J. Opl Res. Soc. Vol. 38, No. 4, pp. 347-351, 1987
Printed in Great Britain. All rights reserved

0160-5682/87 $3.00 + 0.00

Rule Induction with Statistical Data—A Comparison with Multiple Regression

JOHN MINGERS
Polytechnic of the Southbank

Rule induction is a method of automatically developing rules from sets of examples. Quinlan's ID3 algorithm, which was developed for determinate data, has been extended to deal with statistical data. This paper reports on an experimental comparison with multiple regression.

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

INTRODUCTION

Expert systems require expert knowledge to be incorporated within them, and various methods of automating this process are being developed.[1] One approach is that of rule induction from examples, as developed by Quinlan[2] in the ID3 algorithm. ID3 was designed to work with deterministic data but, based on Hart,[3] Mingers has developed the algorithm to deal with noisy statistical data.[4,5] This paper reports on a test of the modified algorithm on data which had been used with regression, thus allowing a tentative comparison between the two.

RULE INDUCTION AND MULTIPLE REGRESSION

Rule induction and regression are similar in that they both use a set of data consisting of a number of examples or cases, each of which consists of a number of observations. Both methods then use induction to determine a relationship between these observations which can be used for predicting one of the variables. Before discussing the differences between them, the question of terminology needs to be resolved, as different terms tend to be used in the two different domains:

Regression	Rule induction
Cases	Examples
Independent variables	Attributes
Dependent variable	Class

With induction, the various examples have attributes and can be classified into classes. With regression, cases consist of the values of independent variables and a dependent variable. This paper will generally use the regression terms.

The differences between the two methods are quite significant. Regression assumes that the data is continuous from an interval scale of measurement (although logical relationships can be modelled using dummy variables). The relationship between the independents and the dependent is taken to be a functional one—in fact, a linear function—and in order to conduct significance tests on the results, the variables are assumed to be independently normally distributed. In contrast, rule induction is essentially classificatory, since the dependent variable is only nominal—i.e. the name of a class. The independent variables may be nominal or interval, and the relationships which are induced are logical rather than functional. There is no requirement that the variables be independent, nor are any assumptions made about their distributions.

It might seem as though the two are applicable in totally different situations, but there is some degree of substitutability. For example, if the regression assumptions are not justified and the dependent variables could be grouped into a number of classes, then induction could be used on a regression problem. Alternatively, if all the attributes were interval variables, regression could be used on a classification problem. In this instance, a classification problem—predicting football results—had been tackled using regression before rule induction was available.

REGRESSION AND FOOTBALL DATA

The problem domain was the fairly intractable one of predicting the football results. The dependent variable used was the net score in a match, and regression was used to explain this in terms of factors such as difference in long-term form, difference in recent form, home and away form, propensity for drawn games etc. Numerous measures and combinations of these explanatory variables had been tested throughout the season. The regressions were generally significant overall, although few individual variables were, because of the high degree of noise in the data. The predicted results were generally sensible, though not startlingly accurate.

To test the rule induction, a data set of 164 matches was used with the collection of variables which had been found best in regression. This clearly gives a bias towards regression. Only marginal changes to the data were needed (see below).

The following variables were used:

SCORE	The dependent variable: the net score in a match for regression. For induction, this was changed to HOME, AWAY or DRAW, thus making it nominal.
POSITION	The difference between the average number of points of the two teams. For induction, this was scaled to be in the range −20 to +20 as the algorithm only deals with integers.
FORM	The difference between the teams on a measure of recent form: recent results calculated as a percentage and exponentially smoothed.
HOME-F	A measure of home form for the home team relative to its overall form.
AWAY-F	A measure of the away form of the away team relative to its overall form.

Regression results for this set of data are shown in Table 1. As can be seen, the best regression uses only two of the variables but is strongly significant overall.

RULE INDUCTION AND FOOTBALL DATA

The data was run through the enhanced version of ID3. This version (ID5) uses the G-statistic (which follows approximately the chi-squared distribution) instead of Quinlan's information measure, and after generating a complete tree, prunes it back to some specified level of G. For further details, see Mingers.[5]

A rule tree generated from this data, and pruned using a cut-off value of 10, is shown in Figure 1.

The first attribute chosen is position. A value of −1 is selected, and any matches less than this are predicted to be aways. This is quite reasonable, as the negative value means that the away team has a higher average number of points, but note that there are still 15 homes and 22 draws at this leaf, so the rule is not particularly reliable. Further classification of these cases has not turned out to be significant, and so has been pruned away. Other problems typical of these statistically-based trees are the small numbers of observations at certain leaves (despite the high degree of pruning and large number of cases) and the lack of reliable rules for certain classes—e.g. a draw.

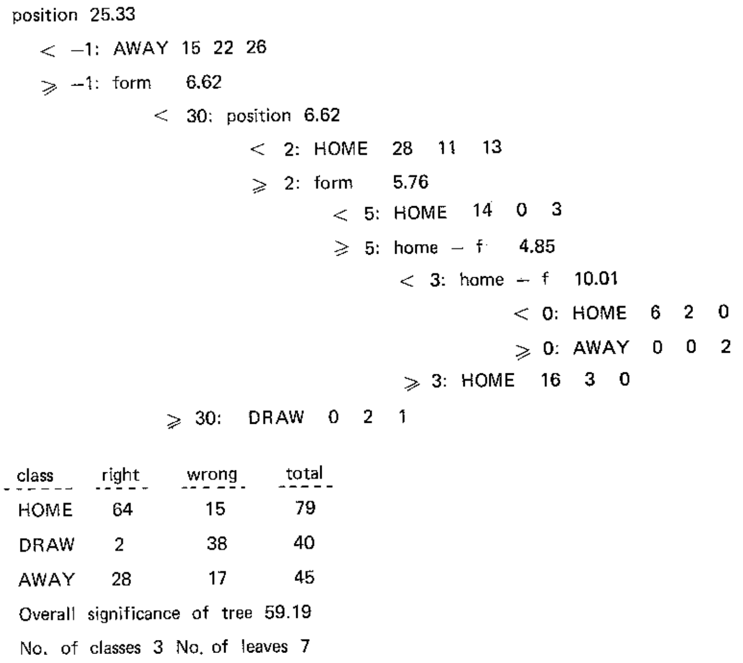
The cut-off value of 10 has heavily pruned the tree, reducing it from 75 leaves down to only 7. Had the value been higher, only the first branch would have remained. Nevertheless, the overall tree is strongly significant (<0.001). The effect of differing cut-off values on the significance of the resulting tree can be seen in Table 2.

This shows the G-statistic for various cut-off values, together with appropriate critical values of chi-squared. As can be seen, the overall rule-tree is highly significant at all the cut-off levels. It appears to be more significant for cut-off values of 6 and 8.

TABLE 1. *Regression coefficients for football data*

Const.	Position	FORM	HOME-F	AWAY-F	R	d.f.
0.70	0.21	−0.031	−0.0021	−0.0059	0.36	159
	(4.8)	(2.1)	(0.12)	(0.37)		
0.60	0.21	−0.030			0.36	161
	(4.3)	(4.8)				

()—t-value.
R strongly significant (<0.001).



```
position 25.33
  < -1: AWAY 15 22 26
  ≥ -1: form    6.62
             < 30: position 6.62
                        < 2: HOME   28   11   13
                        ≥ 2: form     5.76
                                 < 5: HOME   14   0   3
                                 ≥ 5: home - f    4.85
                                         < 3: home - f    10.01
                                                      < 0: HOME   6   2   0
                                                      ≥ 0: AWAY   0   0   2
                                         ≥ 3: HOME   16   3   0
             ≥ 30:   DRAW   0   2   1

class    right    wrong    total
------   -----    -----    -----
HOME      64       15       79
DRAW       2       38       40
AWAY      28       17       45
Overall significance of tree 59.19
No. of classes 3 No. of leaves 7
```

FIG. 1. *Rule tree for football data. Notes*: (*i*) *The three numbers at each leaf are the number of examples in each class at that leaf.* (*ii*) *The number at each branch is the value of the G-statistic at that branch.* (*iii*) *The overall significance is a G-statistic for the tree as a whole, the significance of which can be found from the chi-squared distribution.*

TABLE 2. *Overall significance of the rule tree*

				Critical values	
Cut-off level	No. of leaves	d.f.	*G*-stat	1%	0.1%
2	75	148	344***	191	207
4	49	96	268***	131	145
6	18	34	134***	56	65
8	10	18	84***	35	42
10	7	12	59***	26	33

The significance of the number of correct predictions can also be tested. As shown in Mingers,[5] if it is assumed that a chance mechanism makes predictions based only on the proportion of the various classes in the data, then the number of successful predictions will follow the binomial distribution. Based on the proportion of homes, draws and aways in the sample, the probability of a correct prediction by chance is 0.366, and so the number of correct predictions will be binomially distributed with a mean of 60.1 and standard deviation of 6.2. The actual number of correct predictions can be tested to see if it is significantly better than chance. The results (using the normal approximation) are shown in Table 3.

These also were signiffcant at all levels of pruning, although it should be noted that this test does not take into account the number of leaves in the tree, and it might be expected that many of the lower-level rules would be due to chance and would therefore be of little use for predicting other sets of data. Even with the highest degree of pruning, there were 94 successful predictions, although not many of these were draws however!

TABLE 3. *Significance of football predictions*

	164 matches	$\mu = 60.1$ $\sigma = 6.2$	
Cut-off level	No. of leaves	Successful predictions	z-score
4	49	143	13.4
6	18	111	8.2
8	10	100	6.5
10	7	94	5.5

TABLE 4. *Significance of football predictions for new data*

	182 matches	$\mu = 70.4$ $\sigma = 6.6$	
Cut-off level	No. of leaves	Successful predictions	z-score
6	18	89	2.8
8	10	93	3.4
10	7	97	4.0

Applying the tree to another set of data

Generally, a predictive system should be validated on a different set of data from that with which it was developed. This was done using a further set of 182 matches, and the results are shown in Table 4 in terms of the significance of the number of correct predictions.

As can be seen, the number of correct predictions was significant for different cut-off levels, and was more significant for the more highly pruned trees. This confirms the fact that only the most general rules are valid across different sets of data. That any of these are significant is surprising given the extreme noisiness of the data.

COMPARING REGRESSION AND RULE INDUCTION

It is difficult to compare directly the significance of the rule trees with that of the regression equation, but since the aim is to predict football results, it would be useful to compare the number of correct predictions for the first set of data with which both methods were developed. Predicted results using the regression equation are not immediately available, since the equation predicts the actual net score, and this needs to be converted to home, draw or away. To do this it is necessary to decide on two limiting values of the net score—one to demarcate homes from draws, and the other, draws from aways. For instance:

$$\text{away} < -0.5 \leqslant \text{draw} < 0.5 \leqslant \text{home}.$$

There is no *a priori* best value for these limits, and to provide as severe a test as possible for the rule tree, the optimum values for this particular data were chosen. These were found by looking at all possible combinations of values in steps of 0.25 of a goal to see which gave the most correct forecasts. The best results were as shown in Table 5.

In practically using the regression to make predictions, the limits would have to be chosen beforehand, thus lowering the level of accuracy. In this ideal case, even the best that the regression can do (90 correct) is marginally worse than the ID5 results, even at the greatest level of pruning (94 correct). Clearly, this single result should be treated with caution, but it is significant that the experiment was biased towards the regression in that the variables were those developed for regression rather than induction, and that the regression predictions were optimized for regression.

CONCLUSION

Regression and rule induction have both been used on the same set of (very noisy) data, and rule induction has been shown to have similar predictive power to regression, even though the test was biased in favour of regression. More generally, in comparison with regression, induction seems to have certain advantages. It does not need the same assumptions concerning linear relationships and normality, and is generally much less restrictive in the type of data it will accept. Its output is also in a form which is much more intelligible and useful in many situations.

TABLE 5. *Best multiple-regression predictions*

Net score	Predicted result	Number correct
>0	Home	75
$0 > -0.25$	Draw	5
$\leqslant -0.25$	Away	10
		90

The two methods are different, however, in that rule induction is for classificatory data and regression is for measured data, and they are not therefore direct substitutes in many circumstances. A statistical method more directly comparable is discriminant analysis, and work could usefully be done comparing these.

REFERENCES

1. R. Michalski, J. Carbonell and T. Mitchell (Eds) (1984) *Machine Learning: An Artificial Intelligence Approach.* Springer, New York.
2. J. R. Quinlan (1979) Discovering rules from large collections of examples: a case study. In *Expert Systems in the Micro Electronic Age* (D. Michie, Ed.) Edinburgh University Press.
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. J. Mingers (1986) Expert systems—experiments with rule induction. *J. Opl Res. Soc.* **37,** 1031–1037.
5. J. Mingers (1986) Expert systems—rule induction with statistical data. *J. Opl Res. Soc.* **38,** 39–47.
6. D. Hand (1981) *Discrimination and Classification.* Wiley, New York.