terminating when all examples in the current subset have the same class. The result of this process is represented as a tree in which each node specifies an attribute and each branch emanating from a node specifies the possible values of that attribute. Terminal nodes (leaves) of the tree correspond to sets of examples with the same class or to cases in which no more attributes are available.

2.1 The use of contingency tables

At each node in the development of a decision tree there will be a set of instances and a number of attributes available to classify them. One selects the best attribute by seeing how well each one separates the data into the various classes. Breiman et al. (1984) call this a "goodness of split" measure.[1] To calculate an attribute's goodness of split at a particular node in the tree, the available examples can be set out in a contingency table. This is illustrated in the following example, which we use later to demonstrate the various measures.

Consider a domain in which one must predict the recurrence of breast cancer. One attribute in this domain involves whether radiation treatment was given (yes or no), and another is the age at which the patient's menopause occurred (< 60, ≥ 60, not occurred). Table 1 shows some instances from this domain, and Tables 2 and 3 show the resulting contingency tables for the two attributes.

[1] Most work, including that described here, assumes that the distribution of classes in the sample reflects that of the population and that the costs of misclassifying the data are the same for all classes. Breiman et al. (1984) show how both differing prior probabilities and varying misclassification costs can be incorporated into their algorithm.

Table 2. Contingency table for the radiation attribute.

		Class		
		RECUR	NOT RECUR	
Radiation	YES	0	3	3
	NO	5	2	7
		5	5	10

Note that the class totals will be the same for all the attributes being compared at any node, but the row totals and the number of rows may differ. The goodness of split measures calculate the extent to which the attribute values split the data into the separate classes. A perfect attribute would have each attribute value associated with only one class. Each row would therefore have only one non-zero entry. At the opposite extreme, the values for a useless attribute would not be associated with the classes at all, and all the entries in a row would be the same. Looking at the two examples, it appears that radiation is better than menopause.

In addition to the choice of basic measure, three other problems have received some attention in the literature on decision trees – multi-valued attributes, missing data, and small splits. The remainder of this section reviews various strategies for dealing with each of these issues.

2.2 Measures of goodness of split

Since Quinlan's original work, there have been a number of alternative suggestions for measures to be used in selecting attributes. These will be developed using the notation for a general contingency table shown in Table 4 and calculated for the examples in Tables 2 and 3.

2.2.1 Quinlan's information measure (IM)

Quinlan (1979, 1983) proposed an evaluation function based on a classic formula from information theory that measures the theoretical information content of a code $-\sum p_i \log(p_i)$ – where p_i is the probability of the i-th message. The value of this measure depends on the likelihood of the various possible messages. If they are equally likely (and so the p_i are equal), there is the greatest amount of uncertainty and the information gained will be greatest. The less equal the probabilities, the less information there is to be gained. The value of the function also depends on the number of possible messages. A good analogy is with a horse race – the more runners and the more evenly they are matched, the greater the value of knowing the winner.

Table 3. Contingency table for the menopause attribute.

		CLASS		
		RECUR	NOT RECUR	
AGE OF MENOPAUSE	< 60	3	2	5
	≥ 60	1	2	3
	NOT	1	1	2
		5	5	10

Using Quinlan's (1983) notation, the information needed to classify items, given only the class totals as a whole, is

$$\begin{aligned} M(C) &= -\frac{x_{.1}}{N}\log\left(\frac{x_{.1}}{N}\right) - \frac{x_{.2}}{N}\log\left(\frac{x_{.2}}{N}\right) - \cdots \\ &= -\frac{1}{N}\left(\sum x_{.j}\log x_{.j} - N\log N\right). \end{aligned}$$

The information content of a row, say A_1, is

$$M(A_1) = -\frac{x_{11}}{x_{1.}}\log\left(\frac{x_{11}}{x_{1.}}\right) - \frac{x_{12}}{x_{1.}}\log\left(\frac{x_{12}}{x_{1.}}\right) - \cdots,$$

and the other rows are treated similarly.

The information needed to classify items, given knowledge of the attribute value, is then the average of these expressions, weighted by the frequency of occurrence of each value (the row total):

$$\begin{aligned} B(C|A) &= \frac{x_{1.}}{N}M(A_1) + \frac{x_{2.}}{N}M(A_2) + \cdots \\ &= -\frac{1}{N}\left(\sum\sum x_{ij}\log x_{ij} - \sum x_{i.}\log x_{i.}\right). \end{aligned}$$

The information measure IM is then defined as the gain in information brought about by knowledge of the attribute:

$$\begin{aligned} IM &= M(C) - B(C|A) \\ &= \frac{1}{N}\left(\sum\sum x_{ij}\log x_{ij} - \sum x_{i.}\log x_{i.} - \sum x_{.j}\log x_{.j} + N\log N\right). \end{aligned}$$

Table 4. General notation for contingency tables.

		CLASS				
		C_1	C_2	...	C_c	TOTAL
VALUE OF ATTRIBUTE	A_1	x_{11}	x_{12}		x_{1c}	$x_{1.}$
	A_2	x_{21}	x_{22}		x_{2c}	$x_{2.}$
	$\vdots$					$\vdots$
	A_r	x_{r1}	x_{r2}		x_{rc}	$x_{r.}$
		$x_{.1}$	$x_{.2}$	...	$x_{.c}$	N

For the radiation example in Table 2, we have

$$\begin{aligned}
M(C) &= -\frac{5}{10}\log\frac{5}{10} - \frac{5}{10}\log\frac{5}{10} &&= 0.69315\\
M(A_1) &= -0 - \frac{3}{3}\log\frac{3}{3} \quad = \quad 0\\
M(A_2) &= -\frac{5}{7}\log\frac{5}{7} - \frac{2}{7}\log\frac{2}{7} &&= 0.59827\\
B(C|A) &= \frac{3}{10}\times 0 + \frac{7}{10}\times 0.59827 &&= 0.41879\\
IM &= 0.69315 - 0.41879 \quad = \quad 0.27436.
\end{aligned}$$

Similarly, for the menopause attribute shown in Table 3, we have

$$\begin{aligned}
M(C) &= 0.69315\\
M(A_1) &= -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} &&= 0.67301\\
M(A_2) &= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} &&= 0.63651\\
M(A_3) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} &&= 0.69315\\
B(C|A) &= \frac{5}{10}\times 0.67301 + \frac{3}{10}\times 0.63651 + \frac{2}{10}\times 0.69315 &&= 0.66575\\
IM &= 0.69315 - 0.66575 &&= 0.02740.
\end{aligned}$$

This result agrees with the intuition presented earlier. The radiation attribute provides more information than the menopause attribute and so should be selected for extending the decision tree.

2.2.2 The chi-square contingency table statistic (χ^2)

Hart (1984) and Mingers (1986a, 1987a) have employed another measure to select among attributes – the chi-square (χ^2) statistic. This is the traditional statistic for measuring the association between two variables in a contingency table. It compares the observed frequencies with the frequencies that one would expect if there were no association between the variables. The resulting statistic is distributed approximately as the chi-square distribution, with larger values indicating greater association. In these experiments Yates' correction is used for 2 × 2 tables (Upton, 1986). The basic equation for this function is

$$\chi^2 = \sum\sum \frac{(x_{ij} - E_{ij})^2}{E_{ij}},$$

where $E_{ij} = x_{i.}x_{.j}/N$, i.e., the expected value for each cell in the contingency table. Thus for the radiation attribute we get

$$\chi^2 \quad = \quad \frac{(0-1.5)^2}{1.5} + \frac{(3-1.5)^2}{1.5} + \frac{(5-3.5)^2}{3.5} + \frac{(2-3.5)^2}{3.5} \quad = \quad 4.29,$$

whereas for the menopause attribute we get

$$\chi^2 \quad = \quad \frac{(3-2.5)^2}{2.5} + \cdots + \frac{(1-1)^2}{1} \quad = \quad 0.533.$$

Again, the results favor the radiation attribute.

2.2.3 The G statistic (G)

Mingers (1987a) has also used the G statistic (Sokal & Rohlf, 1981) in the induction of decision trees. This is another statistic designed for use with contingency tables that is based on information theory. It is equivalent to Kullback's (1967, p. 158) information measure $H(A,C)$ and Wilk's likelihood ratio (Upton, 1982). This metric approximates the chi-square distribution and is also closely related to Quinlan's measure. In fact, Mingers (1987a) has shown that

$$G = H(A,C) = 2N \times IM,$$

where N is the number of examples. When comparing attributes, N will be constant and so G and IM will always give the same ordering. In the experiments described later, G has been used instead of IM, since it follows the chi-square distribution.

Returning to the breast cancer domain, we have

$$G = 2 \times 10 \times IM = 2 \times 10 \times 0.27436 = 5.49$$

for the radiation attribute and

$$G = 2 \times 10 \times 0.02740 = 0.548$$

for the menopause attribute, again indicating the former should be preferred.

Table 5. Chi-square probability of finding a value greater than or equal to the statistics from Tables 2 and 3.

ATTRIBUTE	X^2	G
RADIATION	0.04	0.02
MENOPAUSE	0.47	0.46

2.2.4 Using probabilities rather than the statistic (PROB)

Instead of using the value of the χ^2 or G statistics as calculated, one can compute the probability of such a value occurring from the χ^2 distribution on the assumption that there is actually *no* association between the attribute and the classes. The more extreme the calculated value, the less likely it is to have occurred given the assumption. Therefore, the smaller the probability the greater the likely degree of association between attribute and class. Table 5 shows these probabilities for the data from Tables 2 and 3.

The probabilities for the G statistic show there is a 46% chance of a value as high as 0.548 (for the menopause attribute) but only a 2% chance of a value as high as 5.49 (for the radiation attribute). Therefore, radiation is likely to be associated with the class, but menopause is not. Similar probabilities emerge for the χ^2 statistic. This approach has two potential advantages. First, it provides actual levels of significance that can be used in deciding whether to include an attribute at all. Second, it makes allowance for the different number of attribute values via the degrees of freedom. The latter point is discussed more fully in the section on multi-valued attributes.

2.2.5 The GINI index of diversity (GINI)

Breiman et al. (1984) have employed another measure that is similar to IM but based on a different function. Their GINI function measures the 'impurity' of an attribute with respect to the classes. Given the probabilities for each class (p_i), the general GINI function, or measure of impurity, is

$$\sum\sum_{j\neq i} p_i p_j \;=\; \left(\sum p_i\right)^2 - \sum p_i^2 \;=\; 1 - \sum p_i^2.$$

In our case, we estimate the class probabilities with the actual relative frequencies – $x_{.i}/N$. Thus the impurity of the class totals is

$$i(t) = 1 - \left(\frac{x_{.1}}{N}\right)^2 - \left(\frac{x_{.2}}{N}\right)^2 - \cdots,$$

whereas the impurity of row A_1 is

$$i(A_1) = 1 - \left(\frac{x_{11}}{x_{1.}}\right)^2 - \left(\frac{x_{12}}{x_{1.}}\right)^2 - \cdots$$

and similarly for other rows.

The increase in impurity is the class impurity minus the weighted average of the row impurities:

$$
\begin{aligned}
i &= i(t) - \frac{x_{1.}}{N} i(A_1) - \frac{x_{2.}}{N} i(A_2) - \cdots \\
&= \left(1 - \sum \left(\frac{x_{.j}}{N}\right)^2\right) - \frac{x_{1.}}{N}\left(1 - \sum \left(\frac{x_{1j}}{x_{1.}}\right)^2\right) - \cdots \\
&= \frac{1}{N}\left(\sum\sum \frac{x_{ij}^2}{x_{i.}} - \sum \frac{x_{.j}^2}{N}\right).
\end{aligned}
$$

For the radiation attribute from Table 2, this comes to

$$
i = \frac{1}{10}\left(\frac{0^2}{3} + \frac{3^2}{3} + \frac{5^2}{7} + \frac{2^2}{7}\right) - \left(\frac{5^2}{10} + \frac{5^2}{10}\right) = 0.21429,
$$

whereas for menopause it evaluates to

$$
i = \frac{1}{10}\left(\frac{3^2}{5} + \frac{2^2}{5} + \cdots\right) - \left(\frac{5^2}{10} + \frac{5^2}{10}\right) = 0.026667.
$$

This gives the same ordering as the other measures, preferring radiation to menopause.

2.2.6 Gain-ratio measure (GR)

Quinlan (1986b) used a variant of his IM measure that incorporates the notion that an attribute itself will have some information value. This amount will depend upon the distribution of examples among its possible values (i.e., the row totals). The less evenly spread its values, the less information in the attribute. An efficient splitting measure should convert as much as possible of the attribute's information value into the classification procedure. This can be measured by calculating the ratio of the gain in information from using the attribute (i.e., IM) to the information value of the attribute itself. Thus Quinlan proposes a gain-ratio measure

$$
GR(A) = \frac{IM(A)}{IV(A)},
$$

where $IV(A)$ is the information value of attribute A, which is defined as

$$
IV(A) = -\sum \frac{x_{i.}}{N} \log\left(\frac{x_{i.}}{N}\right).
$$

The function has a high score if the examples are spread evenly between the attribute values and a low one if they are not. Thus GR favors those attributes with an unequal distribution of examples. IV is also proportional to the actual number of possible attribute values, so that GR favors attributes with a small number of values. One problem is that GR might choose attributes with very low IV scores, rather than those with high IM scores. To avoid this, Quinlan calculates the average IM for all the attributes in contention at a node and then selects only from among those with above average IM scores.

Returning to the domain of breast cancer, for the radiation attribute we get

$$IV = -\frac{3}{10}\log\frac{3}{10} - \frac{7}{10}\log\frac{7}{10} = 0.61086$$

and

$$GR = \frac{0.27436}{0.61086} = 0.44913,$$

whereas the results for menopause are

$$IV = -\frac{5}{10}\log\frac{5}{10} - \frac{3}{10}\log\frac{3}{10} - \frac{2}{10}\log\frac{2}{10} = 1.02965$$

and

$$GR = \frac{0.02740}{1.02965} = 0.02662.$$

Note that, in comparison with the original IM measures, radiation appears much more attractive in relation to menopause, although the overall ordering is the same.

2.2.7 Marshall correction (MARSH)

This is a correction factor that can be applied to any of the measures. The aim of it is the opposite of the gain ratio, that is to favor attributes which split the examples evenly and to avoid those which produce small splits. Marshall (1986) has suggested multiplying the calculated measure by the product of the row totals, $x_{i.}$, as this will be a maximum when the row totals are equal.

This simple approach has two problems. First, it makes the measure depend on N, the number of examples, and this will differ from node to node in the tree. This does not affect the choice of attribute at a node, but it does preclude between-node comparisons. Second, it depends on the number of values of the attribute. To avoid the first problem, one can use the ratios of $x_{i.}$ to N, giving

$$\frac{x_{1.}}{N} \times \frac{x_{2.}}{N} \times \cdots$$

If there are k attribute values, this expression has a maximum value of

$$\frac{N}{Nk} \times \frac{N}{Nk} \times \cdots = \frac{1}{k^k} \quad \text{when} \quad \frac{x_{1.}}{N} = \frac{x_{2.}}{N} = \cdots = \frac{x_{k.}}{N} = \frac{N}{k}.$$

The correction is therefore between zero and $1/k^k$. To make the factor less dependent on k, one can instead use the expression

$$\frac{x_{1.}}{N} \times \frac{x_{2.}}{N} \times \cdots \times k^k,$$

which has a range between zero and one. Applying this correction to the IM measure on the breast cancer data, the radiation attribute gives

$$\text{Marsh} = 0.27436 \times \frac{3}{10} \times \frac{7}{10} \times 2^2 = 0.23046,$$

Table 6. Summary of goodness of split calculations for two attributes from the breast cancer domain.

MEASURE	RADIATION	MENOPAUSE	RATIO
IM	0.27436	0.02740	10.01
χ^2	4.29	0.533	8.05
G	5.49	0.548	10.01
GINI	0.21429	0.02667	8.03
GAIN RATIO	0.44913	0.02661	16.88
MARSH	0.23046	0.02219	10.38

whereas menopause becomes

$$\text{Marsh} \quad = \quad 0.2740 \times \frac{5}{10} \times \frac{3}{10} \times \frac{2}{10} \times 3^3 \quad = \quad 0.02219.$$

Again, the ordering of attributes remains the same as for the other measures.

2.2.8 Summary

Table 6 summarizes the results obtained with each of the measures except the probability method. This table shows the extent of the differences in terms of the ratio between the measures for the two attributes. The *GR* function strongly favors radiation, since it has only two values and the examples are unevenly spread between them. The measures that are not based on information theory give radiation less weight. This may be because the zero in the first row of radiation has a greater influence in the log calculations. The Marshall correction makes little difference on these data, since the attributes are split roughly equally.

2.3 Strategies for multi-valued attributes

Attributes differ in terms of the number of possible values they have and the nature of those values:

- *Interval* attributes take on integer or real values from a measured scale such as money, age, weight, counts, etc.
- *Ordinal* attributes take on values that can be ranked or ordered, but that do not have constant intervals; e.g., grade on an examination (A, B, C, or FAIL).
- *Nominal* attributes take on values with no inherent ordering, such as type of car, sex, yes/no, or truth value. The latter are often referred to as *logical* attributes.
- *Structured* attributes (Bundy, Silver, & Plummer, 1985) may mix the above types; these are difficult to handle.[2]

[2] For example, if the attribute is an exam grade, how should someone who has not taken the exam be classified? Should it be counted as a fail or as a separate value that cannot be

Hart (1984), Kononenko, Bratko, and Roskar (1984), and Quinlan (1985) have all noted a problem with multi-valued attributes (those having more than two possible values). Such attributes tend to discriminate better among classes simply because they have more possible values. As a result, the basic algorithm is unduly biased towards incorporating them in the decision tree.

To illustrate this with an extreme case, consider age as an attribute, with each year being a distinct value. If each person in the sample happened to have a different age, then the attribute would discriminate perfectly between the classes. However, it would have little use for making predictions on other data sets. Researchers have suggested a number of strategies for dealing with multi-valued attributes. These are reviewed below.

2.3.1 Using degrees of freedom

When using the chi-square distribution to find the probability of a particular χ^2 or G value, the degrees of freedom (v) must be considered. For a contingency table $v = (r-1) \times (c-1)$, where r is the number of rows and c is the number of columns. Here the columns are the classes and the rows are the attribute values. Therefore, this approach, originally proposed by Hart (1984), should make allowance for differing numbers of attribute values. Mingers (1987a) has examined two different methods.

The first method involves normalizing the value by calculating its distance from the mean in terms of standard deviations. For the chi-square distribution, the mean $\mu = v$ and the standard deviation $\sigma = 2v$. Thus the normalized value is $(G-v)/2v$. Ideally, this value should be the same for calculated values having identical probabilities but differing degrees of freedom. In practice they differ slightly because the shape of the chi-square curve varies.

The second, and theoretically better, method of using the degrees of freedom is to calculate the actual probability of a particular calculated value of G or χ^2. The attribute with the smallest probability should then be selected since it has the greatest degree of discrimination. Although this approach does increase run time significantly – by about 20% using an algorithm for the chi-square distribution from Cook, Craven, and Clarke (1985) – it does not make it prohibitive. It also lets one select a significance level for pruning the decision tree.

2.3.2 Binarization

Bratko and Kononenko (1986) and Breiman et al. (1984) take a more radical approach, adapting the ID3 algorithm so that it treats all attributes as though they were binary attributes by grouping the various attribute values together. This is done by dividing the possible values into two sets and treating each set as though it were a single value. All possible 'binarizations' of values are tested for each attribute. If the attribute values are ordinal then only adjoining values can be grouped together, but if they are nominal then any combination is allowed. Most approaches split numeric attributes into just two ranges

ordered like the others? In the latter case, should the variable be split into two variables – one to record whether or not the subject was taken and a second to record the grade if it has been? This area has received little attention, but see Michalski and Chilausky (1980).

of values by testing every possible split and choosing the best. Evaluating more than two ranges would be prohibitively time-consuming. The effect of binarization is that each node of the tree can only have two branches, one for each of the two groups.

2.3.3 The gain-ratio measure

As mentioned above, Quinlan's (1986b) gain-ratio measure is another response to the problem of multi-valued attributes. This function favors attributes with fewer values, other aspects of the data being equal.

2.4 Dealing with missing data

Data sets with particular items missing are a common problem. However, it is important to distinguish between two different reasons for the absence of a value. Generally the item is missing by chance, because it has not been recorded or is unavailable for some reason. However, there are also situations in which there could not logically have been information because of relationships between the attributes. An example of the latter is where one attribute is "number of statistical samples" and another is "are the variances equal?" Clearly, if there is only one sample the second question is meaningless. Such cases will be referred to as *null* values rather than missing values.

The two situations require different remedies. In the second case the method is clear. If a node contains some examples with null values then the relevant attribute(s) cannot be considered for selection at that node. If such an attribute were selected, then situations could arise where classifying further examples would be impossible. Thus, equal variances can only be considered down those branches where there is more than one sample.

The first case is more difficult. Ideally the aim is to determine the missing value, but in most cases some form of estimate must be used instead. Quinlan (1986b) has examined a number of methods for estimating the unknown value, some of them quite complex. These include using the modal value, using Bayesian probabilities, determining the unknown value using a decision tree, treating 'unknown' as a new value for the attribute, and distributing the unknown examples according to the proportion of occurrences in the known examples. His results suggest that, of the more sophisticated methods, only the use of a distribution is an improvement on using the modal value.

The modal method was used for the experiments reported in Section 3. More precisely, if a particular example had a missing value, it was assumed to be (at that node only) the most common value of the attribute among those examples in the same class. Apart from being simple, this method seems the least likely to distort the results.

2.5 Small splits

When an attribute is selected at a node, the data are split into two or more subsets. These may have roughly equal numbers or they may be very unequal, depending on the attribute chosen. Often when the split is very unequal, the small group (possibly with only two or three examples) is purely of one class.

This small split then needs no further division and in general the more small, pure splits that are chosen the smaller will be the tree.

Breiman et al. (1984) and Quinlan (1986b), whose gain-ratio measure favors such splits, argue that this property is desirable because it produces smaller trees. However, from a statistical viewpoint this is not necessarily the case. These very small groups are quite likely to be chance occurrences and therefore unreliable for predicting new sets of data. Statistically, it is generally better to have larger groups of examples with a few in different classes, rather than a small group all in the same class. Marshall (1986), taking this point of view, favors equal splits and uses the correction outlined above to encourage this.

3. Experiments with decision-tree induction

The previous section described a number of different measures of goodness of split and various approaches to the problems of multi-valued attributes and small splits. The main purpose of the current research was to conduct a detailed comparison between these alternatives to determine their effect on the size, predictive accuracy, and usability of the induced decision tree. This section describes the test data used and the methodology of the experiments. In order to give a baseline for the comparisons, the experiments included a condition using no goodness of split measure, so that attributes at each node were chosen randomly.

3.1 Noise and residual variation

All the methods were tested on a number of different sets of data, which had varied characteristics and degrees of noise. These will be described shortly, but first it is necessary to discuss what is meant by noisy data. With deterministic data, an example in the data set can always be *correctly* classified from the known attributes, as can further examples. However, in many real problems there is a degree of uncertainty and/or error present in the data that leads to errors in classification.

Two different sources of uncertainty can be distinguished. One of these is mismeasurement: for a variety of reasons an incorrect value of an attribute or class can occur in the data, including the case of a missing value. This may happen because of incorrect recording or transcription, or because of incorrect measurement or perception at an earlier stage. This source of uncertainty will be termed *noise*. The second situation occurs when extraneous factors that are not even recorded affect the results, leading to variability that cannot be wholly explained in terms of the data available. In statistics this is called *residual variation*, and the same term will be used in this paper. In real-world problems this is often the greatest source of error.

It is also worthwhile distinguishing two different situations that underlie the use of this whole approach – classification and prediction. In classification, an example is a member of a particular class (e.g., a type of plant), and thus has certain attributes (e.g., size, color, shape). The causal relation is from class to attribute and the purpose of induction is to let one classify further examples from their attributes. In prediction, various factors (e.g., disease, patient, treatment) combine to produce some outcome (e.g., recovery or recurrence). The causal relation is from attributes to class, and the purpose is to predict future outcomes from the factors. Although the data appear the same, the effect of residual variation differs in the two cases. In classification the extra factors will affect the attributes, whereas in prediction they will affect the class. Generally, it is harder to compensate for uncertainty in the class than for uncertainty in an attribute.

3.2 Data sets

The experiments drew on four data sets, three from natural domains and one constructed artificially.

Profiles of B.A. Business Studies degree students (BABS). These data relate various attributes of each student, on entry to the course, to the final class of degree achieved. There are 186 observations with seven attributes – age (years), type of entry qualification (A-level,[3] BTEC Ordinary National Diploma, or some other), sex (male/female), number of O-levels, number of points at A-level (0–20), grade of maths O-level (A, B, C, FAIL), and full-time employment before the course (yes/no). There are four possible classes of degree – first, upper second, lower second, or third. Three of the attributes are integer and four symbolic. There is no known noise, but many other factors affecting the results have not been (and probably could not be) measured, giving high residual variation. This is an example of a prediction task.

The recurrence of breast cancer (Cancer). These data, containing 286 examples, are derived from those used in Bratko and Kononenko (1986) and concern the recurrence of breast cancer . There are two classes (recur or not recur) and nine attributes, of which four are integer. These include age, tumor size, number of nodes, malignant (yes/no), age of menopause (< 60, ≥ 60, not occurred), breast (left, right), radiation treatment (yes/no), and quadrant of breast (left, right, top, bottom, center). There are both missing data and residual variation. It is another example of a prediction task.

Classifying types of Iris (Iris). Kendall and Stewart (1976, p. 331) use these data as a test of discriminant analysis. There are 150 examples of three different varieties of Iris, with roughly equal numbers of each. The four integer attributes are measurements such as petal length and petal width, from which the examples can be classified. There is little noise or residual variation.

Recognizing LCD display digits (Digits). This is an artificial domain suggested by Breiman et al. (1984). A digit in a calculator display consists of seven lines, each of which may be on or off. Thus, there are ten classes (one for each digit) and seven binary-valued attributes (one for each line). Residual variation is introduced by assuming that a malfunction leads to a 10% chance of a line being incorrect. Such errors affect the attributes but not the class. Note that the chance of an example being completely correct is $0.9^7 = 0.48$. Three hundred cases were randomly generated. This is another example of a classification task.

[3] A-level and O-level are British national exams taken at ages 18 and 16 respectively. BTEC is the Business and Technology Education Council, which validates national exams.

3.3 Criteria for evaluation

There are three important criteria for evaluating a decision tree – size, accuracy and understandability. Thus, these are natural dependent measures for experiments on decision-tree induction.

Size. Occam's Razor is a generally accepted principle – the fewer terms in a model the better. This particularly holds for statistical models, in which one can always improve explanatory power on the training data by adding extra variables. However, such spurious additions will usually worsen the predictive ability of the model on independent test data. According to this view, one should attempt to minimize the size of the induced decision tree, as measured by the number of nodes or leaves. These two measures are related. Indeed, if the tree is strictly binary (i.e., every node has two branches), then the number of leaves equals the number of nodes plus one. If multi-valued attributes give a tree that is not strictly binary, then the number of leaves will be greater. The number of leaves has been selected as the measure of size in the present experiments because it corresponds to the number of distinct 'rules' contained within the decision tree.

Accuracy. This measure refers to the predictive ability of a decision tree in terms of classifying an independent set of test data. One can measure this ability in terms of the error rate, i.e., the proportion of incorrect predictions that a tree makes on the test data. This is a fairly crude measure, as it does not reflect the accuracy of predictions for the different classes within the data. Classes are not equally likely and those with few examples are usually predicted badly. Indeed, with heavy pruning there may be no rules left for a particular class, so that it can never be correctly predicted. Titterington et al. (1981) discuss measures that take into account the different classes.

More important is whether one uses the pruned or unpruned tree to classify the test data. With non-deterministic data, the basic algorithm can produce a very large tree in an effort to correctly classify every instance. Much of this structure will reflect chance occurrences, leading to inaccurate predictions on other data sets. Therefore one prunes the tree to remove such spurious branches. Ultimately, the accuracy of the pruned tree is most significant, as it would be used in practice. However, this paper also examines the performance of unpruned trees. This ensures that differences between the measures are not swamped by pruning, and also enables the gain from pruning to be estimated.

Understandability. Part of the rationale for expert systems is that they should represent knowledge explicitly so that the expert, and to a certain extent the user, can readily understand it. Certainly this is one advantage of decision trees over other statistical techniques that perform the same function, such as discriminant analysis. However, in comparison with most other representations used in machine learning, decision trees themselves are difficult to understand and therefore to validate (Cendrowska, 1987).

There is general agreement (Quinlan, 1986b; Mingers, 1986a; Shepherd, 1983; Kononenko et al., 1984) that deeper trees are less comprehensible. This is particularly the case with binarized trees, which can become very deep, often testing the same attribute many times down a branch. In addition, in binarizing multi-valued attributes, there is no guarantee that the resultant grouping of the attribute values will appear meaningful to the expert. Moreover, the expert may actually want the different values to be kept separate.

For example, with the degree student data, one of the most important attributes from a course manager's viewpoint is the type of qualification (A-level, BTEC OND or Other). A rule that splits students into these three categories at the same level would be preferred to one that arbitrarily groups pairs of them together. Similarly, in a domain involving expense claims in a company, one attribute (Department) had five values (Finance, Engineering, etc.), with different rules applying to the different departments. Maintaining the values separately produces a smaller and more easily understood tree than does splitting them up in some arbitrary fashion. The understandability of a tree is difficult to quantify or measure, so in these experiments it must be weighed against the results for size and accuracy, which can be quantified. Given two methods that perform equally well in terms of size and accuracy, one should prefer the method that produces the most easily understood trees.

3.4 Experimental method

In all, ten different measures and variants were tested on all four of the data sets, with random selection of attributes included as a control condition. The particular factors of interest, as mentioned above, were size, accuracy, and understandability. In order to get a realistic measure of the accuracy of trees as they would be used in practice, it was necessary to prune them. Of the various pruning methods available, Breiman's error complexity method was used for all the data sets, as this has proved very reliable and also generates a single, best-pruned tree. Empirical studies show that there is little or no interaction between pruning method and type of measure (Mingers, 1988).

To obtain independent test data and reliable results, each original data set was split randomly (70/30) into a training and a test set. The trees were grown and pruned on the training set and then accuracy was measured on the test set. In fact, the test set was not wholly independent since it is used in Breiman's pruning method. This develops a number of pruned trees entirely from the training set, but then selects the best via the test set.

To guard against random splits that happened to be untypical, the whole procedure was carried out nine times, giving nine independent pairs of training and test data for each data set. All the methods were run on the same datasets and the results averaged across the nine pairs.

3.5 Results on decision-tree size

Table 7 shows the sizes of the initial trees for each combination of metric and dataset. It records the number of leaves averaged across the nine testing sets, the first of the dependent variables examined in the experiment.

Analysis of variance (ANOVA) shows the very strong (and expected) differences between domains ($F = 90.8$, $F_{0.01} = 4.5$). The BABS and Cancer data, with high levels of residual variation, lead to very large, bushy trees with over

Table 7. Size of original tree (number of leaves) for different measures and domains.

Measure	BABS	Digit	Cancer	Iris	Total
G Standard	66.8	49.6	65.8	6.9	189.1
Normalized	74.7	49.6	66.0	6.9	197.2
Probability	76.1	49.6	66.4	11.8	203.9
Binarized	62.1	49.6	60.1	6.9	178.7
Marshall	77.7	49.6	75.7	8.0	211.0
χ^2 Standard	78.4	52.0	97.9	16.9	245.2
Probability	88.3	52.3	98.6	15.6	254.8
Binarized	76.9	52.0	84.0	16.9	229.8
Gain Ratio	55.0	32.4	63.9	6.2	157.5
GINI	65.8	49.7	67.0	6.8	189.3
Random selection	111.1	53.8	148.2	23.0	336.1
Mean (excl. random)	72.1	48.6	74.5	10.3	

seventy leaves. In contrast, the Iris data, with very little noise, has small trees. The Digit data are unusual in that the potential size of tree is limited. The seven attributes are all binary, so that each can only be used once along a path. This limits a tree to a depth of seven (and therefore $2^7 = 128$ possible leaves), if all attributes are involved. In fact, as the data becomes partitioned into subsets, several attributes become redundant and so $2^6 = 64$ is a more realistic maximum.

The analysis also shows that there are significant differences between the types of measure ($F = 4.8$, $F_{0.01} = 3.0$). The most obvious difference is that random selection (i.e., no measure) leads to decision trees that are roughly twice as large as in other conditions. The Digit data are an exception for the reason explained above. This shows the extent of the benefit to be gained by using a reasonable evaluation function.

Removing the 'random' condition shows that there are still significant ($F = 5.6$, $F_{0.01} = 4.0$) differences between the more informed selection strategies. There are two main effects here – the gain ratio produces the smallest trees, whereas the chi-square variants produce especially large ones. The gain-ratio result is not surprising, since the measure was designed with this goal in mind, and it confirms Quinlan's (1986b) results. However, the larger size of the chi-square trees was not expected and is not easily explained. One possible reason, mentioned earlier, is that chi square appears to be less sensitive than other measures to rows of the contingency table with zero frequencies. This means that it is less likely to select attributes with small, pure splits and thus will have larger trees.

When the chi-square and gain-ratio results are removed, the rest (G-statistic measures and GINI) are not significantly different at the .01 level. However, note that within the chi-square and G-statistic families, binarization gives the smallest trees and probability the largest. These differences are significant at the .05 level. A detailed comparison of trees generated with and without the

Table 8. Size of pruned tree (number of leaves) for different measures and domains.

Measure	BABS	Digit	Cancer	Iris	Total
G Standard	3.1	11.9	2.7	3.0	20.7
Normalized	3.1	11.9	2.7	3.0	20.7
Probability	2.0	21.4	2.9	7.0	33.3
Binarized	2.0	11.9	2.7	3.0	19.6
Marshall	2.0	11.3	3.0	3.0	19.3
χ^2 Standard	5.8	11.3	4.3	3.0	24.4
Probability	4.2	19.1	2.6	6.8	32.7
Binarized	2.0	11.3	2.8	3.0	19.1
Gain Ratio	2.3	10.1	4.2	3.1	19.7
GINI	3.1	11.0	4.4	3.0	21.5
Random selection	2.2	27.3	2.0	7.8	39.3
Mean (excl. random)	3.0	13.1	3.2	3.8	

Marshall correction indicates that it does reduce the number of small splits without increasing the tree size inordinately.

In summary, the use of any informed selection criteria halves the size of trees in comparison with a random selection strategy. Between the informed measures, there are significant differences, with gain ratio producing the smallest trees, and chi square the largest ones.

Table 8 shows the corresponding size of trees after pruning. Clearly, there is a dramatic reduction in the size of trees, with the BABS and Cancer trees being reduced to only two or three leaves. Decision trees for the Digit domain remain quite large because there are ten different classes and the tree generally maintains at least one leaf per class. ANOVA shows that there are no significant differences between selection methods ($F = 1.5$, $F_{0.01} = 3.0$), including the random selection, although this does generate a particularly large tree for the Digit data.

3.6 Results on decision-tree accuracy

Tables 9 and 10 show the error rates for unpruned and pruned trees, respectively. The error rate is the number of incorrect classifications on the test data, averaged across the nine sets.

Looking first at Table 9, the overall average error rates range from 50% to 7%. This shows both the marked differences in predictability between domains, and the inaccuracy of unpruned trees on independent test data (of course, they are 100% accurate on the training data). More interesting is that ANOVA shows no significant differences ($F = 1.8$, $F_{0.01} = 3.0$) between the measures, including the random selection method. In fact, random selection was not even the worst strategy on this dimension. Actually, this is not as surprising as it may appear. The main effect of a good measure is to reduce the size of the tree, rather than alter its accuracy. The original tree will be 100% correct on the training data (provided there are no contradictions in the data), no

Table 9. Accuracy of unpruned tree for different measures and domains, using percentage of incorrect classifications on the test data.

Measure		BABS	Digit	Cancer	Iris	Total
G	Standard	49.7	33.0	30.7	6.8	120.5
	Normalized	51.4	33.0	33.5	6.8	125.0
	Probability	51.2	33.9	35.0	7.0	127.1
	Binarized	46.8	33.3	30.4	6.8	117.3
	Marshall	52.0	33.3	35.7	7.1	128.1
χ^2	Standard	49.9	33.5	30.5	7.1	121.0
	Probability	50.7	34.4	34.6	9.8	129.5
	Binarized	49.6	33.5	28.9	7.1	119.1
Gain Ratio		55.0	34.5	31.2	6.6	127.3
GINI		48.7	34.8	32.3	6.6	122.4
Random selection		48.9	35.9	34.8	7.9	127.5
Mean (excl. random)		50.5	33.8	32.3	7.2	

matter how the attributes are chosen. The only difference will be the order of the attributes and the number of times attributes are tested down a branch. When used on another set of essentially similar data, the tree performs as well, whether it is large or small.

Table 10 shows equivalent results after tree pruning, showing that this process generally makes a significant improvement. The average error is reduced by about 25% for the BABS and Cancer data and by about 19% for the Iris data. Interestingly, it does not improve accuracy in the Digit domain. One explanation is that the artificially-generated residual variation is very consistent between training and test sets. In 'real' data, the effects are coincidental and are seldom repeated across sets of examples. If they were, then they would be a predictable effect that could be incorporated into the tree. Pruning is done precisely to remove these coincidences. With the Digit data, the fairly limited range of possible errors (each light can only be on or off) is actually repeated in the training and test data sets, making pruning unnecessary. Other samples of data from the Digit domain, created with differing amounts of noise, show the same pattern of results.

Moving to the differences between measures, ANOVA again indicates that these are not significant ($F = 2.4$, $F_{0.01} = 3.0$). This confirms results reported elsewhere (Mingers, 1988) that accuracy, in a particular domain, is almost entirely determined by method and extent of pruning rather than choice of measure. These negative results show that binarization and the gain-ratio measure do not produce more accurate results to counter their possible detrimental effect on the understandability of a tree.

Other methods of handling multi-valued attributes (normalization and probability) are also neither better nor worse. In fact, there is no evidence that multi-valued attributes cause any problems, although the largest number of attribute values was five, and larger numbers might have an effect. There is

Table 10. Accuracy of pruned tree for different measures and samples, using percentage of incorrect classifications on the test data.

MEASURE	BABS	DIGIT	CANCER	IRIS	TOTAL
G STANDARD	38.2	33.0	23.5	5.7	100.4
NORMALIZED	38.2	33.0	23.5	5.7	100.4
PROBABILITY	36.7	35.2	24.3	6.0	102.2
BINARIZED	36.9	33.0	23.5	5.7	99.1
MARSHALL	36.7	33.2	21.5	5.7	97.1
χ^2 STANDARD	37.3	34.4	24.3	5.7	101.7
PROBABILITY	38.1	37.6	25.4	5.7	106.8
BINARIZED	38.4	34.4	23.7	5.7	102.2
GAIN RATIO	36.9	35.4	24.7	5.3	102.3
GINI	38.2	33.4	24.0	5.7	101.3
RANDOM SELECTION	36.7	37.7	27.5	6.7	108.6
MEAN (EXCL. RANDOM)	37.6	34.3	23.8	5.7	

also no evidence that measures which favor small splits perform less well, although the high degree of pruning may remove such nodes from the final tree. Taking the results as they stand, the G statistic with the Marshall correction is marginally best and the chi-square statistic with probability is worst. It would be interesting to see if these results occurred with other data, or if much larger samples revealed significant differences.

4. Conclusions

The results reported in this paper show that the predictive accuracy of induced decision trees, both pruned and unpruned, is not sensitive to the goodness of split measure. This confirms Breiman et al.'s (1984) results. All of the methods tried are quite sophisticated and make good use of the information available. In fact, the results show that accuracy is not improved significantly by using a measure at all. Selecting attributes entirely randomly produces trees that are as accurate as those produced using a measure. However, the choice of measure does significantly influence the size of unpruned trees. Randomly selecting attributes produces trees roughly twice as large as those produced with an informed measure. Between the measures, the gain ratio generates the smallest trees, whereas chi-square produces the largest. After pruning, there is little difference in size.

The overall accuracy depends almost exclusively on the amount of noise and residual variation in the data and on the degree of pruning, not on the type of measure used. With these domains, it ranges from 7% to 37% error rates. Pruning yields increases in accuracy of between 19% and 25%, except for the Digit data, for which there was no improvement. Despite these improvements, the final accuracy with the Cancer and degree student data remained very low, reflecting the high level of residual variation in the data. Multi-valued attributes have not caused problems, and methods that take them into account (binarization, gain ratio, normalization, probability) are no more accurate than those that do not. In addition, methods which favor small splits that may be statistically unsound are no less accurate than those that do not. The understandability of decision trees is an important issue, but it cannot be easily measured or quantified. Researchers have suggested that deep trees, and binarized ones in particular, are less comprehensible. The results show that there are no grounds to prefer these on the basis of size or accuracy.

Further work is needed, both to confirm and extend these results. They should be repeated with different domains, including artificial domains with more diversity than the LCD digits. A variety of questions remain to be answered. Will larger data sets reveal significant differences? Does residual variation cause greater inaccuracy than noise? Are residual effects on the classes worse than on the attributes? Are certain forms of tree more easily understood than others?. What are the effects of differing misclassification costs and unequal prior probabilities? Future experiments should attempt to answer these questions.

The more general limitations of this approach should also be mentioned. First, the method is inevitably limited by the quality and quantity of the data available. As with all statistical techniques, methods for inducing decision trees can only capture patterns that are present in the data.

Second, the form of knowledge representation is very limited. The basic technique can only generate trees of a specific form; i.e., those in which each node tests the value of a single attribute. The method excludes trees that test the values of multiple attributes at a single node, or that examine relations between descriptors. Actually, the decision tree framework can be extended to handle such representations, but searching the space of such structures would be very difficult.

Third, decision trees are generally harder to understand than sets of individual rules or frames, although Corlett (1983) and Quinlan (1987) describe methods for transforming a decision tree into production rules. Also, they are difficult to modify without recreating them entirely, though Schlimmer and Fisher (1986) and Utgoff (1988) have examined methods for incremental tree construction.

Finally, the basic approach employs an algorithm that is only one-step optimal. At each node, it selects the best attribute, but it cannot backtrack once that choice has been made. One can create datasets in which a sub-optimal early attribute leads to a better (smaller) tree overall. Some form of branch and bound or dynamic programming could be used to explore the space of decision trees, but the improvement might not be worth the increased computational time.[4]

In summary, there exists a number of extensions to the standard method for inducing decision trees and future work should explore these variations. Moreover, systematic experimentation with both real and artificial datasets is the natural approach to determining whether such extensions constitute actual improvements over existing methods.

[4] This problem has been investigated by Michalski (1978).

Acknowledgements

Thanks to A. Martin, R. Hurrion, R. Dyson, R. S. Michalski, P. Langley, and the anonymous referees for their helpful comments on earlier drafts of the paper.

References

Bratko, I., & Kononenko, I. (1986). Learning diagnostic rules from incomplete and noisy data. *Seminar on AI Methods in Statistics*. London: Unicom Seminars Ltd.

Bratko, I., & Lavrac, N. (Eds.). (1987). *Progress in machine learning*. Wilmslow, England: Sigma Press.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Bundy, A., Silver, B., & Plummer, D. (1985). An analytical comparison of some rule-learning programs. *Artificial Intelligence*, *27*, 137–181.

Cendrowska, J. (1987). PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, *27*, 349–370.

Cook, D., Craven, A., & Clarke, G. (1985). *Statistical computing in Pascal*. London: Edward Arnold.

Corlett, R. (1983). Explaining induced decision trees. *Proceedings of the Third Technical Conference of the BCS Expert Systems Group*. London: British Computer Society.

Hart, A. (1984). Experience in the use of an inductive system in knowledge engineering. In M. Bramer (Ed.), *Research and developments in expert systems*. Cambridge: Cambridge University Press.

Hunt, E., Marin, J., & Stone, P. (1966). *Experiments in induction*. New York: Academic Press.

Kendall, M., & Stewart, A. (1976). *The advanced theory of statistics* (Vol. 3). London: Griffin.

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

Kullback, S. (1967). *Information theory and statistics*. New York: Dover.

Marshall, R. (1986). Partitioning methods for classification and decision making in medicine. *Statistics in Medicine*, *5*, 517–526.

Michalski, R. S. (1978). *Designing extended entry decision tables and optimal decision trees using decision diagrams* (Technical Report No. 898). Urbana: University of Illinois, Department of Computer Science.

Michalski, R. S., & Chilausky, C. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, *4*, 125–161.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (1983). *Machine learning: An artificial intelligence approach.* Los Altos, CA: Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (1986). *Machine learning: An artificial intelligence approach* (Vol. 2). Los Altos, CA: Morgan Kaufmann.

Mingers, J. (1986a). Inducing rules for expert systems – statistical aspects. *The Professional Statistician*, *5*, 19–24.

Mingers, J. (1986b). Expert systems – experiments with rule induction. *Journal of the Operational Research Society*, *37*, 1031–1037.

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

Mingers, J. (1987b). Rule induction with statistical data – a comparison with multiple regression. *Journal of the Operational Research Society*, *38*, 347–352.

Mingers, J. (1988). *A comparison of methods of pruning induced rule trees* (Technical Report). Coventry, England: University of Warwick, School of Industrial and Business Studies.

Quinlan, J. R. (1979). Discovering rules from large collections of examples: A case study. In D. Michie (Ed.), *Expert systems in the micro electronic age.* Edinburgh: Edinburgh University Press.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach.* Los Altos: Morgan Kaufmann.

Quinlan, J. R. (1985). Decision trees and multi-valued attributes. In J. Hayes & D. Michie (Eds.), *Machine intelligence* (Vol. 11). Chichester, England: Ellis Horwood.

Quinlan, J. R. (1986a). The effect of noise on concept learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2). Los Altos: Morgan Kaufmann.

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

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*, 221–234.

Schlimmer, J. C. & Fisher, D. (1986). A case study of incremental concept induction. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 496–501). Philadelphia, PA: Morgan Kaufmann.

Shepherd, B. (1983). An appraisal of a decision-tree approach to image classification. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (pp. 473–475). Karlsruhe, West Germany: Morgan Kaufmann.

Sokal, R., & Rohlf, F. (1981). *Biometry.* San Francisco: Freeman.

Titterington, D., Murray, L., Murray, G., Spiegelhalter, D., Skene, A., Habbema, J., & Gelpke, G. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society, A Series*, *144*, 145–175.

Upton, G. (1982). A comparison of alternative tests for the 2 × 2 comparative trial. *Journal of the Royal Statistical Society, A Series*, *145*, 86–105.

Utgoff, P. (1988). ID5: An incremental ID3. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 107–120). Ann Arbor, MI: Morgan Kaufmann.