13
Decision Trees and Multi-Valued Attributes

J. R. Quinlan
School of Computing Sciences,
New South Wales Institute of Technology, Australia

Abstract

Common induction systems that construct decision-trees have been reported to operate unsatisfactorily when there are attributes with varying numbers of discrete possible values. This paper highlights the deficiency in the evaluation of the relevance of attributes and examines a proposed solution. An alternative method of selecting an attribute is introduced which permits the use of redundant attributes. Results of experiments on two tasks using the various selection criteria are reported.

1. INTRODUCTION

As knowledge-based expert systems play an increasingly important role in artificial intelligence, more attention is being paid to the problem of acquiring the knowledge needed to build them. The traditional approach involving protracted interaction between a knowledge engineer and a domain expert is viable only to the extent that both these resources are available; this approach will not meet the apparently exponential growth in demand for expert systems. A solution to this dilemma requires rethinking the way knowledge-based products are built. An example of this reappraisal of methodology appears in Michie (1983), and is based on the principle of formalizing and refining the knowledge implicit in collections of examples or data bases.

Dietterich and Michalski (1983) give an overview of methods for learning from examples. There are many such, all based on the idea of inductive generalization. One of the simplest of these methods dates back to work by Hunt in the late fifties (Hunt *et al.*, 1966). Each given example, described by measuring certain fixed properties, belongs to a known class and the 'learning' takes the form of developing a classification rule that can then be applied to new objects. Simple though it may be, derivatives of this method have achieved useful results; Kononenko *et al.* (1984), for example, have managed to generate five medical diagnosis systems with minimal reference to diagnosticians.

In the course of their work, Kononenko *et al.* uncovered a deficiency in the basic machinery being used, and this paper focuses on that shortcoming. We first formalize the inductive task and present analytical support for the existence of a deficiency. The solution proposed by Kononenko *et al.* is discussed and an alternative method introduced. A series of trials of the various methods on two classification tasks is then reported, revealing their relative merits.

2. CONSTRUCTING DECISION-TREES

We imagine a universe of *objects*, each described in terms of a fixed number of *attributes* or properties. Each attribute has its own (small) set of discrete attribute values. Each object belongs to one of several mutually exclusive classes. For example, we could have the following scenario:

objects:	people
attributes:	colour of hair (with attribute values red, brown, fair, grey, black)
	colour of eyes (brown, black, blue)
	height (tall, medium, short)
classes:	unremarkable appearance, unusual appearance.

The concept learning task of interest here can now be stated briefly as:

given:	a collection of objects (and their descriptions) whose class membership is known
find:	a classification rule, couched in terms of the attributes, that will assign any object to its class.

The given set of objects is usually referred to as the *training set.* The method used to find a rule is induction in which observations about the training set are generalized so as to apply to other, as yet unseen objects. While it is possible to ensure that the developed rule works for all members of the training set, the correct performance of the rule on other objects cannot usually be guaranteed. Instead, we rely on heuristic guides such as Occam's Razor: among all rules that accurately account for the training set, the simplest is likely to have the highest success rate when used to classify objects that were not in the training set.

In this discussion we will assume that there are only two classes, Y and N, although all the results can be extended to an arbitrary number of classes in a straightforward way. We will also limit ourselves to classification rules expressed as decision trees. Each interior node of such a tree is a test, based on a single attribute, with a branch for each possible outcome of the test. Each leaf of a decision-tree is labelled with a class. An object is classified by starting at the root of the tree,

performing the test, taking the branch appropriate to its outcome, and continuing the subtree at that branch. When a leaf is eventually encountered, the object is taken to be a member of the class associated with the leaf.

Forming a decision-tree by induction from a training set comes down to deciding, from the descriptions and known classes of objects in the training set, which attribute-based test to use for the root of the decision tree. For this test (with v outcomes, say) will partition the training set into v blocks, one for each branch emanating from this root node. Each block can be treated as a training set in its own right and the same procedure applied recursively until all (sub) training sets contain objects of a single class. A method of choosing a test to form the root of the tree will be referred to as a *selection criterion.*

One obvious choice for an attribute-based test is to branch on the value of an attribute, creating a separate path for each possible value it can have. Choosing a test for the root could be carried out by the trivial algorithm: choose the first attribute first, then the second attribute and so on. However, the decision-tree built by this procedure would not be expected to reflect any structure in the training set and so would have poor predictive performance. A better strategy, employed for example by ACLS (Michie, 1983; Shapiro, 1983; Shepherd, 1983) and ID3 (Quinlan 1982, 1983 a, b), is to use an information-theoretic criterion as follows. If the training set contains y objects from class Y and n from class N, the information that needs to be supplied by a classification rule for the set can be related to the relative frequencies of class membership by the function

$$I(y, n) = -\frac{y}{y+n}\log_2\left(\frac{y}{y+n}\right) - \frac{n}{y+n}\log_2\left(\frac{n}{y+n}\right).$$

Now, let A be an attribute with possible values $A_1, A_2, \ldots, A_v$, and let y_i and n_i denote the numbers of objects of class Y and N respectively that have the ith value A_i of A. If attribute A was chosen as the root of the decision-tree, with a branch for each of its v possible values, the information that would need to be supplied by the (sub) tree corresponding the the branch for A_i is similarly

$$I(y_i, n_i).$$

Weighting each branch of the decision-tree by the proportion of objects in the training set that belong to that branch, we can write the expected information requirement after testing attribute A as

$$E(A) = \sum_{i=1}^{v} \frac{y_i + n_i}{y+n} I(y_i, n_i).$$

Naturally, the expected information needed after testing attribute A is

generally less than the information needed before any attribute is tested. The information gained by branching on attribute A is just

$$\text{gain}(A) = I(y, n) - E(A).$$

The information-based criterion referred to earlier can be expressed simply as: choose the attribute whose information gain is maximal. In the following, this will be called the *original criterion.*

3. MULTI-VALUED ATTRIBUTES

Kononenko *et al.* (1984) have developed an inductive inference system ASSISTANT and used it to built classification rules in several medical domains. At one stage of its evolution, their system used the original criterion of information gain to select attributes as above. In the course of their experiments they encountered a problem when the attributes being compared had different numbers of values. In one study, medical specialists were of the opinion that the attribute 'age of patient', with nine discrete ranges, was being chosen over more relevant attributes with fewer values. The choice of an inappropriate attribute results in excessive fragmentation of the training set; structure in the set becomes harder to detect and the performance of the classification rule on unseen objects may be degraded. In this case, the opinion of the specialists was borne out by the fact that, when the attribute was omitted altogether, the induced classification rule gave better results.

Let us analyse the problem in more abstract terms. Suppose we form an attribute A' which is identical to A except that two of the attribute values, A_1 and A_2 say, are collapsed into a single value A'_{1+2}. A' then has $v-1$ values $A'_{1+2}, A'_3, \ldots, A'_v$, where there are now y_1+y_2 and n_1+n_2 objects from classes Y and N respectively that have value A'_{1+2} of the new attribute A'. Let us examine the difference in information gain between A and A'. Since $I(y, n)$ is unchanged, this difference is

$$E(A') - E(A).$$

For values of $i > 2$, corresponding terms in this difference cancel so the difference reduces to terms related to A_1 and A_2 on one hand and A'_{1+2} on the other. The difference can then be written as

$$\frac{y_1+y_2+n_1+n_2}{y+n} I(y_1+y_2, n_1+n_2) - \frac{y_1+n_1}{y+n} I(y_1, n_1) - \frac{y_2+n_2}{y+n} I(y_2, n_2).$$

The minimum value of this difference can be found by equating its partial derivative with respect to y_1 to zero. The minimum value occurs when

$$\frac{y_1}{y_1+n_1} = \frac{y_2}{y_2+n_2} = \frac{y_1+y_2}{y_1+n_1+y_2+n_2}$$

which gives the minimum value of the difference as zero. The upshot of this analysis is that the information gain attributable to A will generally exceed that attributable to A', the two gains only being equal if the proportions of class Y and class N objects in the two merged attribute values are identical.

Now let us look at the situation from the other side of the coin. Suppose that the values of attribute A are sufficiently 'fine' for the classification task at hand. If we were arbitrarily to increase the number of values of A by subdividing existing values, we would not expect to increase the usefulness of A for a classification rule; on the contrary, we would intuitively expect the excessive fineness of A to obscure structure that may exist in the training set. But the above analysis shows that the information gain of the new, finer A will generally be increased, thereby boosting its chances of being chosen as the most relevant attribute. By analogy, there would seem to be a bias in the information gain criterion towards attributes with larger numbers of values. This analysis supports the empirical finding of Konenenko *et al.*

4. BINARY TESTS

The remedy implemented in ASSISTANT is the requirement that all tests have only two outcomes. If we have an attribute A as before with v values $A_1, A_2, \ldots, A_v$, the decision-tree no longer has a branch for each possible value. Instead, a subset S of the values is chosen and the tree has two branches, one for all values in the set and one for the remainder. The information gained is then computed as if all values in S were amalgamated into one single attribute value and all remaining values into another. In this selection criterion, referred to as the *subset criterion,* the test chosen for the root uses the attribute and subset of its values that maximizes the information gain. Kononenko *et al.* report this modification led to smaller (but less structured) decision-trees with an improved classification performance. In one medical domain, for example, the decision-tree formed from a training set of 1300 objects was reduced from 525 to 157 nodes, and its classification accuracy on 545 unseen objects improved from 62 per cent to 66 per cent.

Limiting decision trees to a binary format is reminiscent of the original concept learning system CLS (Hunt *et al.*, 1966). In that system, each test was of the form 'attribute A has value A_i', with two branches corresponding to true and false. This is clearly a special case of the test implemented in ASSISTANT, which permits a set of values, rather than a single value, to be distinguished from the others. CLS, however, did not use an information-theoretic measure to evaluate tests, but rather employed a lookahead scheme based on a system of measurement and misclassification costs. Nevertheless, designating a single value and

evaluating tests using information gain as before seems worthwhile exploring as a comparator for ASSISTANT's selection criterion, and will be referred to as the *single-value criterion.*

If all tests must be binary, there can be no bias in favour of attributes with large numbers of values and so the objective has certainly been achieved. It could be argued, however, that ASSISTANT's remedy has undesirable side-effects that have to be taken into account. First, it could lead to decision-trees that are even more unintelligible to human experts than is ordinarily the case, with unrelated attribute values being grouped together and multiple tests on the same attribute. More importantly, the modified procedure can require a large increase in computation. An attribute A with v values has 2^v value subsets and, when trivial and symmetric subsets are removed, there are still $2^{v-1}-1$ different ways of specifying the distinguished subset of attribute values. The information gain realized with each of these must be investigated, so a single attribute with v values has a computational requirement similar to $2^{v-1}-1$ binary attributes. This is not of particular consequence if v is small, but the approach would appear infeasible for an attribute with 20 values. There are applications for which such a large number of attribute values is not unreasonable; for example, the attribute 'family' for Australian spiders would have 39 values (Clyne, 1969).

5. NORMALIZING THE GAIN

Another method of overcoming the problem posed by attributes with different numbers of values would be to normalize the information gain in some way. This was attempted by Kononenko *et al.* (1984): if an attribute had v values, the normalized gain was computed as the 'raw' gain divided by $\log_2(v)$. The results achieved with this procedure were unsatisfactory, as very important attributes with large numbers of values were now discriminated against, at least near the root of the tree. For example, an attribute with eight values would have to achieve three times the information gain of a binary-valued attribute if it were to be the chosen attribute.

6. GAIN RATIO

This paper suggests an alternative information-based criterion that resembles a normalized gain, although the rationale for the criterion is quite different.

Consider again our training set containing y and n objects of class Y and N respectively. Let attribute A have values $A_1, A_2, \ldots, A_v$ and let the numbers of objects with value A_i of attribute A be y_i and n_i, respectively. Enquiring about the value of attribute A itself gives rise to

information, which can be expressed as

$$IV(A) = -\sum_{i=1}^{v} \frac{y_i + n_i}{y + n} \log_2\left(\frac{y_i + n_i}{y + n}\right).$$

Notice that this information measure is unrelated to the utility of A for classification purposes. For example, if

$y_1 = y_2 = \cdots = y_v$; and

$n_1 = n_2 = \cdots = n_v$

attribute A would be useless as the root of the decision-tree, and yet the information from determining the value of attribute A would be maximal.

$IV(A)$ thus measures the information content of the answer to the question, 'What is the value of attribute A?' As discussed earlier, gain(A) measures the reduction in the information requirement for a classification rule if the decision tree uses attribute A as root. Ideally, as much as possible of the information provided by determining the value of an attribute should be useful for classification purposes or, equivalently, as little as possible should be 'wasted'. A good choice of attribute would then be one for which the ratio

$$\text{gain}(A)/IV(A)$$

is as large as possible. This ratio, however, may not always be defined—$IV(A)$ may be zero—or it may tend to favour attributes for which $IV(A)$ is very small. We therefore propose the following criterion: from among those attributes with an average-or-better gain, select the attribute that maximizes the above ratio. This will be called the *ratio criterion.*

7. EMPIRICAL INVESTIGATION

The various criteria for selecting attributes as discussed in earlier sections were embodied in the straightforward tree-constructing procedure and evaluated on a family of tasks. This family was derived from an existing classification task, with a universe of 551 objects described in terms of 39 two-valued attributes for which the smallest known decision-tree contained 175 nodes (although smaller trees were discovered in the course of these experiments). In order to observe the effects of multi-valued attributes in stark relief, related tasks were synthesized by collapsing four of the attributes into a single attribute; these tasks thus had 36 attributes, one of them having 16 values and the remainder two values.

Three different choices of the four attributes to be combined into a single attribute were as follows:

*D*1: the two most important attributes were combined with two attributes of limited use

*D*2: the attributes were chosen to produce the most even distribution over the 16 values of the combined attribute

*D*3: the attributes were chosen to produce the most uneven distribution, subject to the requirement that all 16 values were represented.

Each selection criterion was evaluated on the original problem (D0) and on all the derived tasks. The same procedure was followed in each case. First, the entire 551 objects were presented as the training set to observe the size of the resulting decision-trees. Next, 20 randomly selected subsets containing 50 per cent of the 551 objects were set up and used as training sets. Since these training sets were incomplete, the decision-trees formed from them were not exact: each was tested on the remaining 50 per cent of the objects to measure the number of classification errors that resulted. Finally, to simulate forming more inaccurate classification rules, a similar procedure was followed using 20 per cent of the objects for the training set and evaluating the decision trees on the remaining 80 per cent.

The results of these experiments are summarized in Table 1 and Figures 1 and 2. Table 1 shows the sizes of the decision trees obtained from all 551 objects. For the original task (D0) in which all attributes are two-valued, the subset and single-value selection criteria are identical to the original, but noticeable differences emerge on the derived tasks. The ratio criterion does very well on the original task, giving a decision-tree of 143 nodes that is considerably smaller than any other known correct tree for this task. The same selection criterion, however, produces a much larger decision tree for task D1.

The most important characteristic of a good selection criterion, though, is that it should lead to decision-trees that accurately classify unseen objects. Figure 1 refers to the experiments in which decision-trees formed from half of the 551 objects were tested on the remaining half. For each task and selection criterion, the figure shows the 95 per cent confidence interval for the mean number of classification errors over the

Table 1. Number of nodes in correct decision tree.

Selection criterion	D0	D1	D2	D3
Original	175	205	187	187
Subset	175	205	169	163
Single-value	175	179	167	185
Ratio	143	265	179	179

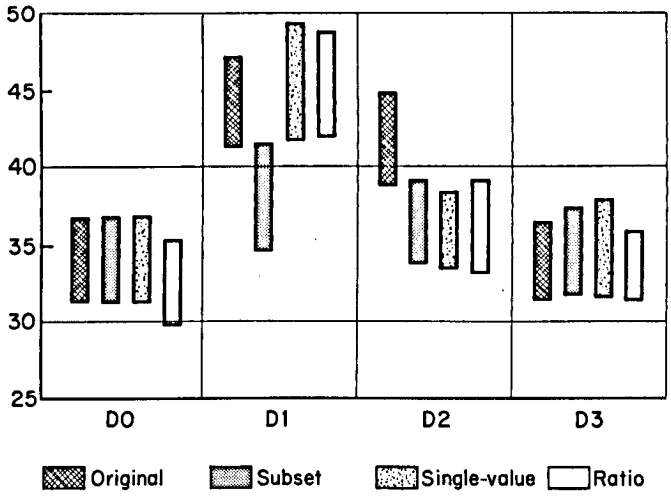

Figure 1. Mean number of errors with training-set of 275 objects.

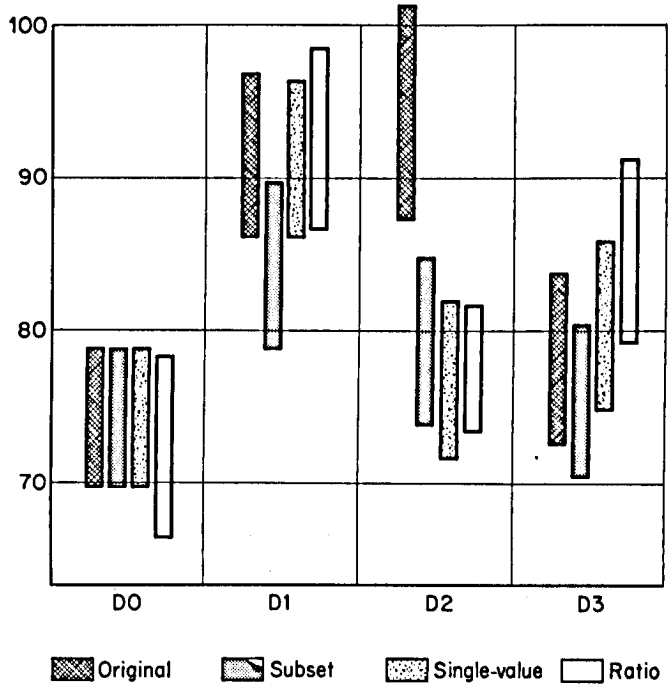

Figure 2. Mean number of errors with training set of 110 objects.

20 trials. These indicate that the subset criterion is significantly better on D1 while the original criterion is clearly worse on D2. Figure 2 refers to the similar experiments in which the training set contained 20 per cent of the objects and the resulting decision-trees were then tested on the 80 per cent of unseen objects. Once again a similar pattern emerges.

These results support the finding of Kononenko *et al.* that the original selection criterion can be somewhat deficient in a task with multi-valued attributes. For task D2 in Figure 2, changing from the original to the subset selection criterion improved the mean classification accuracy on unseen objects from 79 per cent to 82 per cent, and this difference would probably increase if more multi-valued attributes were involved.

8. REDUNDANT ATTRIBUTES

All selection criteria appear to have more difficulty with the task D1, as seen in both the size of decision-tree for the complete training set and the errors made by decision-trees constructed from partial training sets. Recall that this task aggregates both important and unimportant attributes, and thereby models a common real-world situation in which coarse values of an attribute are all that is required for classifying most objects, but much more precise values are needed for rarer cases. Two examples should illustrate the idea. In a thyroid diagnosis system (Horn *et al.*, 1985) many cases can be classified by knowing simply whether the level of each measured hormone is normal or not, but some cases require the level to be divided into as many as seven subranges. The study on Australian spiders mentioned earlier divides the 39 families into six groups, where the group alone often provides sufficient information for classification purposes.

The obvious remedy is to incorporate redundant attributes, each measuring the same property at different levels of precision appropriate to different classification needs. In the examples above, we might have both hormone level (seven values) and whether normal (two values), and both spider family (39 values) and group (six values). It would seem that the human experts, who provide the attributes in the first place, would have little difficulty in specifying these different precision levels useful for particular subsets of objects.

Let us now see what effect the introduction of a redundant attribute might be expected to have on the decision trees produced by the various selection criteria. Suppose A is some attribute with a full complement of values and A' is a redundant attribute with a lower level of precision, i.e. at least one value of A' corresponds to a subset of values of A. We have shown earlier that the information gain using A' can never exceed that using A, so the original selection criterion will never prefer A' to A. That is, adding the redundant attribute A' will have no effect on the

decision-tree formed. When the subset selection criterion is used, it is apparent that any subset of the values of A' can also be expressed as a subset of the more finely divided values of A, so including redundant attribute A' will not increase the range of tests available. However, some value of A may not be represented in a small training set while the corresponding coarser value of A' is represented, so tests derived from small sets of objects may be more accurate using A' rather than A. In the case of the single-value criterion, however, adding A' may have a beneficial effect by broadening the range of possible tests, as one attribute value of A' may correspond to a subset of the values of A. Finally, the attribute information $IV(A')$, will generally be less than $IV(A)$, so the introduction of A' would be expected to have an effect on the ratio criterion, although whether this effect is beneficial or otherwise is not clear. To summarize: the addition of a redundant attribute will not change the decision-trees produced by the original criterion, should not have a dramatic effect on those produced by the subset criterion, but may alter the trees produced using the single-value and ratio criteria.

These observations were tested by rerunning the previous trials, this time including both the original binary attributes as well as the composite 16-valued attributes. Each original attribute is then redundant in the

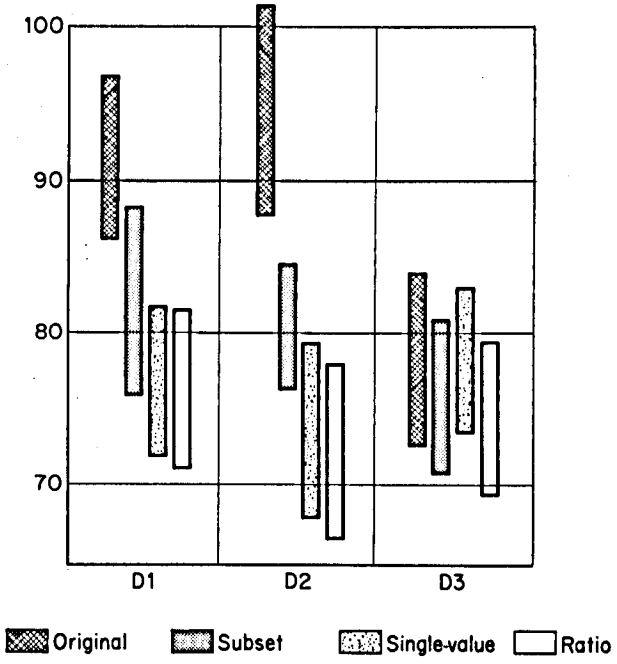

Figure 3. Mean number of errors, original attributes included, with training-set of 110 objects.

sense above, because each of its values corresponds to eight values of the composite attribute. Each of the tasks D1, D2 and D3 now has 40 attributes, one with 16 values and the rest with two. As before, the trials included 20 'runs' on each task, selecting a fixed proportion of the set of 551 objects as a training set and testing the tree produced on the remainder.

Figure 3 summarizes the results when 110 objects (20 per cent) were used as a training set, and shows the mean number of errors when the trees were used to classify each of the other 80 per cent of the objects. If this figure is compared to the corresponding sections of Figure 2, the following points emerge. As expected, the inclusion of the additional redundant attributes has no effect on the trees produced using the original selection criterion. There are small changes in the mean error with the subset criterion, and significant improvements with the single-value and ratio criteria. There is a particularly noticeable decrease in mean errors with the ratio criterion on D1 and D3. Notice also that the ratio criterion now gives marginally lower errors on all tasks than the other criteria.

9. CONFIRMING EXPERIMENT

The results above were put to the test on a completely unrelated classification task. In this task there were 1987 objects described in terms of 14 attributes, five with three values and the remaining nine with two. Despite the larger number of objects, this is a much simpler classification task as a correct decision tree with only 48 nodes was previously known.

Twenty training sets were prepared by randomly selecting half of the 1987 objects. For each of the four selection criteria, decision-trees were formed from each of these training sets and tested on the remaining objects. As would be expected from the simpler concept being formed and from the larger training sets, there were relatively few errors when the trees were used to classify the unseen objects. It was observed that, for three of the multi-valued attributes, one attribute value was more important than the others. Following that philosophy of redundant attributes discussed above, a redundant binary attribute was added for each of these three. In this redundant attribute, the two less important values were merged into a single value. The runs were repeated, this time using the augmented set of 17 attributes.

The results are summarized in Figure 4. Notice that, since no attribute has more than three values, selecting a non-trivial subset of values is equivalent to selecting a single value, so the subset and single-value criteria give identical results. In the first runs, the original and ratio criteria emerge as less useful than the others, because the decision-trees formed using them give a higher error rate. When the three redundant

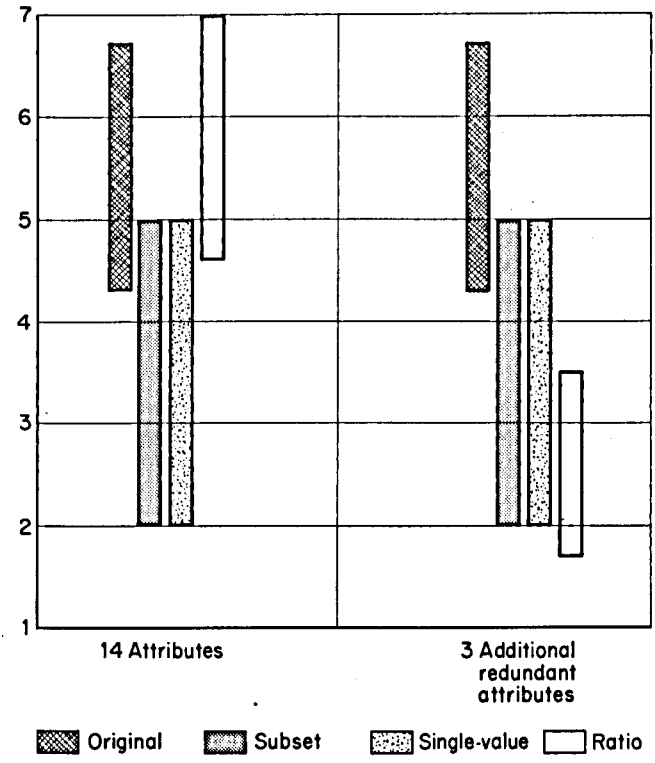

Figure 4. Mean number of errors with training-set of 993 objects.

attributes are added, however, only the ratio criterion is affected: it now gives significantly better results than any of the other three criteria.

10. CONCLUSION

Several observations can be made regarding these results. Analysis and experiments both support the findings of Kononenko *et al.* (1984) regarding the deficiency of the original selection criterion when attributes with differing numbers of values are present. This deficiency will tend to favour attributes with larger numbers of values, producing a decision-tree that has a higher error rate when classifying unseen objects.

The solution proposed by Kononenko *et al.* is to restrict tests in decision trees to binary outcomes, i.e. whether or not the value of an attribute is in a designated set. This has been found to reduce the size and improve the accuracy of decision-trees. However, the computational requirements of the subset selection criterion may make it infeasible for tasks containing attributes with many values. This technique has been compared to a similar binary restriction explored by Hunt *et al.* (1966),

the single-value criterion, which makes no such exponential computational demands. The single-value criterion has also been found to generate slightly more accurate decision-trees than the original criterion.

We have also proposed and investigated a selection criterion based on the ratio of information gain to attribute information. This has been found generally to perform about as well as the single-valued criterion. However, it has two noteworthy advantages. It does not restrict the decision-tree to a binary format which may be awkward and unnatural for some applications. More importantly, it is able to benefit from the provision of redundant attributes whose levels of detail, as expressed by their number of possible values, can be chosen to suit different classification needs. When suitable redundant attributes are provided, the ratio criterion has been observed to outperform the other three criteria, even though its computational requirements are roughly equivalent to those of the original selection criterion.

REFERENCES

Clyne, D. (1969) *A guide to Australian spiders.* Nelson, Sydney.

Dietterich, T. G. and Michalski, R. S. (1983) A comparative review of selected methods for learning from examples. In *Machine learning* (eds R. S. Michalski, J. Carbonell and T. Mitchell). Tioga, Palo Alto, Calif.

Horn, K., Compton, P., Lazarus, L., and Quinlan, J. R. (1985) The implementation of an expert system for the interpretation of thyroid assays in a clinical laboratory. *Australian Computer Journal* **17**, 1.

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

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

Michie, D. (1983) Inductive rule generation in the context of the Fifth Generation, *Proc. Int. Machine Learning Workshop,* University of Illinois at Urbana-Champaign.

Quinlan, J. R. (1982) Semi-autonomous acquisition of pattern-based knowledge. In *Machine intelligence 10* (eds J. E. Hayes, D. Michie, and Y.-H. Pao). Ellis Horwood, Chichester.

Quinlan, J. R. (1983a) Learning efficient classification procedures. In *Machine learning: an artificial intelligence approach* (eds R. S. Michalski, J. Carbonell, and T. Mitchell). Tioga, Palo Alto, Calif.

Quinlan, J. R. (1983b) Learning from noisy data. *Proc. Int. Machine Learning Workshop,* University of Illinois at Urbana-Champaign.

Shapiro, A. (1983) The role of inductive learning in expert systems. Ph.D Thesis, University of Edinburgh.

Shepherd, B. A. (1983) An appraisal of a decision tree approach to image classification. *Proc. IJCAI-8* (Karlsruhe).