J. Opl Res. Soc. Vol. 38, No. 1, pp. 39–47, 1987
Printed in Great Britain. All rights reserved

0160-5682/87 $3.00 + 0.00

Expert Systems—Rule Induction with Statistical Data

JOHN MINGERS
Polytechnic of the Southbank

Rule induction has been proposed as a way of speeding up the acquisition of knowledge for expert systems. Quinlan's ID3 algorithm has been used successfully but can only deal with determinate data. This paper explores extensions to deal with statistical data.

Key words: expert systems, knowledge acquisition, rule induction, statistical data

INTRODUCTION

In a previous article,[1] the use of rule induction in expert systems was discussed. This was based on work by Quinlan,[2] who had produced an algorithm (ID3) to perform the process of rule induction on very large sets of data. The data consists of a set of examples from some particular knowledge domain. Each example has a number of attributes and can be classified into a particular class. The algorithm attempts to find a tree of rules (in the form of IF. . .THEN clauses) which will correctly classify all the examples on the basis of their attribute values. Quinlan was working with data which was deterministic—i.e. the data *could be classified* purely from the attributes—but this paper extends the method to deal with stochastic data.

Stochastic data will contain a degree of random, chance variations within it as well as basic relationships between the variables. The ID3 algorithm assumes that the data is determinate, and so attempts to produce a rule tree able to classify every single instance. Some, if not most, of the rules in the tree will be based on pure chance. The algorithm has no concept of statistical significance, and so cannot distinguish between a significant and an insignificant improvement in the information measure.

This difficulty has been identified by Hart,[3] who suggested using the chi-square contingency table test in the place of the information measure. Hart's reasoning is as follows:

The information measure (IM) is used to select attributes which are good discriminators within the current set of data. The IM is actually calculated from a contingency table,[1] and since χ^2 is doing essentially the same thing—detecting the degree of association—it could replace the IM. This would have two advantages. First, once the selected attribute was insignificant (at, say, the 10% level), generation of that branch could be terminated. Secondly, it could take into account, via the degrees of freedom, the number of possible values of the attribute. The more values, the more likely an attribute is to discriminate well by chance, and therefore the IM tends to be biased towards attributes with a number of possible values and against attributes with only two.

This approach was mentioned in the original work by Hunt *et al.*[4] on which Quinlan's method is based, and Quinlan has also looked at chi-square,[5] although as a complement rather than a replacement for the IM.

This paper reports on investigations into these ideas; discusses the use of chi-square and its problems; looks at alternatives and considers the statistical distribution of the information measure and the G-statistic; and presents results of testing the system as well as measures of significance for the resulting rule trees.

USING THE CHI-SQUARE TEST

The ID3 algorithm, with the enhancements mentioned previously,[1] was modified to calculate χ^2 instead of IM, and will henceforth be called ID5. Initial tests on small, deterministic problems confirmed that it worked and generally generated the same or similar trees to the IM.

Sample problem with noisy data

To test the use of χ^2 on noisy data, the last two graduating years of the Business Studies degree were used (see Table 1), with the aim of trying to predict students' final grade of degree from certain of the attributes they possessed when starting the course.

Initial runs were performed using χ^2 purely as a means of choosing attributes—not of judging their significance—on the two years separately and on the data as a whole. These runs were compared with runs using the IM measure. The general conclusions were first that both methods produced extremely bushy trees, indicating little if any reduction in the data, though identifying important rules. Generally, IM produced smaller trees than χ^2. Secondly, they both produced similar trees, at least in the first few levels. Beyond these, however, the trees were quite different.

Using χ^2 to cut off the rule-tree

The next stage was to use the χ^2 value to cut off the tree when the best value at a branch was insignificant. In the resulting tree, the leaves would no longer have items all of the same class. The class of a leaf would therefore become probabilistic, based on the relative frequencies of the items there.

To determine the significance correctly, an algorithm to calculate the probability of a particular χ^2 value would need to be incorporated. The choice between attributes could then be based on the smallest (i.e. most unlikely) probability, and the tree stopped when the probabilities were all greater than the chosen significance level. Unfortunately, this would dramatically increase the run time of the program. A second alternative is to incorporate a section of the χ^2 tables for varying degrees of freedom and significance levels, and have the program use these. This approach is adequate for cutting off the tree but does not allow for the selection of an attribute from a number with the same level of significance if they have differing degrees of freedom. A third, approximate, method is to modify the calculated χ^2 value to take into account the degrees of freedom (ν) by, for example, working out the number of standard deviations away from the mean. This has the advantages of being easily understandable and able to be used both for the selection of attributes and for cutting the tree.

Problems with simple chopping of the tree

At this point, a major problem with the approach of stopping the generation of a branch when there were no more significant attributes became apparent. On looking at a whole tree, with the χ^2 values at each node, there were situations where a later node was significant, even though earlier nodes on that branch had not been. These later nodes would never have been reached because generation would have stopped at the earlier, insignificant nodes. The basic reason for this is that some attributes really work together jointly. Either one, by itself, is insignificant, but taken together they classify the data well. For example, in the following set of rules, neither A1 nor A2 can, by itself, classify the data at all.

TABLE 1. *Sample data for degree students*

Age	Sex	Qualification	A-level score	O-level maths	No. of O-levels	Work exper.	Degree
18	M	A	8	C	7	N	2.2
18	M	A	9	C	5	N	2.1
17	F	A	6	C	7	N	2.1
21	M	A	6	C	6	Y	2.2
19	M	A	6	C	8	N	2.2
18	M	A	3	C	7	N	3rd
17	F	A	2	B	8	N	2.1
28	F	B	—	C	5	Y	2.2
33	F	O	—	F	—	Y	2.2
20	M	A	8	A	10	Y	1st
17	F	A	9	B	9	N	2.1
17	M	A	11	C	8	N	2.1

Notes: (i) Qualification (for entry) can be A-levels (A), BEC OND (B) or other (O).
(ii) O-level maths is the grade (A–C) or fail (F) or not taken (F).
(iii) Work experience means significant experience, not just holiday jobs.
(iv) '—' means not relevant, e.g. A-level score if you have not taken A-levels.

```
IF  A1
    = 1  THEN  IF  A2
                       = 1  :Class1
                       = 2  :Class2
    = 2  THEN  IF  A2
                       = 1  :Class2
                       = 2  :Class1
```

This means that a more sophisticated approach is needed—pruning rather than merely chopping off branches. This involves initially generating the entire tree and then, starting at the tip or leaf of each branch, working backwards until the first significant node is found. That node will then be kept. However, this does not mean that all the later nodes can automatically be dropped. Before a node can be dropped, it is necessary to check all those nodes which come after it on various branches to see if any of them are significant. If any are, the node must remain. In other words, a node must be kept either if it itself is significant, or if any nodes on branches stemming from it are significant.

This is part of the more general problem of dependence or relationships between the attributes. The above situation is one of logical dependence, where the effects of attribute values depend on prior attributes in the tree, and is dealt with successfully. This approach to rule induction, however, cannot cope with rules involving functional dependence between the attributes. For example,

$$\text{IF attrib1} = \text{attrib2 THEN}\ldots$$

or

$$\text{IF attrib1} \geqslant 2 \times \text{attrib2 THEN}\ldots$$

This is because it has no method of generating the possible functional relationships. This is a very difficult problem, as there is an infinite variety of possible functional relationships between any or all of the attributes, and it is not addressed in this paper.

Other difficulties with χ^2

The major one is that of statistical unreliability because of the small number of observations involved in the lower levels of the tree. In generating the tree, the number of data items left at any particular point gets progressively smaller until there may only be two or three awaiting classification. The χ^2 statistic is particularly sensitive to this, and it is not a problem which can really be overcome, although Fisher's exact test could be used when the frequency is small.

It was also found that using the normalized value certainly did correct the bias towards multi-valued attributes to the extent that they were almost never chosen, and the tree became extremely deep and even harder than normal to understand and interpret. This brings out the fact that the statistic is being used to perform two different functions—first, the *selection* of attributes in generating the tree and, secondly, judging their *significance* in pruning the tree. In evaluating possible measures, these different functions have different criteria. For significance, it is purely the statistical reliability of the measure which is important. In selecting attributes, however, wider considerations come into play concerning how good or bad the resultant set of rules is as a whole, and statistical reliability may be less important than other considerations such as understandability or acceptance by the expert.

To summarize, the difficulties with χ^2 were the production of significantly larger trees than with IM; the extreme sensitivity to small, expected frequencies; and, particularly with the normalized value, the favouring of two-valued attributes, leading to very narrow trees with many levels. This led to a reconsideration of the IM—this too must follow some statistical distribution which would enable it to be used for pruning the tree—and other alternatives.

THE DISTRIBUTION OF THE INFORMATION MEASURE AND OTHER ALTERNATIVES

Reference to Kullback[6] showed a statistic for contingency tables based on information theory. Further, Sokal and Rohlf[7] discuss a statistic called the G-statistic, also based on information theory. The relationships between these statistics, Quinlan's IM, the χ^2 test statistic and the χ^2 distribution were investigated. The details are shown in the Appendix. The main results can be summarized as follows:

(i) Kullback's statistic and the G-statistic are identical and will henceforth be referred to as G. The distribution of G is a very close approximation to the χ^2 distribution.

(ii) G is proportional to the IM—in fact,

$$G = 2N \times \mathrm{IM},$$

where N is the total number of observations. This answers the question about the distribution of IM.

(iii) The χ^2 *statistic* itself only approximates the χ^2 *distribution* as Sokal and Rohlf[7] shows (p. 697). Moreover, Sokal and Rohlf shows that in the tails, the G statistic can be a better approximation than the χ^2 statistic.

(iv) G is approximately but not exactly equal to the χ^2 statistic. Hart[3] is mistaken, therefore, in suggesting that the χ^2 statistic is equal to N times Quinlan's IM (p. 119); rather, it is approximately equal to $2N$ times IM.

As a result of the above, and the fact that G is less sensitive to small frequencies, it was decided to use G in the algorithm rather than χ^2. It was further decided to use the calculated value without normalization. Strictly, this is incorrect as it ignores the degrees of freedom and means that pruning to a specified significance level will not be accurate. However, pragmatically the final rule tree as a whole is what matters most, and those with multi-valued attributes were judged better. As for significance, how should *the* correct level of significance (10%? 5%? 7.2%?) be judged anyway? Problems with small samples would be further improved by pruning the tree severely to ensure that there were reasonable numbers at each leaf.

RESULTS

Data on business studies students

The attributes used and a sample of the data were shown in Table 1. The entire tree, with no pruning, had 148 nodes. A pruned tree with a cut-off value of 8 is shown in Figure 1.

Clearly this pruned tree represents a tremendous simplification over the full one, but how sensible is it and how can we evaluate it in general and in comparison with trees based on other cut-off values?

The first attribute chosen is qualification, and it suggests that those either with BEC or with 'other' qualifications will get 2.2s, the 'other' being based on only three students, one of whom actually got a first. For those with A-levels, A-level score is chosen next with a test value of 9 points. Those with over 9 are predicted as 2.1. Those with less than 9 are then split on sex, females being predicted as 2.2, males being analyzed further, but in all cases bar one still being predicted as 2.2.

Some of the rules seem sensible—e.g. the rules concerning qualifications and A-level score—but this tree illustrates a number of general problems.

(i) A number of the leaves are still based on a very small number of observations, despite the heavy pruning—for example, the split into 2.2 or third with only three observations in each case—and this would clearly have little or no statistical validity. This happens for different reasons. First, because the statistic (whichever is used) is independent of the actual number of observations—it only measures the split between them—it tends to favour situations which can be reduced to only one class, even if there are very few cases. It would be preferable to favour situations which were less dichotomized but with more cases. Secondly, it may be that there are only a few examples of a particular attribute value—e.g. students with qualifications other than BEC or A-level.

```
qualification   8.84
   a  : A' level score   12.19
         < 9: sex  10.23
                 m  : O' maths   7.99
                        a : 2.2    1  0  3  0  0
                        b  : O' levels   8.55
                              <  8:2.2    0  0  14  0  0
                              ≥  8:2.2    0  3   8  2  0
                        c  : O' levels   2.30
                              <  9:2.2    0  3  18  2  0
                              ≥  9: O' levels    8.32
                                     <  10:third    0  1  0  2  0
                                     ≥  10:2.2      0  0  3  0  0
                        f  : null
                 f : 2.2    1  9  15  0  0
         ≥  9:2.1    0  7  2  1  0
   b  : 2.2    0  1  11  2  0
   o  : 2.2    1  0   2  0  0

   class      right      wrong      total
   -----      ---        -----      -----
   first       0           3          3
   2.1         7          17         24
   2.2        74           2         76
   third       2           7          9
   fail        0           0          0

   overall significance of the tree  58.42
   No. of classes 5      No. of leaves 10
```

FIG. 1. *Pruned rule-tree for degree data. The value after an attribute, e.g. qualification 8.84, is the G-statistic. The five numbers at each leaf are the number of students in each class at that point. The class chosen is that with the greatest number.*

Note that this occurs even with a reasonable total sample size. Using rule induction on small amounts of statistical data cannot possibly give results that can be relied on any more than any other statistical technique. For example, Brooks and Alty[8] induce rules concerning students' computing ability on the basis of 17 examples, most of the rules being based on only one student. The generalizations they make from this tiny, unrepresentative sample might easily gain more credence through the mystique of terms like expert systems and rule induction.

(ii) In some cases, an attribute is significant, but when the classes are determined, the class is the same whatever the attribute value. For example, with a 'B' at O-level maths, there is a further split on number of O-levels (8 or more), yet both cases yield a 2.2. This cannot be avoided during pruning but could be removed after.

(iii) With a pruned tree, it may turn out that no rules are generated for a particular class or classes. For instance, in this example there is no rule for a first. This reflects an inherent unpredictability in the data and the small number of instances.

(iv) How should the level of pruning be chosen? With insufficient pruning there will be many spurious rules. With too much pruning the rule tree will be over-general. The next section deals with this by defining measures of significance for a tree.

Measuring the significance of the results

The trees can be evaluated statistically in terms of both the significance of the tree as a whole and its predictive ability. Taking the first point, the observations at all the leaves can be taken together as one large contingency table, with a row for each leaf and a column for each class, since the various leaves split the data exhaustively and exclusively. The G-statistic can then be calculated for this table as a measure of the significance of the tree as a whole. In the above example, its value is 58.42. Table 2 shows the significance of trees generated from this data for different cut-off values.

These results show first that the trees can be extremely significant (at the 0.1% level), and secondly that the significance varies with the degree of pruning. With little pruning, the large number of leaves is based essentially on chance, while with a high level of pruning, the ability to classify the data is small. In this example, the most significant trees are for cut-off values of 6 or 8.

It is also possible to assess the significance of the number of correct predictions which the tree makes. In this example, 83 out of 112 were correct. How much better than chance is this? Each prediction can either be true or false, and the predictions are independent, so the number of true ones should follow the binomial distribution. What will be the probability of success by chance? That depends on how much chance knows to begin with. If we assume no knowledge at all, then each class would be equally likely and the probability of one successful prediction would be 1/4 (ignoring the fail class with no examples). However, the classes are not at all equally likely, and it seems a fairer test (and more consistent with the contingency table) to at least assume knowledge of the proportion of each class. On this basis, i.e. assuming that the chance predictions are made in the same proportions as the actual data, then

$$\begin{aligned}&\text{Prob (successfully predicting a 3rd)} =\\ &\text{Prob (student being a 3rd)} \times \text{Prob (prediction being a 3rd)}\\ &= P_3 \times P_3\\ &= (P_3)^2.\end{aligned}$$

Similarly, for the other classes, so

$$\begin{aligned}\text{Prob (successful prediction)} &= \Sigma(P_i)^2 \qquad [i = 3,\ 2.2,\ 2.1,\ 1]\\ &= (3/112)^2 + (24/112)^2 + (76/112)^2 + (9/112)^2\\ &= 0.51.\end{aligned}$$

So on this assumption, the number of successes will follow a binomial, with $n = 112$, $p = 0.51$. This is approximately normally distributed, with $\mu = 57.1$, $\sigma = 5.3$. Table 3 details the results for various cut-off levels.

There are two points to note. First, each level is strongly significant, but this takes no account of the number of leaves in each case. Secondly, to put this in perspective, if the only rule were that everyone got a 2.2, then 76 predictions would be correct.

TABLE 2. *Significance of degree data rules*

Cut-off level	No. of leaves	Degrees of freedom	G-statistic	Significance levels 5%	1%	0.1%
0	61	180	196	212	227	244
2	60	177	194	209	224	241
4	28	81	113*	103	114	126
6	11	30	65***	43.8	50.9	59.7
8	10	27	58***	40.1	47.0	55.5
10	5	12	31**	21.0	26.2	32.9
12	4	9	21*	16.9	21.7	27.9

TABLE 3. *Significance of degree predictions 1984*

Cut-off level	112 students No. of leaves	$\mu = 57.1$ $\sigma = 5.3$ Successful predictions	z-score
2	60	110	9.9
6	11	84	5.0
8	10	83	4.8
12	4	81	4.4

Table 4. *Significance of degree predictions 1985*

Cut-off level	67 students $\mu = 29.7$ $\sigma = 4.1$ Successful predictions	z-score
2	27	−0.8
6	35	1.2
8	35	1.2

Testing the rule tree

Clearly, it is best to test predictive systems on a separate set of data, and this was done using the 1985 students' results. There were 67 students—2 firsts, 23 2.1s, 38 2.2s and 4 thirds.

The probability of a successful chance prediction is therefore 0.44 (calculated as above). Using the above rule trees on this data produced the results shown in Table 4.

These results are obviously disappointing as they are not statistically significant. The set of rules which was best at explaining the '83–'84 data has not managed to produce more correct predictions than would have been produced by chance. This reflects the lack of a stable pattern in this very noisy data. One of the main rules was that those with 9 or more A-level points should get a 2.1. On inspecting the '85 data, there were *no* students with 9 or more points, so the rule was useless for that data.

When a rule tree was developed for the 1985 data, it was quite different from the 1984 one. Even the first attribute chosen changed from qualifications to O-level maths. The problem is not peculiar to this approach but is implicit in any inductive method, e.g. multiple regression, where it is common that good explanatory systems are not good predictors. The question is whether rule induction is better or worse than other methods.

In summary, the use of the G-statistic and tree-pruning has been demonstrated, and measures for the significance of the resulting tree and the number of successful predictions have been developed. That the predictions have not been successful in this case is largely a function of the data, but it would be interesting to known if a more traditional technique, such as multiple regression, would have done better.

CONCLUSION

This paper has developed and explored extensions to ID3 to enable it to deal with stochastic data, following on from suggestions by Hart. It has been shown that:

(i) Statistical data can be analysed using χ^2 instead of the information measure. However, it has been found better to use the G-statistic—another measure based on information theory. The relations between these various measures are set out.

(ii) It is not adequate to cut off a tree when a branch becomes insignificant during its development, as both Hart and Quinlan suggest, in case it should be significant further on. It must instead be developed fully and then pruned back. Even then, simply pruning to some specified significance level generates trees which are not totally satisfactory.

(iii) Measures of significance both for a rule tree and for predictions made from a tree have been developed, and these enable different trees with different levels of pruning to be compared.

There remain, however, difficulties in using rule induction. The problems specifically concerned with using statistical data have been discussed in the last section. The most important of these is that of statistical reliability, as some of the rules may be based on a very small number of cases, even when there is a large amount of data in total. If there is only a small amount of data, there is no more value in using rule induction than any other statistical technique. Problems concerning both statistical and determinate data include:

—discovering rules involving functional relationships between the attributes;
—the difficulty of understanding and interpreting rule trees in their standard form;
—the question of an 'optimum' rule tree—would this be the smallest, the most efficient,[9] the most easily understood, the expert's choice, or what?
—the selection and development of suitable attributes in the first place.

Thus the potential of rule induction to deal with statistical data has been demonstrated, and the degree of success warrants further investigation. This does not mean, however, that it is a panacea for knowledge acquisition—it will always be limited by the quality and amount of data available; it still requires a good deal of expert time to supply the attributes and examples, and to validate the rule tree; and it will only really be suitable for fairly straightforward, well-defined domains.

AREAS FOR FURTHER RESEARCH

(i) More tests on other sets of data, preferably with differing degrees of noise. Also, comparisons with other statistical techniques such as multiple regression, discriminant analysis and categorical/classification methods.[10,11]

(ii) Methods for more intelligently pruning the tree, or post-processing it, to avoid redundant tests, taking into account the relative strength/reliability of the various rules, and displaying them in a more appropriate form.

Acknowledgements—I would like to thank Mike Parker and anonymous referees for help and comment.

APPENDIX

Relations Between the Information Measures and χ^2

Given that we are considering the association between one attribute and the class, the data (or subset thereof) can be set out in a contingency table:

		Class				
		C_1	C_2		C_c	Totals
	A_1	x_{11}	x_{12}			$x_{1.}$
Attribute values	A_2	x_{21}	x_{22}			$x_{2.}$
				x_{ij}		
	A_r					$x_{r.}$
	Totals	$x_{.1}$	$x_{.2}$		$x_{.c}$	N

where x_{ij} is the number of examples in class j with attribute value i.

Quinlan's Information Measure

Using notation from Quinlan,[2] the information content of the class totals as a whole is

$$\begin{aligned} M(C) &= -\frac{x_{.1}}{N}\log\frac{x_{.1}}{N} - \frac{x_{.2}}{N}\log\frac{x_{.2}}{N} - \dots \\ &= -\frac{1}{N}\,(\Sigma x_{.j}\log x_{.j} - N\log N). \end{aligned}$$

The information content for the row A_1 is

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

and similarly for the other values of A.

Taking an average of these A values weighted by the frequency of occurrence of each (the row totals) gives

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

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

$$I = M(C) - B(C|A)$$
$$= \frac{1}{N}(\Sigma\Sigma x_{ij} \log x_{ij} - \Sigma x_{i.} \log x_{i.} - \Sigma x_{.j} \log x_{.j} + N \log N). \qquad (1)$$

Kullback's Information Measure

Using notation from Kullback,[6] Table 2.1, p. 158,

$$H(A, C) = 2\Sigma\Sigma x_{ij} \log \frac{N x_{ij}}{x_{i.} x_{.j}}$$
$$= 2(\Sigma\Sigma x_{ij} \log x_{ij} - \Sigma x_{i.} \log x_{i.} - \Sigma x_{.j} \log x_{.j} + N \log N)$$
$$= 2NI \quad \text{substituting from (1).}$$

Thus Quinlan's measure is

$$\frac{H(A, C)}{2N}.$$

The G-Statistic

From Sokal and Rohlf[7], Box 17.6,

$$G = 2(\Sigma\Sigma x_{ij} \log x_{ij} - \Sigma x_{i.} \log x_{i.} - \Sigma x_{.j} \log x_{.j} + N \log N).$$

It can be seen that the G-statistic is the same as Kullback's information measure $H(A, C)$.

The χ^2-Test

This is the most common measure for contingency tables:

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

where

$$E_{ij} = \frac{x_{i.} x_{.j}}{N}.$$

Kullback[6] (p. 159) shows that his measure, $H(A, C)$, is approximately but not exactly equal to this.

REFERENCES

1. J. Mingers (1986) Expert systems—experiments with rule induction. *J. Opl Res. Soc.* **37,** 1031–1037.
2. J. R. Quinlan (1984) Learning efficient classification procedures and their application to chess end games. In *Machine Learning: An Artificial Intelligence Approach* (R. Michalski, J. Carbonell and T. Mitchell, Eds). Springer, New York.
3. A. Hart (1984) Experience in the use of an inductive system in knowledge engineering. In *Research and Developments in Expert Systems* (M. Bramer, Ed.). Cambridge University Press.
4. E. Hunt, J. Marin and P. Stone (1966) *Experiments in Induction.* Academic Press, New York.
5. J. R. Quinlan (1983) Learning from noisy data. *Proceedings of the International Workshop on Machine Learning.* Department of Computer Science, Illinois University.
6. S. Kullback (1967) *Information Theory and Statistics.* Dover, New York.
7. R. Sokal and F. Rohlf (1981) *Biometry.* Freeman, San Francisco.
8. A. Brooks and J. Alty (1985) The use of rule induction: a knowledge acquisition technique for expert systems to interpret HCI experiments. In *People and Computers: Designing the Interface* (P. Johnson and S. Cook, Eds) Cambridge University Press.
9. R. O.'Keefe (1986) Simulation and expert systems—a taxonomy and some examples. *Simulation* **46,** 10–16.
10. G. Kass (1980) An exploratory technique for investigating large quantities of categorical data. *Appl. Statist.* **29,** 119–127.
11. A. Gordon (1981) *Classification: Methods for the Exploratory Analysis of Multivariate Data.* Chapman & Hall, London.