

INDUCTION OF SETS OF RULES FROM ANIMAL DISTRIBUTION DATA: A ROBUST AND INFORMATIVE METHOD OF DATA ANALYSIS

David R.B. Stockwell and Ian R. Noble
Ecosystem Dynamics Group, Research School of Biological Sciences,
Australian National University, Canberra, ACT 2601

1. INTRODUCTION

We are interested in developing a method of modelling using sets of rules. There were two main reasons for choosing sets of rules as a representation for analysing and modelling biotic response to the environment. Firstly, ecologists frequently think about patterns in the environment as rules, e.g. 'If community x is present then species y is likely to be present.' This may explain the great interest in rule-based expert systems in ecological management [1]. Secondly, multivariate techniques either often perform poorly or are not applicable to animal distribution data. For example, the derived components of Principle Components Analysis (PCA) may explain a small amount of the variance, providing no reduction in the intrinsic dimensionality of the data set [2]. If categorical data are present, e.g. vegetation types, multivariate methods are clearly an inappropriate form of data analysis.

Until recently, modellers have developed rules by the slow, error-prone and subjective elicitation of knowledge from experts. Where data is available, decision tree induction methods can automate the development of rules, for prediction of animal and vegetation distributions [3],[4]. Despite advantages over multivariate methods for prediction (such as lack of assumptions about the frequency distribution of the data and relative insensitivity to outlying values), some methods are not very robust, producing many alternative trees under data perturbations [3]. In addition, decision trees enforce a hierarchical structure which may be an inappropriate form of model in some cases.

We first describe a system for producing rule sets, called GARP (Genetic Algorithm for Rule Set Production). Modelling systems should at least support prediction, exploration, and explanation, so we describe the application of sets of rules to these tasks. In addition, rule sets appear to have two unique advantages: rule sets are robust (i.e. stable under data perturbations) and informative (i.e. allow the construction of complex representations from simple components without global assumptions). We explain why by comparison of the structure and assumptions of rule sets to decision trees and multivariate methods.

2. MODELLING VIA RULE SETS

A rule has the basic form: 'if something is true then something necessarily follows.' The 'if' part of the rule is called the *precondition*, the 'then' part the *conclusion*. The preconditions of rules in GARP are simple *conjunctive* expressions: e.g.

$$V_1 = v_1 \ \& \ V_2 > v_2 \ \& \ \dots \ \& \ V_m = \{v_{m1}, v_{m2}\}$$

where v_1, v_2, \dots, v_m are values of the variables V_1, V_2, \dots, V_m .

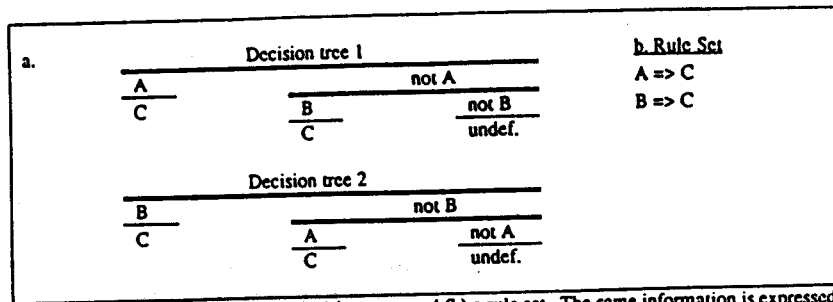


FIGURE 1. A comparison of (a) decision trees and (b) a rule set. The same information is expressed two different ways in the decision trees.

The variables 1 to m are a subset of the total number of variables, i.e. they are not repeated. The precondition selects a subset of the data set. For example, given a data set $\{ \langle 1,1,1 \rangle, \langle 0,1,1 \rangle, \langle 1,0,1 \rangle \}$ then the selector $V_1=1$ selects the data set $\{ \langle 1,1,1 \rangle, \langle 1,0,1 \rangle \}$. The conclusion is an assignment of a classification value to the selected subset. This assignment results in two additional pieces of information, the distribution of incorrect and correct assignments, and a measure of the quality of the rule. Rules are ideal for representing *higher-order* interactions, i.e. correlations between particular values of different variables that are not found throughout all values of the variables.

The following is an example of a typical rule developed by GARP. The data are from observations from water bodies in disused quarry pits in Gloucestershire, England. The variables include observations of numbers of birds, water parameters and habitat characteristics.

$$0 \leq \text{age} = [20,30), \text{aq} = [10,20) \quad (24,48) \quad 2.03$$

This rule means 'If age of pit is between twenty and thirty years and the number of aquatic species present is between ten and twenty, then juvenile birds are absent.' This rule gets 24 examples incorrect and 48 correct. The significance of the rule is 2.03 (i.e. there is <1% probability of being formed through a random event.)

A *rule set* is an unordered list of rules. The set of rules constitutes the model of the system M; a single rule a partial model M_i . To show how rule sets can represent information concisely, compare a simple rule set with its equivalent formulations as decision trees (Figure 1).

In this form of representing decision trees, the nodes of the decision tree are indicated by a double line, and the single lines indicate a class assignment. Decision trees one and two are alternative ways of representing the same information contained in the rule set. The decision trees however place priority on the first selector. In decision tree 1, A has priority, and in decision tree two, B has priority. In comparison, rule sets place equal priority on each rule. Thus decision trees contain information on the order to apply the rules, information that may be irrelevant to modelling the system. We explain later how this many to one relationship contributes to the instability of decision trees found in [3].

3. INDUCTION OF RULE SETS

The space of possible rules is invariably too large to exhaustively search all possible rules. Genetic algorithms have been found useful. The idea is to 'evolve' models by modification with specially defined operators. Starting with an initial set of inferior rules, operators modify the rules in

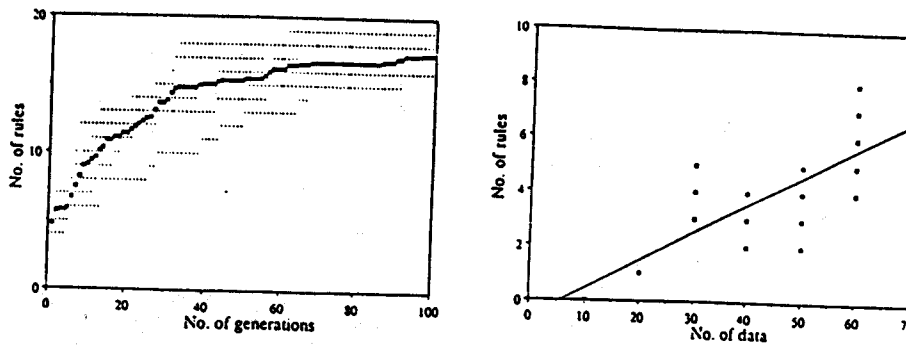


FIGURE 2. The number of rules discovered by GARP with respect to (a) number of generations and (b) number of data. While there is little gain in running the algorithm longer, more rules would be gained by increasing the amount of data.

ways that may or may not lead to an increase in quality. After each modification the quality of the rule is tested and a size limited set of best rules so far maintained.

Typically the algorithm will produce better models (i.e. sets of rules covering more examples) given longer run times and more training data. On the juvenile waterbird data set, the number of rules found by GARP with significance greater than 90% with respect to number of generations and number of data is shown in Figure 2. The number of rules discovered can be seen to plateau around 100 generations, while the relationship of rules to data is still increasing.

The early genetic algorithms used lists of simple binary classifiers as selectors (i.e. $V_i=0$ or 1). Heuristic operators, such as mutation and cross-over, derived from the genetic evolution of DNA, were used for searching the space of possible binary classifiers [5]. Successful learning can be achieved using heuristic operators defined on more complex representations [6]. The operators in the present version of GARP are: the *random* operator, generating a rule with a random number of conjunctions and values e.g. (null) $\rightarrow V_1=\{1,2\}, V_2=1$, the *mutate* operator, changing the value of a variable in a rule at random. e.g. $V_1=\{1,2\}, V_2=1 \rightarrow V_1=\{0,1\}, V_2=1$, and the *concatenate* operator, concatenating two existing rules e.g. $(V_1=\{1,2\}), (V_2=1) \rightarrow V_1=\{1,2\}, V_2=1$.

The steps in the algorithm are: (1) Test all rules of selector length one and place a preset number of best rules into the current rule set. (2) Apply each of the defined operators to randomly selected rules in the current rule set. (3) Place those modified rules that have greater than preset significance into the current rule set. (4) Order the current rule set, display the current rule set, and eliminate the least useful rules if the number of rules exceeds a maximum limited by available memory. (5) Repeat steps 2 to 4 a set number of times. (6) Print out the current rule set.

Genetic algorithms are believed to be most useful in applications where the modeller has little reliable background knowledge [7]. Empirical studies report that genetic algorithms are most useful for finding rules in large complex spaces, i.e. noisy, high-dimensional, and discontinuous with many local optima [8]. Even though genetic algorithms are computationally expensive relative to methods that employ more background information, they also have an inherent parallelism that can be exploited by massively parallel computational architectures [9].

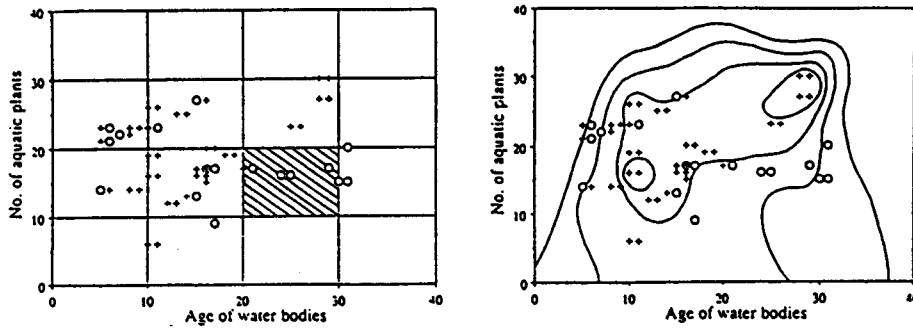


FIGURE 3. (a) A rule identified by GARP can help in finding 'interesting' regions of the variable space, shown hatched above. (b) The points can be contoured to help formulation of functions describing the data.

4. PREDICTION WITH RULE SETS

In prediction, a choice is made between possible classes, given certain information. The best class to predict is the one with the highest expected probability. Hence, a good predictive decision rule is to choose the class C_i given by the rule M_j that maximises $E(P(C_i|M_j))$. Thus, $P(C_i|M_j)$ calculated from given data is a measure of the usefulness of the rule M_j .

However, estimation of $E(P(C_i|M_j))$ from $P(C_i|M_j)$ is misleading when there are low numbers of samples, spurious associations, or high prior probabilities of the class C_i . Instead of $P(C_i|M_j)$, GARP estimates the statistical relevance of a rule, i.e. the increase in probability of a class given M_j relative to the prior probability of that class (i.e. $P(C_i|M_j) > P(C_i)$). A satisfactory measure of statistical relevance with a standard normal distribution is:

$$z = (n_1 - np_1) / \sqrt{np_1(1-p_1)} > \alpha \text{ and } n_1 > 4. \quad (1)$$

where n_1 is the number of correct classifications, p_1 is the expected fraction of correct classifications, and n is the total number of data [10].

Note that we are only concerned with rules that give a positive value for z as these ensure that $P(C_i|M_j) > P(C_i)$. Rules that significantly decrease the class frequency may be expressed as 'if A then not C.' However, introduction of negation introduces difficulties not dealt with in the existing system.

5. EXPLORATION OR HYPOTHESIS GENERATION

Rules produced by GARP are useful for exploring data sets. Rules identify 'interesting' regions in the space defined by the variables. A region is *interesting* when the distribution of classes is significantly different from the overall distribution of classes, as shown in the hatched region of Figure 3a. Having identified combinations of variables with interesting regions we could contour the surface and hypothesize a new classes of functions to fit the data (Figure 3b). Alternatively a probability surface could be found using the algorithm of Bayes and Mackey [11]. Note that it is less presumptuous to perform rule induction prior to fitting the probability surface as rule sets can describe parts of the regions of a probability surface without the assumption of monotonicity or restriction to analysis in one or two dimensions.

6. EXPLANATION WITH RULE SETS

6.1 Robustness of rule sets

A *robust* modelling system produces a similar model in repeated, identical, though independent situations, capturing the idea of a theory confirmed by repeated independent experiments. Robustness can be estimated using measures of the stability of the structure of the model under random perturbations of the data, as introduced by resampling regimes. In decision tree induction systems, the initial choice of the root variable is the major determinant of the tree produced, as shown in Figure 1. If random perturbations affect the choice of root variable, then different, though logically equivalent trees may be produced. In rule sets however perturbations act on single rules. Hence the rule set undergoes partial changes, not complete restructuring making rule sets more robust than decision trees.

6.2 Construction of structures with rules, or informativeness

Often little is known about the response of an animal to its environment. With little background knowledge, we can validly make very few, or only weak assumptions. Multivariate methods make the strong assumption that the class of functions generating the observed data is *global* i.e. defined on the entire range of the variables. Often function characteristics such as linearity and monotonicity are assumed. Decision trees incorporate assumptions about the priority of the variables, and most algorithms are biased to produce shorter trees, and nodes that partition the data set in particular types of ways. In contrast, rules are locally defined and make no such global assumptions.

In rule sets, each of the individual components is a significant hypothesis. Hence we can use the rules to deduce the existence of more complex structures. Rules can potentially be formed into networks, functions or decision trees on the evidence that a number of rules suggest global patterns in the data. This approach to modelling allows construction of global structures from local features, without the global assumptions. In contrast global structures cannot be deduced from multivariate models, as the global structures are assumed prior to the analysis. It is the ability to deduce new information in flexible ways that constitutes *informativeness*.

Global information can be incorporated after development of the rule set. For example if the application suggested we should assign a priority to a single variable (e.g. geology in [4]), and that variable was present in a number of rules, we could organise the rule set into a decision tree. If the purpose of the rules is prediction only we can remove those rules that are specializations of existing rules without affecting predictive accuracy. Figure 4 illustrates these processes on a set of rules.

7. OTHER POSSIBILITIES OF RULE SETS

Transformation of the above data using PCA led to a large number of vectors explaining significant amounts of variance providing little reduction in the dimensionality of the system. GARP on the other hand produced a large number of rules with multiple preconditions, suggesting higher-order interactions are common. This initial investigation suggests a relationship between PCA and rule set induction that could be exploited in a hybrid system - multivariate models to describe global regularities and rules to represent the higher-order interactions.

The rule sets developed from GARP could be used in large-scale ecological simulations. Rules are a natural representation of changes in the states of variables (e.g. if $V_1=1$ at t_1 then $V_1=0$ at t_2). There

Juveniles				Rules
age={0,10}	age={10,20}	age={20,30}	age>30	1<-age={10,20}&C>300
undef.	C>300	aq={10,20}	0	0<-age={20,30}&aq={10,20}
	1	0		0<-age>30
				0<- age>30&area={0,10}

FIGURE 4. Construction of a decision tree from a rule set. The fourth rule is a specialisation of the third and is not needed for prediction.

is no impediment to implementation of rule sets on a variety of parallel computing architectures given appropriate algorithms [12].

8. CONCLUSIONS

Induction of rules sets is a modelling system more robust than decision tree induction and more informative than multi-variate methods. Rule sets can be produced efficiently from data with many variables using a genetic algorithm such as GARP. Rule sets may be used for prediction, exploration or explanation, the major requirements of a useful and valid modelling system. This method of analysis may be particularly suited to analysis of the response of animals to the environment, as preliminary investigations of waterbirds found a number of higher-order interactions. Work is proceeding on automating the construction of rules into more complex tree and network structures.

9. REFERENCES

- [1] Noble, L.R., The role of expert systems in vegetation science, *Vegetatio*, 69, 115-121, 1987.
- [2] Recher, H.F., Kavanagh, R.P., Shields, J.M., and P. Lind, Ecological association of habitats and bird species during the breeding season in southeastern New South Wales, *Australian Journal of Ecology*, 16:3, 337-352, 1991.
- [3] Stockwell, D.R.B., Davey, S.M., Davis, J.R. and L.R. Noble, Using induction of decision trees to predict greater glider density, *AI Applications in Natural Resources Management*, 4:4, 33-43, 1990.
- [4] Moore, D.M., Lees, B.G., and S.M. Davey, A new method for predicting vegetation distributions using decision tree analysis in a geographic information system, *Environmental Management*, 15:1, 59-71, 1991.
- [5] Holland, J.H., Holyoak, K.J., Nisbett, R.E., and P.R. Thagard, *Induction: processes of inference, learning and discovery*, MIT Press, Cambridge, Massachusetts, 1986.
- [6] Greffentette, J.J., Ramsey, C.L., and A.C. Schultz, Learning sequential decision rules using simulation models and competition, *Machine Learning*, 5, 355-381, 1990.
- [7] DeJong, K., Learning with genetic algorithms: an overview, *Machine Learning*, 3, 121-138, 1988.
- [8] Fitzpatrick, J.M., and J.J. Grefenstette, Genetic algorithms in noisy environments, *Machine Learning*, 3, 101-120, 1988.
- [9] Robertson, G.G., and R.L. Riolo, A tale of two classifier systems, *Machine Learning*, 3, 139-159, 1988.
- [10] Mendenhall, W., Scheaffer, R.L., and D.D. Wackerly, *Mathematical Statistics with Applications*, Wadsworth Inc., Belmont, California, pp549, 1981.
- [11] Bayes, A.J., and B.G. Mackey, (in press), Algorithms for monotonic functions and their application to ecological studies in vegetation science. *Ecological Modelling*.
- [12] Stockwell, D.R.B. and D.G. Green, Parallel computing in ecological simulation, *Mathematics and Computers in Simulation*, 32, 249-254, 1990.