Machine Learning and the Problem of Prediction and Explanation in Ecological Modelling

David Stockwell
Statement of Originality

The work presented in this thesis is my own. Contributions by co-authors to multiple author manuscripts are stated in the introduction to the thesis. Contributions by others are recorded in the acknowledgements at the end of each manuscript.

David R.B. Stockwell

Ecosystem Dynamics Group
Research School of Biological Sciences
Australian National University
Canberra, A.C.T. 2601
Acknowledgments

I would like to thank my supervisors David Green, Ian Noble, and Bob McKay and my advisers Stuart Davey, Richard Davis, and Hugh Possingham. Amongst their many valuable contributions, the most valuable include private and group discussions, guidance on the course of the project, advice, collaboration on the manuscripts, and assistance with manuscript preparation. Others who read drafts at various times also contributed to the refinement of my ideas: Zoran Aleksic, Peter Chesson, Mike Dale, Ian Davies, Habiba Gitay, Peter Menzies, and Mike Dale. A number of short discussions with people were also influential: Peter Cochrane, Charles Zammit, Andrew Moore, and Clark Glymour. Suppliers of data are acknowledged in each manuscript where applicable.

I appreciate the influence of members of the Research School of Biological Sciences A.N.U. - particularly Les Watson, George Micklos, Barry Osmond, Peter Firth, Chris Buller - not only for their roles in the running of the school and providing a stimulating and supportive environment for academic work, but also for involvement in committees and consultations that played a part in my academic development. Thank you also to the staff of the Ecosystem Dynamics Group for general assistance throughout the project, Ian Noble, Marilyn Ball, Peter Chessen, Ian Davies, Jack Egerton, Sandy Berry, and Bill Parker.

I also thank the people from the Australian National University staff who assisted in my academic development: for writing skills - John Clanchy from the Study Skills Centre; for formulating research proposals - Margot Pearson from the Centre for Educational Development and Academic Methods, and Sylvia Hibberd from the Research Grants Office; and for seminar and panel discussion skills - Geoff Mortimore, David Grantham, John Martin, Linda Groom and other members of The University Toastmasters Club.

I gained a great deal of support from my student colleagues and friends, David de Pury, Jane Symonds, and Russell Cameron. Finally, my greatest appreciation goes to my friend and colleague, Gerlese Sachse-Åkerlind, for her commitment, understanding, and companionship throughout the ups and downs of the project.
Abstract

Machine learning is a field of artificial intelligence research where computer programs simulate the acquisition and application of human knowledge. This thesis applies machine learning to model development, where model development is defined as the formation of symbolic representations of reality given data or examples. The prediction/explanation problem is defined as the problem of achieving prediction and explanation simultaneously with a model. The goal of the research was to address the prediction/explanation problem by developing a machine learning system that produced predictive and explanatory models of the response of animals to the environment.

A number of machine learning systems were examined. Decision tree induction algorithms and a Bayesian classifier system gave good predictions, but the explanatory ability was low. Theoretical examination revealed that prediction and explanation are complex tasks, with a number of different components, defined by a formal learning paradigm. A range of distinct forms of prediction and forms of explanation were identified. A system for induction of rule-sets called GARP was developed and used to illustrate the character, cause and methods of controlling the prediction/explanation problem in habitat analysis. Resampling was used to determine predictive accuracy and rules for providing explanations. Finally GARP was applied to problems in wildlife management such as finding the factors that predict the distribution of wildlife, the environmental factors necessary for wildlife, and the factors controlling wildlife distribution.

The results show that there are a number of fundamental difficulties in developing models for prediction and explanation, some of which contribute to reduction in both prediction and explanation, and some that reduce one or the other. However no fundamental limitations to prediction and explanation were found given the definitions of prediction and explanation used in this study. Adoption of machine learning methods and the formal learning paradigm assists in overcoming these difficulties and finding the underlying model reliably. The pragmatic and scientific benefits of this work are the development of machine learning both as a tool for automated modelling and as a practical method of inquiry into the acquisition of knowledge.
Technical Summary

The first manuscript is a review of some existing applications of machine learning to ecology and functions as a tutorial introduction to the field of machine learning (Manuscript 1). The second manuscript introduces the prediction/explanation problem and sets out a research strategy for understanding and finding a solution to the problem in ecological modelling (Manuscript 2).

The next three papers examine two existing methods for predictive and explanatory adequacy: decision tree induction and the Bayesian classifier. Qualitative modelling of biotic response using decision trees is introduced (Manuscript 3). In a comparison of a number of modelling methods, including linear models and expert system development, decision trees were found to predict well but were poor at explaining biotic response (Manuscript 4). A new system was developed based on a Bayesian classifier (Manuscript 5). This matrix-based, probability method allows rapid development of expert systems due to incremental and integrated knowledge acquisition. However, it contained a number of limiting assumptions that could make prediction unreliable, and its potential for providing explanations from analysis of data was low.

Following the failure of the previous methods, it was clear an analysis of the nature of prediction and explanation was needed. Using the Bayesian classifier in an empirical study, a number of forms of prediction were identified: e.g. resubstitution, resampling, forecasting, extrapolation and transmission (Manuscript 6). Similarly explanation was found to be a complex and difficult concept to define, although the notion of an underlying structure provides some guidance (Manuscript 7). Rules were identified as explanatory structures that support prediction. Support for rules as explanations in ecology was obtained through a survey (Manuscript 8).

A rule-set induction system was developed with the potential for prediction and explanation (Manuscript 9). Issues of system design for achieving prediction and explanation were examined in an empirical evaluation of natural and artificial data sets (Manuscript 10). Finally, the explanatory ability of the system was compared with linear regression for explaining the presence and absence of waterbirds on disused quarry pits (Manuscript 11).

Achieving explanation and prediction using a machine learning approach is discussed in the conclusion (Manuscript 12). The characteristics and causes of the prediction/explanation problem are discussed. While there are a number of fundamental difficulties in achieving combined prediction and explanation, no fundamental limitations to achieving this goal were discovered. These difficulties can be ameliorated through the increased computational resources, more well defined model development methodology and data driven approach to model development that machine learning methods provide.
# Contents

Section 1. Introduction and review of the problem and methods

| Manuscript 1. | Machine learning in ecological modelling | 13 |
| Manuscript 2. | Prediction versus explanation? - a research program for machine learning | 37 |

Section 2. Evaluation of existing methods

| Manuscript 3. | Modelling of qualitative data via machine learning | 53 |
| Manuscript 4. | Using induction of decision trees to predict greater glider density | 59 |
| Manuscript 5. | LBS: Bayesian learning system for rapid expert system development | 71 |

Section 3. Theoretical development of prediction and explanation

| Manuscript 6. | Learning to predict: an empirical evaluation of the Bayesian classifier as a general predictive system | 91 |
| Manuscript 7. | A structural view of scientific explanation by explanatory modelling systems in ecology | 111 |
| Manuscript 8. | A survey of forms of explanation used in ecology | 133 |

Section 4. Rule set induction for prediction and explanation

| Manuscript 9. | Induction of sets of rules from animal distribution data: a robust and informative method of data analysis | 151 |
| Manuscript 10. | The effect of bias on prediction and explanation in induced rule sets | 157 |
| Manuscript 11. | Using rule sets to explain animal response to the environment | 175 |

Section 5. Conclusions

| Manuscript 12. | Nature and solutions to the prediction/explanation problem in ecology | 193 |
| References | 207 |
| Appendix 1 | Publications during doctoral research | 221 |
Section 1. Introduction and review of the problem and methods

Machine learning is a field of artificial intelligence research where computer programs simulate the acquisition and application of human knowledge. This thesis applies machine learning to model development, where model development is defined as the formation of symbolic representations of reality given data or examples (Manuscript 1). The prediction/explanation problem is defined as the problem of achieving prediction and explanation simultaneously with a model (Manuscript 2). The goal of the research was to address the prediction/explanation problem by developing a machine learning system that produced predictive and explanatory models.
Introduction

Modelling is a process of formalising knowledge about the world. Two main uses of models in the physical sciences from the time of Newtonian mechanics have been to improve the precision of predictions and the clarity of our explanations of the nature of the world. However, problems of prediction and explanation have plagued many areas of ecology. For example, failure to predict is a significant problem in applications of ecological models to environmental impact assessment (Bradbury et al. 1984). One reason for this is that predictions or explanations by themselves are difficult due to characteristics of ecology such as noisy data, ranges of spatial and temporal scales, and difficulty of conducting laboratory experiments on natural environments.

A more subtle and interesting problem is the difficulty of achieving prediction and explanation with the same model. The evidence of years of modelling experience has shown that some types of model are better for predicting while others are better for explanation (Loehle 1983). Two examples are simulation models and curve-fitting. Simulations often have disappointing predictive accuracy but generally explain well as they represent ecological phenomena realistically (Gross 1989). Models developed through statistical curve-fitting predict, but are often criticised for not explaining (i.e. representing realistically) the phenomena of interest (James and McCulloch, 1990). The problem of achieving prediction and explanation with the same model is referred to in this thesis as the prediction/explanation problem. As prediction and explanation are the major goals of modelling, and characterise successful models, understanding and solution of the prediction/explanation problem could be regarded as the major desideratum of the field of ecological modelling.

The prediction/explanation problem is becoming more serious with increasing availability of large-scale databases and more powerful computers. Models are being linked to databases to perform prediction and analysis in a more automated fashion, e.g. geographic information systems (GISs) in land management and planning agencies (Green 1991). Automated modelling could increase the rate and reliability of processing of information from databases for prediction and explanation. Because humans are less involved in the automated parts of the process, the problems of prediction and explanation become more acute, with a greater need for reliability and understanding of the modelling process. Most statistical packages are not reliable in the sense that the methods require a great deal of human involvement in deciding the appropriate method, monitoring the program for computational errors (e.g. lack of convergence and matrix singularity) and then application or use of the model, usually in a different package. Linear regression is an example. Automated procedures for finding the minimum number of parameters, such as stepwise reduction, are unreliable for finding the optimal model (James and McCulloch, 1990).
Most statisticians regard human intervention as necessary for choosing meaningful parameters for the reduced model. A question posed by automated modelling of this method of data analysis might be, under what conditions does there exist a reliable procedure for finding the optimal model?

**Basic definitions, assumptions and arguments**

Modelling is the activity of developing and applying a formal language that represents some aspects of reality. The symbolic formal languages may be mathematical models, logic based theories, or procedural code simulated on a computer. Modelling languages perform a great deal of useful work in science: assisting the exploration and communication of ideas about the world, and predicting and explaining real world phenomena. Modelling encompasses all formalisations, and is not restricted to the development of formalisms that represent features of the world. In this thesis the term, 'model', not only corresponds to formal representations of real world objects and their interactions, but also includes forms of logical expressions such as rules, procedural structures such as decision trees, and matrices of beliefs.

**Prediction and explanation**

To predict, in common language, is to foretell some future event. Prediction in modelling is more general, covering all assertions about states of affairs that are deduced from given information. A prediction is a proper deduction, it asserts some information that is not explicitly part of the given information. For example, the calculation of an output value from a mathematical function given input values can be a prediction.

Explanation may refer to one or many events. Explanation of a singular event, in common language, is to say why a particular event should occur by providing a justification for the event explained. This form of explanation is modelled in logic by a chain of reasoning that leads deductively to a conclusion. Explanation of a set of events, however, is to make the nature of the system known in detail. For example, a representational model illustrates, by correspondence with real world objects and their interactions, the system that generated a particular set of observations. A 'black-box' calculation tool does not provide an explanation. Explicitness provided by explanation in a modelling system is an aid to understanding.

**The prediction/explanation problem**

The prediction/explanation problem is the problem of achieving both prediction and explanation with a symbolic formal language.
An important point of reference in the whole discussion of models is the view of the scientific realists. This view can be characterised for modelling purposes as two assumptions:

- there exists a unique optimal model, and
- the optimal model is the underlying 'true' model of reality.

The second assumption is extensively dealt with in debates on scientific realism in the philosophy of science and is only touched on in this thesis (Jones 1991). A pragmatic approach is taken where models are used as predictive and explanatory tools. Whether the model is a 'true' model of reality is a secondary concern. The prediction/explanation problem suggests that the first assumption is false, as it suggests that a predictive model optimised for prediction and an explanatory model optimised for explanation are not the same, and therefore not unique. This answer is undesirable for at least three reasons. It suggests that:

- multiple models are necessary in developing computer systems for automated modelling,
- the formation of theories is blocked, and
- it contributes to misunderstanding in communication of the insights of models.

Examples of these three problems follow.

1. Data-based computer modelling methods such as non-linear curve-fitting use iterative techniques for optimising an objective function. It has been shown theoretically that in systems with structural complexity, i.e. systems with many interactions between variables, no representational model can simultaneously optimise generality, realism and precision (Levins 1970). This is a type of prediction/explanation problem where generality and realism are explanatory qualities, and precision is a predictive quality. One way for an automated modelling system to deal with this problem would be to develop different models optimised for different purposes, e.g. a general model, a realistic model and a predictive model. Systems for developing and manipulating multiple models for different purposes may be possible, but lead to a complex and inelegant system.

2. The prediction/explanation problem may block theory formation. To see that both prediction and explanation are necessary for a scientific model or theory, consider the occupations of people who perform one without the other. To predict without explaining is to be an oracle, not providing justification for information provided. To explain without predicting is to be a journalist, reporting and justifying past events but not speculating on future events, or the nature of processes that determine those events. The identities of people that perform one but not the other are not scientists, given a view of science that sees one of the main roles of scientists...
as to produce justified predictions. Thus models, in their role as proxy scientists, fail if they do not combine prediction and explanation.

3. To the user of models, lack of models that mutually satisfy these goals contributes to confusion between the uses of models. An example where confusion over the uses of models has led to unsatisfied expectations is the field of environmental impact assessment (Bradbury et al. 1984). Studies of the proposed impact of development on biological communities often lead to no more than descriptions of the status of the species in the region, and descriptions of the patterns and relationships between species. While the use of these methods adds to understanding of a system, the methods of analysis do not generally predict well. Expectation that they should predict the impact of a development leads to dissatisfaction with ecology, ecologists, or ecological modellers.

The general approach of the thesis

The above examples illustrate the importance of developing modelling methods that mutually satisfy these goals. Often new approaches can give insights and solutions to old problems. This thesis is an examination of the prediction/explanation problem using recently developed machine learning methods. These methods are characterised by being knowledge-based; the model primarily represents knowledge about a situation, rather than a representation of a situation. A machine learning method uses data and heuristic search methods to develop models, i.e. it is an inductive method based on forming generalisations from experience presented in the form of data.

The thesis is concerned with the problems of modelling phenomena and only secondarily with the phenomena themselves. The applications of modelling the response of wildlife to the environment and diagnosis of diseases are used to explore and illustrate important issues in modelling via machine learning.

The investigations of this thesis are primarily empirical, conducted through experimentation with machine learning modelling systems. Machine learning is useful in an empirical study of modelling because a number of different models can be developed in different areas of ecology. Conventional modelling methods where a single hypothesis is tested, or single models developed, are incapable of establishing the average behaviour of repeated model developments. A useful spin-off of this research is the computerised model development system, a useful tool for data analysis.

An understanding of the characteristics of the prediction/explanation problem in knowledge-based methods would contribute both to ecology and to machine learning (Davey and Stockwell, 1991). To achieve this goal the existence of the problem needs to be demonstrated, its causes understood, and how it might be alleviated
demonstrated. If possible we need to develop a system that achieves the goals of prediction and explanation simultaneously. This thesis sets out to achieve these goals.

The remainder of the introduction is composed of two sections. The first is a brief description and introduction to the main words and concepts of machine learning and automated model development. The second section is a summary and guide to the structure of the thesis, including a very brief summary of each chapter, the motivation for the chapter, the conclusion, and the originality of authorship.

**Definition of terms used in the thesis**

**Machine Learning**

Machine learning methods are primarily inductive methods of developing formal languages. An inductive method draws conclusions through generalising from observations, or data. A 'high-level' language is a formal language easily comprehended by human minds, e.g. simple, abstract and logical, as opposed to a structure easily comprehended by a computer system, e.g. a binary file.

Machine learning offers a way of developing models that both explain and predict. The justification for this view comes from the two main characteristics of machine learning: induction and 'high-level' languages. Developing models through induction guarantees consistency with a given set of facts. The output will at least predict the set of facts used to develop the model. Secondly, because the outputs of machine learning are easily comprehended by people, they will more readily explain adequately.

Machine learning offers a new methodology for efficient model development. Models developed without data, from accepted theories or known facts, may explain by providing plausible justifications for events, but there is no guarantee they will be consistent with given facts. The only reliable solution to inconsistencies is to reject the whole model. While falsification methodology may arrive at a correct structure in the long run, by elimination of inconsistent models, it is obviously inefficient. It would seem a more efficient process to at least generate hypotheses that are consistent with given facts.

Statistical predictive models such as curve-fitting techniques develop models consistent with the facts in the sense of providing a 'good fit' to the given data points. However the 'low-level' mathematical structure may be difficult to interpret. Often restrictive assumptions are necessary for making the mathematical process tractable. If these assumptions are not met then in a strict application the method is invalid.
Automated model development

What is an efficient methodology? In the context of large amounts of data being generated by geographic and remote sensing systems, in domains with little background knowledge to guide the development of plausible theories, an iterative process of generating single hypotheses, collecting data for testing them, falsifying and revising hypotheses seems inefficient. Automated modelling envisions a more streamlined approach to model development applicable to the information age. A general statement of the task is:

- given a large amount of information, with many variables with a range of types (continuous, ordered and categorical),
- some queries, or questions the model is supposed to answer,
- very little background knowledge (i.e. nothing is known about the general characteristics of the domain),
- produce a model that supports both prediction and explanation with a minimum of human interaction that could be used in some automated system, such as an expert system or in a geographic information system.

A sceptical response is that it is impossible - that a human expert is always needed to monitor and correct automated modelling systems - that "garbage in equals garbage out." The sceptic would point to instances where application of automated statistical methods (such as stepwise variable reduction in linear regression) has lead to poor models. My response is that the sceptic is avoiding the question. The issue is what is the problem and what can be done? Studying the prediction/explanation problem is a means to bring reliable automated scientific data analysis closer to reality.

The structure of the thesis

The thesis takes the form of a number of closely related manuscripts presented in sections (Table 1). A list of the titles, brief description, motivation, conclusions, and statement of originality of authorship where applicable to multi-author papers follows Table 1. Each manuscript has been prepared for publication in its respective journal. These journals include the fields of ecological modelling, artificial intelligence (AI), and machine learning. The manuscripts are written for the differing audiences in these fields. A small amount of duplication has been necessary to allow the individual chapters to stand alone.

The manuscripts are organised into sections corresponding to separate stages in the research strategy: a review of the problem and methods (Manuscripts 1 and 2), an
evaluation of existing machine learning methods (Manuscripts 3, 4, and 5), theoretical
development of prediction and explanation (Manuscripts 6, 7, and 8), a proposed
predictive and explanatory modelling system, rule set induction for prediction and
explanation (Manuscripts 9, 10 and 11), and conclusions (Manuscript 12).

Table 1. The manuscripts in the thesis cover a range of themes. The themes included in
each manuscript are indicated by asterisks.

<table>
<thead>
<tr>
<th>No</th>
<th>Introduction</th>
<th>Review</th>
<th>Empirical</th>
<th>Theory</th>
<th>Multi-model comparison</th>
<th>Applications</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Section 1. Introduction and review of the problem and methods.

0. Introduction (this chapter) - Stockwell D.R.B.

1. Machine learning for model development in ecology - Stockwell D.R.B.

Description: A review of the main areas of machine learning and their possible
applications to ecology.

Motivation: To provide a tutorial for readers less familiar with machine learning.

Conclusions: Machine learning, a field in the general computing field of AI, is
shown to have general application to data analysis and modelling in ecology.
2. Prediction versus explanation? - a research program for machine learning - Stockwell D.R.B.

Description: A selection of views on problems of prediction and explanation in ecology, their possible causes and a suggested approach to a research program to resolve the issue.

Motivation: The origins, assumptions, and goals of a research program need to be stated clearly.

Conclusions: The problem may be addressed through the development of a system that predicts and explains.

Section 2. Evaluation of existing methods

3. Modelling of qualitative data via machine learning - Stockwell D.R.B.

Description: Introduces and justifies the use of decision trees as a form of qualitative modelling in ecology.

Motivation: Due to the number forms of machine learning and the potential application areas there is a need to delineate the scope of the study to a specific field of application. As an initial step in research it is valuable to examine the methods that are currently being used.

Conclusions: Decision trees are a form of modelling with the potential to be predictive and explanatory, particularly where biotic response to the environment is a step function.


Description: A comparison of decision trees with models developed through regression, principle components analysis, and expert system development by knowledge acquisition.

Motivation: To evaluate a number of existing decision tree induction methodologies and learn their strengths and weaknesses.

Conclusions: Decision trees show increased predictive performance over development of rule-based expert systems by knowledge acquisition, or fitting of linear models. They have weaknesses as a form of representation and the structure can be unstable under resampling.

Originality: The first author conceived and carried out the study. Davey supplied the data and wildlife expertise for the expert system model, KA and PCA. Davis and Noble contributed to the conceptual development and preparation of the manuscript.
5. **LBS: Bayesian Learning System for rapid expert system development - Stockwell D.R.B.**

**Description:** A description of a predictive machine learning based system based on a Bayesian classifier, supporting the whole life cycle of expert system development and application.

**Motivation:** To describe and evaluate another modelling system called Learning Base System (LBS).

**Conclusions:** The system proved adequate for prediction, and was particularly useful for integrated and incremental knowledge acquisition. It had limited explanatory capacity.

**Section 3. Theoretical development of prediction and explanation.**

6. **Learning to predict: an empirical evaluation of the Bayesian classifier as a general predictive system - Stockwell D.R.B.**

**Description:** An inquiry into the nature of prediction using the Bayesian classifier system (LBS).

**Motivation:** To clarify the requirements of the task of prediction given the experience of applying LBS to problems.

**Conclusions:** The problems of prediction include the relationship between the training set and the test set, the choice of appropriate variables, and the adherence to assumptions of the model.

7. **A structural view of scientific explanation by explanatory modelling systems in ecology - Stockwell D.R.B.**

**Description:** A review and development of the theory of scientific explanation in ecology from a structuralist viewpoint.

**Motivation:** To clarify the nature of explanation.

**Conclusions:** Explanation can be broadly viewed as an explication of underlying structure, where forms of explanation represent different views of that structure.

8. **A survey of forms of explanation used in ecology - Stockwell D.R.B. and Sachse-Åkerlind G.**

**Description:** A survey of forms of explanation used by people with biological education.

**Motivation:** To provide support for a form of explanation in ecology for subsequent implementation in a learning system.

**Conclusions:** The survey shows that people accept ecology as a pluralistic field, admitting many forms of explanation. Causal forms of explanation are the most popular however. Analogy is not regarded as explanation. The choice of forms of explanation is not determined by the preferences of the respondent.
Section 4. Rule set induction for prediction and explanation


Description: Rules sets are introduced as a form of representation in modelling biotic response to the environment.

Motivation: To show that sets of rules fulfil the requirements of a modelling system.

Conclusion: Possible advantages over decision trees such as robustness and informativeness are proposed. Sets of rules offer a simple explanation where decision trees perform poorly.

Originality: The first author conceived and carried out the study. Noble contributed to the conceptual development and preparation of the manuscript.

10. The effect of bias on prediction and explanation in induced rule sets - Stockwell D.R.B.

Description: I describe the system called GARP, a genetic learning algorithm for rule set production (GARP) and its application to developing predictive rules from large, high dimensional datasets.

Motivation: To find the conditions and forms of bias that maximise predictive accuracy in the GARP algorithm.

Conclusions: The development of rule sets for optimising prediction needs at least two measures of utility, one for prediction and one for explanation. Bias can stabilise the development of the rule set, but can also interfere with prediction if it is inappropriate.

11. Using rule sets to explain animal response to the environment - Stockwell D.R.B. and Gitay H.

Description: An application of rule sets to modelling the response of water birds to the environment.

Motivation: To evaluate the method on real world data, and compare it with existing, competing methods.

Conclusions: GARP achieves the aims of the research, exhibiting the properties of prediction combined with explanation, in qualitative modelling of animal response to the environment.

Originality: The first author conceived and carried out the study. Gitay prepared the data set and provided wildlife expertise in interpreting the results.
Section 5. Conclusions

12. Nature and solutions to the prediction/explanation problem in ecology - Stockwell D.R.B.

Description: The prediction/explanation problem is described with reference to the modelling methods used in the study.

Motivation: To provide a description of the prediction/explanation problem and its causes.

Conclusions: While there are a number of fundamental difficulties in achieving combined prediction and explanation, no fundamental limitations to achieving this goal were discovered.
THE PREDICTION/EXPLANATION PROBLEM
1. Machine learning in ecological modelling

Author: David R.B. Stockwell
Publication status: Submitted to *Ecological Modelling*

Abstract: This paper reviews existing applications of machine learning in ecology and proposes some potential new applications. Machine learning provides a wide range of forms of representations for modelling, increasing the number of modelling methods available to modellers. Machine learning also contributes to an objective approach to model development using data and a formal modelling methodology.

Introduction

The most distinctive feature of machine learning as a practical method of data analysis is the rich variety of forms of representation available for modelling. Forms of representation used in machine learning include not only quantitative mathematical equations but also qualitative symbol systems, e.g. terms in classical logic, procedural programs or formal grammars. The process of modelling by machine learning differs from more conventional methods of statistical modelling. Machine learning often manipulates complicated symbolic descriptions, rather than simply optimising the values of parameters for a mathematical equation (Michalski 1983a). Thus, the output of machine learning can be qualitative as well as quantitative, modelling with the structure of objects, events, and their relationships. This output can form higher-level models required for modelling the complex dynamical phenomena seen in ecology (Ulanowitz 1988). Even though machine learning is a young science, with little integration, there are many strong results emerging on the extent of learning possible by algorithms with finite computational resources, e.g. computers and possibly human minds (Langley 1988a,b).

The aim of this paper is to show through examples that machine learning techniques can be of value in model development in ecology. The first section introduces the relationship of machine learning to scientific methods of enquiry. The second section deals with techniques of machine learning systems for modelling in ecology. The forms of representation in machine learning and their application to representing ecological phenomena are shown. Machine learning algorithms for model development and treatments of uncertainty follow. The final section describes how machine learning achieves the goals of modelling through application to description, prediction, and explanation in ecology.

Machine learning and scientific inquiry

There is a simple game called Mastermind® closely resembling a typical
modelling problem. Two players face each other at either end of a rectangular board containing a rectangular grid of holes for holding coloured pegs (Figure 1). One player, the adversary, places a single row of coloured pegs behind a screen hidden from the view of the other player, the theorist. The other player moves by placing a row of pegs in a row of the grid. The aim is to place a row of pegs on the board that match the colours and positions of the pegs behind the screen. At each attempt, the adversary must reply with white pegs for the number of correct colours and black pegs for the number of correct colours in the correct position. The game stops when the adversary replies with all black pegs, indicating a correct guess by the Theorist.

Figure 1. The game of Mastermind is an example of a typical machine learning problem. The aim of the game is for the Theorist to discover the hidden pattern of coloured pegs, as defined by the Adversary. Feedback on the similarity of the pattern proposed by the Theorist is given in the form of small black and white buttons: white when a peg is the right colour, and black when a peg is the right colour in the right place.

As an analogy to modelling, the hidden row of pegs chosen by the adversary is the underlying model of the system that the modeller, or theorist, is trying to find. The guesses are partial models, and the black and white buttons are indications of the performance of the partial model at prediction and explanation. A response of all black buttons by the adversary corresponds to successful validation of the model. The example illustrates many of the components of an idealised learning system (Kelly and Glymour, 1990):

- a set of possible worlds, (i.e. possible configurations of colours and numbers of pegs behind the screen),

- a language to express the hypotheses concerning the possible worlds, (i.e. the rows of coloured pegs available to the theorist), and a language to express the evidence the theorist
obtains about the world (i.e. the black and white buttons)

- a protocol for presentation of evidence to the learner (i.e. the adversary is honest and error free in the placing of the black and white pins in response to the guesses of the theorist)

- a theorist that produces guesses based on the evidence provided by the current protocol (e.g., a computer program outputs rows of coloured pins consistent with the evidence and guesses seen so far),

- a criterion of hypothesis adequacy (e.g. all evidence pins are black, meaning the pins are of the same colour and in the same position as those behind the screen), and

- a criterion of convergence of the learner to the goal (e.g., that the learner guesses correctly before running out of rows of holes on the board).

There are many ways of modifying the basic game. An empirical scientist proposes hypotheses instead of rows of pins, and receives evidence in the form of results of experiments. A computational theorist studies strategies for guessing the correct underlying combination for a range of languages and presentations of evidence. A philosopher of science may be concerned with justifying the discovery of truth, i.e. criteria for convergence on the true underlying model. An applied modeller would be concerned with finding a particular underlying pattern, given error-prone or noisy evidence from the adversary.

**Induction as scientific enquiry**

Machine learning is mainly induction - the process of generalising from observed data. However induction is not a valid logical process; one can make mistakes through generalisation. The observation of a flock of white swans does not guarantee that all swans are white. Despite this, the principle of induction is fundamental to human cognition. Russell stated that;

"The general principles of science, such as the belief in the reign of law, and the belief that every event must have a cause, are as completely dependent upon the inductive principle as are the beliefs of daily life." (Russell 1978, pp38)

Induction has a place in scientific inquiry. Scientists are doing induction when they propose hypotheses (Popper 1965, pp417). The emphasis of machine learning on induction does not necessarily mean that machine learning is unreliable. Machine learning programs can simulate methods of scientific enquiry such as the Popperian falsification of hypotheses (Kelly and Glymour, 1990).

The main difference between machine learning systems and human model development is the limitation to deducing theories of the world purely from observations. While an experimental scientist can falsify theories by experiment,
normally machine learning systems can't. The machine can only 'observe'. The problem with forming theories of our world by applying induction to databases of observations is that the records in the database are only positive examples. For example a species checklist of an area only contains the species that occur. The absence of a species does not mean it cannot occur, simply that it hasn't been found, or has failed to establish due to chance events or inability to migrate. Theories may be consistent with the data and yet be invalid, simply because no data to falsify them occurs, or occurs rarely. A scientist can overcome this problem by conducting experiments in controlled conditions that may never appear in the field. In the example of species inventories given above, an ecologist may conduct transplant experiments to determine whether the absence of a species in an area is due to a physiological intolerance of the species for the environmental conditions. Consequently applications of machine learning systems cannot dispense with human input. For example, large data bases could be analysed for promising hypotheses that are later validated by real world experimentation.

Modelling methodology

Modellers ensure the validity of models by 'keeping a close link to reality', and 'adopting rigorous validation procedures'. However, there seems to be little methodological guidance on practical questions such as "How many variables should I use in this linear equation, and how should I go about finding them?". Consider the curve fitting problem, e.g. linear regression. The problem of finding the best member of a family of equations is well understood; simply find the values of coefficients that minimise some objective function such as the sum of squares of the residuals. The problem of finding the best family of equations (e.g. which variables, and the form of the relationship, linear or non-linear functions) is much less well understood (Turney 1990a). There is very little technical literature on specifying the initial family of equations, and hence little formal guidance on this very important stage in model development (Glymour 1981, p323). The form of the model is normally justified subjectively, by appealing to the experts in the field, to previous published models, and to intuitive estimates of simplicity.

The methods of developing and justifying models are of concern to many machine learning researchers. In the formal approach to scientific discovery, algorithms and bounds on the complexity of the learning process are developed for learning descriptions of a certain form (e.g. logic or formal languages) from examples of a certain form (e.g. examples of a concept). This large and rapidly developing field offers exciting results on the limits to automated scientific discovery (Osherson et al. 1991). The research also aids the development of better practical algorithms by giving reliable knowledge of their performance under all conditions. Negative results are also useful; the limits to learning with a particular algorithm can be sobering. Simply taking a poorly defined concept such as hypothetical reasoning, or explanation, and define it in a formal (usually logic) system aids the conversion of
practical problems into working systems.

**Techniques of machine learning**

**Forms of representation**

There are many forms of representing knowledge through models. The advent of qualitative models in ecology, such as formal languages, suggests that the more well-known forms of numeric, quantitative models are not sufficient for modelling in biological systems (Green *et al.* 1986). The word 'model' has a much broader meaning than the restricted forms of algebraic, and analytic functions. It can refer to a simplified description of any kind, e.g. the boxes and arrows of a nutrient cycle, or a characterisation of a typical environment for an organism (i.e. niche). Below is a comprehensive listing of the forms of representation in machine learning as formulated by Carbonell *et al.* (1983) with examples and references to ecological applications.

The major initial classification of representations is the implicit or explicit distinction. A system with an implicit representation can behave as though it is learning by improving its performance, without communicating its internal representations of knowledge in any way (e.g., neural nets). A system with an explicit representation manipulates some formal system of symbols that have some meaning to a user. While neural nets show that explicit symbol manipulation is not necessary to model intelligent behaviour, explicit representations seem to have an advantage in the communication of knowledge: they allow the user to understand the behaviour of the system more fully and to use learned knowledge in other applications.

**Implicit representations**

Neural nets are composed of a number of simple interconnected units. Each unit has a number of parameters such as an excitation level, a relaxation rate, and strengths of connections between other units. The parameters learn through variation in the values of the parameters. The learning is usually a result of input signals modifying the strength and configuration of connections in the net. Neural nets can display simple learning behaviour such as pattern recognition. While neural nets are useful for modelling perceptual tasks such as digit recognition, they are not very suitable for working with complex logic based ideas, as the patterns and weights of connections do not provide a readily recognisable representation. One strong point of neural net is the handling of uncertainty, such as the combination of symbolic and continuous domains (Anderson 1990), and fuzzy set membership (Claudill 1990).
Explicit representations

Machine learning has inherited many formalisms from artificial intelligence (Carbonell et al. 1983): (1) parameters in an algebraic expression, (2) formal grammars, (3) logic based expressions, (4) graphs or networks, (5) computer programs or procedural encoding, (6) frames or schemas. Learning systems can potentially learn all these representations.

1. Numerical parameters. The search for parameters in a pre-defined algebraic model is a simple type of learning. For instance in multiple linear regression, we train a system on a set of data to develop algebraic expressions of a fixed functional form:

<table>
<thead>
<tr>
<th>Representation</th>
<th>y = f(x_1, x_2, ..., x_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>y = a_1 x + a_2</td>
</tr>
</tbody>
</table>

Most of the statistical methods of exploratory modelling (e.g. ordination and regression) would fit into this category. These methods are efficient in terms of computing time, but restricted in the form of representation. A machine learning method that develops algebraic models of variable form is the GMDH algorithm (Ivakhnenko 1966). The algorithm builds a tree of nodes, with each node a simple algebraic function. Training the system with data increases the complexity of the tree, forming more complex algebraic expressions. The algorithm has limitations such as low noise tolerance, and restriction to continuous interpolation and low dimensional real-valued data sets (Green et al. 1988b).

2. Formal grammar. The theory of formal languages originated in the study of linguistics and computer languages. Formal grammars have a number of strengths (Green et al. 1986). They can represent patterns created by dynamic qualitative systems in a concise way. They provide a flexible and natural representation of patterns in nature as shown by application to cellular development and animal behaviour.
Rewriting rules
R1. Sentence $\rightarrow$ Clause Predicate
R2. Predicate $\rightarrow$ Verb Clause
R3. Clause $\rightarrow$ Article Noun
R4. Article $\rightarrow$ the
R5. Noun $\rightarrow$ fox; rat
R6. Verb $\rightarrow$ ate

Derivation
S1. Sentence
S2. Clause Predicate (R1)
S3. Article Noun Verb Clause (R2, R3)
S4. Article Noun Verb Article Noun (R3)
S5. the fox ate the rat (R4, R5, R6)

Figure 2. A sentence is derived from intermediate structures by applying rewriting rules. Refer to text for explanation.

The basis of a formal grammar are rewriting rules that describe how intermediate objects such as parts of speech, phrases, and clauses could be manipulated to form sentences. A formal language is a set of symbols and rules for rewriting them. Consider the example is Figure 2. All symbols that can be rewritten as other symbols start with capital letters. Words that start with small letters cannot be rewritten (terminals). The example shows a set of grammatical rules, or a syntax for generating the sentence: "The fox ate the rat." The derivation of the sentence is listed. The derivation steps 1 to 4 providing models at different levels of abstraction. The final derivation, step 5, is the called the ground instance.

There is a need for automated methods for identifying formal languages from segments of languages. Machine learning methods for inducing formal language models from data would have an analogous role to statistics in mathematical modelling. The identification of formal grammars from a sequence of symbols was one of the earliest research areas in machine learning (Gold 1967). The development of efficient algorithms for identifying formal languages from noisy data is an active research area, although identification of anything other than simple languages is computationally intractable (Langley 1987).

3. Formal logic-based descriptions. Logic-based methods have been central to the study of machine learning. Logic is the natural representation of human argumentation, and important for communicating, supporting and questioning formalised domains such as mathematics (Tennant 1990). As logic has such a central role it is worth describing in more detail the application of logic to ecological systems. Formal logic has been used to model a description of the environment of the King Penguin, *Aptenodytes patagonicus* (Niven and Abel, 1991).
Table 1 The basic connectives of logic and examples are based on their use in natural language. The expressions 'not', 'and', 'or', 'if-then' and 'if and only if' are given symbols and precise meanings in logic.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>( \text{It is raining} )</td>
</tr>
<tr>
<td>( q )</td>
<td>( \text{It is snowing} )</td>
</tr>
<tr>
<td>( r )</td>
<td>( \text{It is cold} )</td>
</tr>
<tr>
<td>( s )</td>
<td>( \text{It is cloudy} )</td>
</tr>
<tr>
<td>( \neg p )</td>
<td>( \text{It is not raining} )</td>
</tr>
<tr>
<td>( p \land r )</td>
<td>( \text{It is raining and cold} )</td>
</tr>
<tr>
<td>( p \lor q )</td>
<td>( \text{It is raining or snowing} )</td>
</tr>
<tr>
<td>( q \Rightarrow r )</td>
<td>( \text{If it snows then it is cold} )</td>
</tr>
<tr>
<td>( p \Leftrightarrow s )</td>
<td>( \text{It rains if and only if it is cloudy} )</td>
</tr>
</tbody>
</table>

The basic element in logic is the proposition, any declarative sentence that can be either true or false, but not both. For example "It is raining" is a proposition. Propositions are usually denoted by capitals such as \( P, Q, \) and \( R \). We can build compound propositions by using logical connectives as shown in Table 1. The meaning of the logical connectives is similar to the meaning of simple connectives in natural language as the examples illustrate.

More complex logical expressions can be constructed using parentheses. Assigning \( P \) to the proposition "It is raining", \( Q \) to the proposition "It is snowing" and \( R \) to "It is cold", and using parentheses we can form the compound proposition \((P \lor Q) \Rightarrow R\) for sentence "If it is raining or snowing then it is cold." We then need a way to determine the truth value for a compound proposition given any assignment of true or false to the propositions. A truth table is a systematic listing of all possible combinations of true and false. The truth table (Table 2) shows the truth value of the compound proposition as a function of the truth values of the individual propositions.

Table 2. A truth table supplies the truth value for the compound proposition for each of the truth values of the elementary propositions.

<table>
<thead>
<tr>
<th>Elementary expressions</th>
<th>Derived Expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G )</td>
<td>( H )</td>
</tr>
<tr>
<td>( T )</td>
<td>( T )</td>
</tr>
<tr>
<td>( T )</td>
<td>( F )</td>
</tr>
<tr>
<td>( F )</td>
<td>( T )</td>
</tr>
<tr>
<td>( F )</td>
<td>( F )</td>
</tr>
</tbody>
</table>

The language of propositional logic is not complex enough to express the structure of many simple statements such as: (1) snow is colder than rain, (2) every plant needs rain, or (3) the sum of \( x \) and \( y \). Three more notions handle cases such as this. The resulting logic is first-order (or classical or predicate) logic. The proposition (1) above can be represented by the a binary predicate \( COLDER(snow, \)
rain). Other examples of predicates are: \( \text{PREDATOR(seal, penguins)} \) for "The Leopard seal is a predator of King penguins", \( \text{ODD(x)} \) for "The value x is an odd number." A two place predicate, often used in mathematical statements, is called a relation. e.g. \( \text{EQUALS(x,4)} \) for "\( x = 4 \)", and \( \text{GREATER(x, y)} \) for "\( x > y \)". The constants, variable or function symbols such as snow, rain, seal, penguin, 4, x, and y, are called elementary terms. A function is a predicate that returns a term, not a truth value and is usually written in lower case. For example, \( \text{mate(seal\_1)} \) represents a seal that is the mate of \( \text{seal\_1} \) not the truth of the statement that \( \text{seal\_1} \) has a mate.

Two other symbols complete the language of classical logic. The symbols \( \forall \) and \( \exists \) are called \textit{universal} and \textit{existential} quantifiers. The statement (3) above can be written using the universal quantifier \( (\forall x)(\text{PLANT}(x) \rightarrow \text{NEEDS}(x, \text{rain})) \) while the statement \( (\exists x)(\text{PLANT}(x) \land \text{NEEDS}(x, \text{rain})) \) using the existential quantifier means "there exists a plant (or some plant or at least one plant) that needs rain."

Specific machine learning induction systems develop specific types of logical models. The main types are:

- \textit{atoms}, simple unary and binary predicates (e.g., Colour = red, or a linear range 5<Length<10).
- pure conjunctive forms use only the 'and' connective between propositions, e.g. \( P_1 \land P_2 \land \ldots \land P_n \).
- pure disjunctive forms use only the 'or' connective, e.g. \( P_1 \lor P_2 \lor \ldots \lor P_n \).
- internal disjunction, where the disjunction is confined to the values of predicates e.g. \( ((A=a \text{ or } a \text{ or } a) \text{ and } (5<B<10)) \).
- DNF or disjunctive normal form, e.g. \( C_1 \lor C_2 \lor \ldots \lor C_n \) where \( C_i \) is a pure conjunctive concept, and
- CNF or conjunctive normal form, e.g. \( D_1 \land D_2 \land \ldots \land D_n \) where \( D_i \) is a pure disjunctive concept.

4. \textbf{Graphs and networks.} A graph is any structure consisting of a finite number of nodes connected by arcs. Most models of systems with flows of physical substances employ some type of directed graph representation. Graphs can represent the flows of nutrients in ecosystems, where the nodes are sinks of nutrients, and the links are flows of nutrient. Graphs come in various forms; the \textit{acyclic} graph is particularly useful (Figure 3). A graph is acyclic if it contains no closed loops.
A tree-structured graph has a single root node from which links radiate outward to other tree-structured graphs. Tree-structured models are widely used in ecology and biology to represent evolutionary relationships between species, or the morphological features necessary to classify a species. A decision tree classifies an unknown species into a unique class by guiding a user through a series of decisions, called nodes, traversing the tree-structured graph down to successive decision points, called branches, until there are no more decision points, called a leaf node. The species name that labels that leaf node is then assigned to the previously unknown species.

Many machine learning systems develop decision trees. Different algorithms generate decision trees with a variety of forms, primarily determined by the number of possible branches at each decision point (Figure 4). Figure 4a illustrates a tree generated by the algorithm ID3 which has large number of branches at each split (Quinlan 1986). Figure 4b illustrates a tree generated by CART which has binary splits, producing a bi-furcating tree (Breiman et al. 1984). Figure 4c illustrates a linear decision tree, a structure that models a sequential list of decisions.

Decision trees are a flexible form of representation. A number of methods exist for translating them into statements in logic, or sets of rules (Kalkanis and Conroy, 1991). For example, decision lists have been shown to be a generalisation of noise-free CNF and DNF concepts (Rivest 1987). Efficient algorithms exist for inducing decision trees from noisy data (Clarke and Niblett, 1989). The conciseness of decision trees and their properties of being closed and complete, are useful for general data analysis. A complete tree assigns a class to every possible object, thus providing total coverage of the space of possible objects. A consistent tree classifies every example into a single unique class. These properties guarantee that the
application of a decision tree is equivalent to a mathematical function.

5. Programs and procedural encoding. Many learning systems learn the ability to carry out a process efficiently rather than to reason about it. Procedural encodings take the form of familiar computer language directions, such as looping or conditional statements. Procedural representations may be useful for modelling the repetitive behaviour of animals, such as the foraging of animals (Figure 5). Procedural representations are implicit in all simulation models of animal behaviours.

```plaintext
procedure hummingbird_feeding
  until sufficient_nectar=true do
    find_flower
    until flower=empty do
      feed
    end
  end
```

Figure 5. Procedural representation of a hummingbird foraging for nectar.

6. Frames. Frames provide a larger unit of representation for organising less complex representations such as values, rules, or procedural code. Objects in the object-oriented programming languages are a form of frame-based system (Figure 6). A frame is composed of a number of labelled slots that can be filled with a variety of information, e.g. values, rules, or procedures. As a frame can contain another frame, frames are a flexible form of representation useful for problems involving hierarchies and varied data types.

```plaintext
frame hummingbird
  slots
eats: nectar
  forages: procedure(hummingbird_feeding)
lives: nest
  energy level: 50%
deamon
  if energy level < 0% then destroy hummingbird
```

Figure 6. Frame based representation of a hummingbird.

Frames represent the objects relevant to ecology in intuitively acceptable manner, e.g. trees and animals, (Pierret-Golbreich 1988) and landscapes (Green et al. 1990). Experiments with models of the moose, *Alces alces*, using object-oriented programming show that the technique captures the individual, heterogeneous and event-driven nature of animal/habitat interactions (Saarenmaa et al. 1988). The organised representation of complex objects also makes frames very useful. A series of rules and operations describing animal behaviour can be an unorganised series of
facts, inferences and procedures. Clustering all the types of knowledge relevant to an object into a single group results in a considerable increase in clarity. Frames also handle default values and exceptions easily. The main disadvantage of frames is that the formal and general rules for their interactions, such as unification, are difficult to define. Consequently, frame applications tend to have many aspects specific to the problem domain. For example in the hummingbird example in Figure 6, the foraging slot contains procedures specific to the particular simulation implementation.

Types of learning

Another method of classifying machine learning methods is by the types of learning used. Learning methods can be ordered by the difficulty of learning, which is directly related to the degree of assistance given by the teacher (Carbonell et al. 1983). At one extreme the teacher is entirely responsible for the input of knowledge, such as the acquisition of rules in an expert system by interviews with an expert. At the other extreme, in unsupervised learning, the system is free to explore the domain of application, finding patterns, discovering theories, and using those results for acquiring further knowledge. The main forms of learning in order of user input are: (1) rote learning, (2) learning from instruction, (3) learning by analogy, (4) learning from examples, (5) explanation based learning, and (6) learning from observation or discovery.

Rote learning. Most statistical model formation, databases and expert systems would come under this category. The system translates the data into an internal representation for later access in a form resembling the original data.

Learning by instruction. The learner transforms the knowledge from the input language to an internal representation, and integrates it with existing knowledge. The teacher must present the facts in an organised way that augments the students' knowledge. This method of learning has assisted information management (Hass and Hendrix, 1983), and the building of production systems (Rychener 1983).

Learning by analogy. Analogy uses a superficial similarity of a new description to a previously learned description. The system may convert an old piece of knowledge into a new one that is more useful. This form of learning has aided the formation of plans based on past experience (Carbonell 1983), and the solution of proofs in geometry (Anderson 1983).

Learning from examples. In learning from examples a learner is presented with a set of examples and counter-examples of the concept to be learned. For example, in learning the concept 'prime' a learner might be presented with a set of prime numbers, labelled as positive examples, and a set of non-prime numbers, labelled as non-examples of the concept 'prime'. Once the concept is developed that is true for all of the positive examples and none of the negative examples, the concept can be used to
classify further, unclassified examples. In developing concepts for predicting wildlife density, the places where animals occur can be used as positive examples, and places where they don’t occur as negative examples. Learning from examples has been found to outperform other methods such as linear regression and expert systems for predicting the density of the glider species, *Petauroides volans* (Stockwell et al. 1990, and Manuscript 4).

The task of discovering compound propositions in formal logic has received the greatest attention in the literature. Specific machine learning systems target specific logical types of concepts for learning. Haussler (1988) reviewed the theoretical properties of the most efficient algorithms known for learning simple predicates, tree structured predicates (e.g., Shape = Rhomboid or circle where Rhomboid = square or rectangle), pure conjunctive propositions, pure disjunctive propositions, internally disjunctive predicates, k-DNF (i.e. at most k disjunctions of pure conjunctive concepts), and k-CNF (i.e. at most k conjunctions of pure disjunctive concepts) concepts.

Decision trees have been the focus of many applied studies in learning from examples, possibly because these algorithms can efficiently analyse large amounts of data. An efficient algorithm is one that does little searching for alternative models and hence can analyse large numbers of examples of with many attributes (Breiman et al. 1984). Perhaps the most extensively researched algorithm is ID3 (Quinlan 1986). As an example of an empirical result in machine learning, empirical testing of ID3 for growing (developing decision trees), and pruning (removing superfluous branches) decision trees has shown that measures for pruning trees greatly influence the predictive accuracy of trees, whereas measures for selecting the decision node influence only the final tree size, not predictive accuracy (Mingers 1989a,b).

Machine learning algorithms for learning from examples have also been inspired by biological evolution. Evolution is a process whereby a species adapts to new or changed conditions by permanent changes to the DNA code propagated through the offspring of the parents. Thus evolutionary adaptation is a form of learning that goes on within groups of individuals over long time scales. DNA is the explicit language or knowledge representation used by living organisms. The types of changes to DNA in evolution (i.e. crossover, and mutation) provide a model for efficient probabilistic search algorithms (Holland et al. 1986). The Holland classifier is a string of usually binary attributes used to describe the target concept being learnt. A genetic learning algorithm uses adaptive search heuristics. First a set of classifiers is generated at random. The best strings are selected according to a measure of utility, and then modified with heuristics called recombination operators. Tests of performance have shown they are efficient ways of searching very large and complex spaces (De Jong 1988).

Explanation based learning. Explanation-based learning uses a significant amount of domain knowledge to guide its learning of concepts (Mitchell et al. 1985,
Dejong and Mooney, 1986). Given an example, the system gives an explanation consistent with the domain knowledge. The advantages of this method are efficiency, reduction in the search space by the inclusion of the additional knowledge, and consistency with general principles (Mitchell et al. 1985). As few as one example is needed to develop concepts with this method. In comparison, decision tree induction requires many examples. The search is also more efficient due to the constraints imposed by the domain knowledge. The method is also claimed to overcome brittleness (failing suddenly outside intended domain) and improves the accuracy of concept definition (Flann and Dietterich, 1989). The method compares well with other machine learning methods in the discovering of proofs in Whitehead and Russell's Principia Mathematica (O'Rorke 1989).

**Learning from observation or discovery.** Learning by discovery is a more exploratory form of machine learning that requires little background knowledge. The object of learning by discovery is usually a theory, a set of concepts that specify the properties of objects in a domain. Examples of successful applications of learning by discovery include the learning of theorems in number theory (Lenat 1983) and rediscovering 19th century chemistry (Langley et al. 1983).

Even though the scope of a particular theory might be restricted to a particular spatial scale, the aim of theoretical ecology is to explain the properties of all living objects at that scale, i.e. the universal properties (Pickett and Kolasa 1989). Universally quantified sentences express universal laws (e.g. $\forall x P(x)$). When there are an infinite number of $x$'s we cannot possibly examine all examples with finite computing or experimental resources. However, we can falsify a universal law with a single example, called a counter-example.

Because of the impossibility of validating laws over infinite domains, scientific methodologists have advocated the method of falsification of theories as a valid method of approaching the truth (Popper 1965). It is theoretically possible to identify such concepts if the learning system converges to the correct concept eventually. The notion of convergence to a theory is the sort of convergence proposed by Karl Popper. Formalisation of the notion of theory discovery has uncovered a number of interesting results on the limitations to learning (Kelly and Glymour, 1990). More formally a learning system AE (almost everywhere) converges to a theory on a learning environment, if and only if there after a certain number of examples are presented the learning system conjectures only true hypotheses. Only certain kinds of concepts are learnable under the AE convergence criterion for the discovery of theories and a falsification methodology. Some inductive problems are soluble only with a closed world hypothesis. One example of a closed world hypothesis is that given the parts of an object, the object has no further parts. In a biological context, a researcher in community ecology may assume that there are no more levels of detail (such as genetic composition), or no more interactions relevant to that problem.

A different line of research into theory discovery is inferring the existence of
unmeasured, or latent, variables in a more general acyclic graph structure (Glymour 1988a). The aim is to determine a network of variables where the links are correlations, and postulate unobserved or hidden variables between groups of variables that appear to have common causes (Figure 7). The relations between the hidden variables are used to determine the entire hidden structure of the system. This method is potentially useful in ecology where many variables cannot be measured directly, and are merely hypothesised. Glymour relates examples of the discovery of latent variables for intelligence and ability in psychological tests. Similar treatments could be made in ecology using concepts such as evolutionary fitness as hidden variables.

Figure 7. Given a correlation between two variables A and B represented as a directed graph (a), the discovery of a hidden variable corresponds to the formation of a new graph structure such as (b). This hidden variable may have been unobserved but explains the data better.

Uncertainty

In the scientific world, statements with certainty are rare. There are many forms of uncertain knowledge. There are formal logic uncertainties such as: x is A or B (either or both), x is not A (but could be something else), there exists some x that are A (but we don’t know which ones), and stating that all x are A when the number of x’s are infinite. There can be uncertainty in the outcome of procedures because of methodology: inaccurate or imprecise measurements by devices (random variation), systems highly sensitive to initial conditions (chaos), ignorance of important parameter values (latent variables), fundamental uncertainty (e.g., Heisenberg uncertainty principle). There is uncertainty in the meaning of concepts: fuzzy set membership (e.g., tall vs. short), vague descriptions (e.g., intelligence), and uncertainty of belief in ideas: confidence a person has in an outcome, confidence gained through an explanation of an event.

One common approach to uncertainty is to ignore it and to use a deterministic representation. This may result in a poor estimate of confidence in the results. Another approach is to use only almost certain knowledge, but this severely limits the domains of application. Consequently practitioners in AI, and particularly expert system developers, have many ways of dealing with uncertainty. Systems for reasoning with uncertainty are frequently discussed in the AI literature (Saffiotti 1987, Rothman 1989). An extensive monograph attempting a unified approach to uncertainty modelling shows the extent of the subject (Goodman and Nguyen, 1985).
As ecological information always contains uncertainty. Incorporation of forms of uncertainty in ecological models is essential for providing the user with confidence in the applicability of expert systems and learning systems in ecology. Two major areas of research into uncertainty in induction are: (1) coping with noisy training examples, and (2) learning uncertain concepts. After randomly sampling examples and non-examples of a concept, an identification procedure should conjecture concept with a high probability of being the correct concept. A distance between concepts (such as the conjectured and correct concept) can be defined as the proportion of examples on which the two concepts differ in their classification. A identification procedure does 'probably approximately correct' or PAC learning of a concept $L$, if:

$$\text{Prob}[d(L_-, L_h) > \alpha] < \beta$$

where $d(L_-, L_h)$ is the difference between the output of the procedure, $L_h$, and the true concept, $L_-$, and Prob is a measure of probability. The constant $\alpha$ is the allowed tolerance, and the constant $\beta$ is a confidence parameter (Valiant 1984a,b).

While classification errors or noise make learning of formal logic concepts more difficult, efficient algorithms exist for correctly PAC-identifying most simple concepts from data partially corrupted with random classification errors (Angluin and Laird, 1988).

Application of machine learning in ecological modelling

Until now the term model has not been precisely defined. The wide variety of forms of models provided by machine learning makes a universally acceptable definition difficult. Another approach is to define models not by their form, but by their use. Models can be developed for a number of different purposes, the most prominent being description, prediction and explanation. This section describes existing and potential applications of machine learning to description, prediction and explanation.

Description.

Any science, such as ecology, has many forms of description. Some are quantitative measures, e.g. biomass, the weight of organic matter, or indexes such as diversity. Some are qualitative, such as lists of the species occurring in a community or the growth form of a tree. More complex forms of descriptions, such as distributions (e.g. the age distribution of an animal population), and structured objects (e.g. sites are represented as vectors of values of variables), are also widely used. Formal logic provides powerful and mixed descriptions of sets of objects. For
example, internally conjunctive forms (e.g. $A_1 = v_1 \wedge v_2$) can describe combinations of requirements, such as food and shelter. Disjunctive forms (e.g. $A_1 = v_1 \vee v_2$) can describe alternatives, such as food sources. Intervals (e.g. $v_1 < A_1 < v_2$) can describe survival of an organism between a maximum and a minimum temperature.

A tree-structured variable has a finite set of hierarchically ordered values. For example, the variable 'vegetation' in Figure 8 represents the major growth forms of plants on land. When the variable 'vegetation' takes a value, represented by the nodes of the tree, this value is a set of all values below it on the tree.

![Figure 8. A tree-structured attribute has a finite set of hierarchically ordered values. The variable can take any single value from any node of the tree. The value is then a set of all species in nodes below it.](image)

Two forms of description are the characteristic description and the discriminate description (Michalski 1983a). A characteristic description specifies all the common properties of objects in a class. A discriminate description that distinguishes a given class from a limited number of other classes. A typical example of a characteristic description is a maximal characteristic description, a logical expression containing the most number of properties that are true for every member of the class. For instance, a taxonomic description of a species is a characteristic description which lists all the morphological features of all members of the species. A discriminate description specifies one or more ways to distinguish a given class from all other classes. Most interesting is a minimal discriminate description - the shortest logical statement that is true for all members of the class and not true for any non-members of the class. Taxonomic keys are based on discriminate descriptions in order to allow them to be used to distinguish between species.

Two of the basic operations in learning descriptions are specialisation and generalisation (Michalski 1983a). A specialised description covers (is true for) fewer elements of the class. For instance, if the rule $A=1 \Rightarrow \text{Class}$, covers all members of a set, this logical description will be specialised by the conjunction of another value of a different attribute to the rule, e.g. $A=1 \wedge B=2 \Rightarrow \text{Class}$. In generalisation, the description covers more elements than the more specialised rule. The rule $A=1 \Rightarrow$
Class is generalised by the disjunctive of the attribute $A: A=I \lor 2 \Rightarrow \text{Class}$. In a structured variable such as the 'vegetation' example in Figure 8, movement down the tree is specialisation, and generalisation corresponds to moving up the tree.

**Application of description to classification.**

The standard approach to classification is to define a measure of distance between objects, then to arrange objects into classes such that the members of a class are similar and the classes themselves are dissimilar. There are many measures of similarity within classes, mostly based on standard arithmetic measures between vectors of objects. For example:

given an object is a site with attributes $(x_{i1}, \ldots, x_{in})$ and a set of those objects $\{(x_{11}, \ldots, x_{in}), (x_{21}, \ldots, x_{2n}), \ldots, (x_{mn}, \ldots, x_{mn})\}$

a metric defines the distance between objects: e.g. the square metric

$$d(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$$

where $x_i \in \{1,0\}$.

**Table 3.** Machine learning provides an alternative method of classification to conventional metric based methods. conceptual clustering uses elements of the language to suggest new clusters, and a syntactic operation, the maximally specific generalisation (MSG) to perform clustering. In the example the final clusters are identical, but because the conceptual clustering method is based on languages, provides a description of the clusters instead of simply listing the items in each cluster.

<table>
<thead>
<tr>
<th>Metric classification</th>
<th>Conceptual clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects</td>
<td>Concepts (species A,B,C,D)</td>
</tr>
<tr>
<td>a=(1,1,0,1)</td>
<td>ABD</td>
</tr>
<tr>
<td>b=(1,0,1,1)</td>
<td>ACD</td>
</tr>
<tr>
<td>c=(1,1,1,1)</td>
<td>ABCD</td>
</tr>
</tbody>
</table>

**Calculate metric**

<table>
<thead>
<tr>
<th></th>
<th>Performance MSG on each pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>d(a,b)=2</td>
<td>MSG(a,b)=AD</td>
</tr>
<tr>
<td>d(a,c)=1</td>
<td>MSG(a,c)=ABD</td>
</tr>
<tr>
<td>d(b,c)=1</td>
<td>MSG(b,c)=ACD</td>
</tr>
</tbody>
</table>

**Result of merging closest objects**

$\{(1,1,1,1),(1,0,1,1)\}$

$\{(1,1,0,1)\}$

Select maximally specific MSGs

ACD

ABD

A machine learning approach to classification is called conceptual clustering and takes into consideration the language used to characterise the classes of objects (Michalski 1983b). A class is described by a simple expression in the description language. The results of the classification are classes with simple descriptions in the concept language. An example of conceptual clustering is shown in Table 3. The
concept describing a class is the list of species common to a site. Clusters are formed by finding the maximally specific generalisation (MSG); the list of symbols occurring in both sub-concepts. The sites with the longest MSG are joined and described by their MSG. In the metric method, the sites are represented by attribute lists. The distance between two sites is determined by the square metric. In the example given the machine learning method results in the same clusters as the metric method. The MSG clusters sites with the most shared species. The resulting description of each class lists the common species in each object in that class. Unlike the final description in the metric method, the description could then be used to classify further sites into vegetation types.

**Application of description to databases.**

Induction uses data, and databases manage data. To exploit the availability of data due to contemporary computer methods of storage and manipulation, new modelling tools are needed (Green 1991). Conventional databases and statistical modelling methods are designed to handle simple forms of information efficiently, i.e. perform data analysis. However much of the data now collected could also be more properly called information - chunks of knowledge gleaned from small sets of data in a qualitative form. One example is the use of vectorised representations in mapping instead of arrays of pixels. The roads in a map, or polygons formed by areas of contiguous forest types are complex types of data not amenable to mathematical methods of analysis based on simple objects, such as spatially identified site attributes.

Knowledge bases such as expert systems are databases with greater expressiveness and functionality (Bic and Gilbert, 1986). In databases, the need to store and retrieve large amounts of information constrains the forms of representations to a few simple types, e.g. hierarchical or relational records. Knowledge bases deal with a smaller number of more complex types. They contain models in the sense that they are selective abstractions which help users to describe, predict and explain events (Clancey 1989). Machine learning methods may provide automated means of extracting knowledge from knowledge bases.
Table 4. Given a data-base of tree heights and girths, a machine learning system can be used to find rules relating to the data. The predicate 'tree(_, _, _)’ refers to an item of data. The rule for determining if a tree is a canopy tree states that height must be greater than 20 m.

<table>
<thead>
<tr>
<th>Database</th>
<th>Induced rule for Knowledge Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canopy Tree</td>
<td>Height(m)</td>
</tr>
<tr>
<td>Yes</td>
<td>23</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
</tr>
<tr>
<td>Yes</td>
<td>50</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>No</td>
<td>15</td>
</tr>
<tr>
<td>Yes</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 4 contains an example of a relational database of that might be used in a forestry application to store the dimensions of trees. The user queries the database to obtain information about a specific variable (Canopy Tree). In a typical knowledge base in the form of a Prolog program, a rule is a Prolog clause (e.g. trees higher than 15m are canopy trees). By using rules the user sees a view of the database more closely resembling natural language (Bic and Gilbert, 1986). Large geographic databases requiring day to day maintenance will have algorithms that perform higher level analysis such as consistency checking and error correction (Green 1991). Thus machine learning research may be essential in providing the self-monitoring activities for sophisticated databases.

**Prediction.**

The linear statistical model is widely used for prediction. A linear model equates the expected value of the predicted variable with the sum of several weighted variables. For example, prediction involves the calculation of a value for a variable, Y, given a variable, X, using a linear model, Y = a + bx + ε, where ε is a random variable. The main test of whether a variable is useful for predicting another is high correlation. Regression is the use of correlation to fit a statistical model to data.

Decision trees have been the main tool of prediction in machine learning (Quinlan 1986). Accurate prediction in a decision tree requires that the expected probability of the class at the leaf of the tree is high (Buntine 1991). A general model of prediction of this form is shown below where the outcome Y is expected with high probability given the precondition, X.

Given P(Y|X) and P(X)
infer P(Y)

More general structures such as graphs are also being used to predict (Pearl 1986). The variables are nodes in a directed acyclic graph where the arcs signify direct dependencies between the linked variables. Because of the very general form.
INTRODUCTION AND REVIEW

of graph structures and the number of possible interactions between variables, updating a graph structure can be computationally demanding. Pearl (1986) developed efficient computation methods for updating graphs using evidence.

Applications of prediction - Geographic Information Systems

The use of geographic information systems and remote sensing is rapidly increasing because of advances in computerised techniques for handling large amounts of information. An important task in remote sensing is image processing - the process of modifying and analysing pictures. With machine learning it is possible to make inferences about images automatically. The example in Table 5 is an edge recognition problem. After developing a decision tree on a subset of the image, the classification rule recognised edges considerably faster than conventional techniques (Goodman and Smyth, 1990). Edge recognition is an area process - the intensity values of neighbouring pixels determine the presence or absence of an edge. Given standard matrix notation the Sobel operator $E_l(i,j)$, an arithmetic function of a $3\times3$ cell matrix $M$, signals an edge when a threshold value goes beyond a pre-set value (Duda and Hart, 1973, p271). The dispersion operator is based on the ordering of the intensity of adjacent cells. The decision tree induced on the test set uses the dispersion $W_i$ of $M$ (Table 5).

Table 5. Examples of forms edge recognition used in automated pattern analysis. The Sobel and dispersion operators are numeric methods, based on arithmetic operations on the surrounding pixels. The machine learning method is a decision list, where values of the dispersion operator are used to decide the presence of an edge.

<table>
<thead>
<tr>
<th>Conventional techniques</th>
<th>Machine Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sobel operator</strong></td>
<td><strong>Decision list</strong></td>
</tr>
<tr>
<td>$E_l(i, j) =</td>
<td>\Sigma (x - x_{ij})</td>
</tr>
<tr>
<td>$E_l(i, j) =</td>
<td>\Sigma (x - x_{ij})</td>
</tr>
<tr>
<td>$E_l(i, j) = E_l(i, j) + E_l(i, j)$</td>
<td>if $W_4&lt;16$ then non-edge</td>
</tr>
<tr>
<td><strong>Dispersion operator</strong></td>
<td>if $W_4\geq 8$ then edge</td>
</tr>
<tr>
<td>$W(i) = x_{(n+1-i)\times j}, 0&lt;i&lt;(n/2+1)$</td>
<td>if $W_3&lt;10$ then non-edge</td>
</tr>
<tr>
<td>where $x$ is the pixel with the $i$th magnitude in $M$.</td>
<td>else edge.</td>
</tr>
</tbody>
</table>

Decision tree induction has been shown to do edge classification more accurately and efficiently than the Sobel or dispersion operators. The commercial mapping package SIMPLE uses CART for map analysis, and includes statistical and topographic modelling packages (Walker 1988). The predictive performance of this package compares well with logistic regression in developing maps of the distribution of a number of Macropods (kangaroos) in Australia (Walker 1990). CART has also successfully classified vegetation types in a large-scale geographic application (Moore et al. 1990). The usual methodology is to develop a decision tree from a set
of ground-truthed examples, sampled from the area to be analysed. The tree that results is then used to classify the whole image. This process increases the accuracy and speed of processing geographic images.

**Explanation**

Perhaps the major advance of expert systems was the addition of the capacity to explain a solution to a user. An expert system explains by tracing back through the chain of production rules and facts that led to a successful result (Table 6). Two factors contributing to explanatory ability are (1) the clear separation of the representation of knowledge and manipulation of that knowledge, and (2) the modularity of logical inferences, single rules representing 'chunks' of knowledge (Davis et al. 1977).

Table 6. Explanation in a rule-based system proceeds by listing the clauses that were used to reach a conclusion. Given the facts, rules and goal below, the answer, 'tiger' is explained by listing the rules used in sequence.

<table>
<thead>
<tr>
<th>Facts</th>
<th>Goal</th>
<th>Result</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: does(eat_meat).</td>
<td>← it_is_a(X)</td>
<td>X = tiger</td>
<td>From R2 it_is_a(tiger) as it_is(carnivore) and has(dark_stripes).</td>
</tr>
<tr>
<td>F2: has(dark_stripes).</td>
<td></td>
<td></td>
<td>From R3 it_is(carnivore) as does(eat_meat).</td>
</tr>
<tr>
<td>Rules</td>
<td></td>
<td></td>
<td>From F1 does(eat_meat).</td>
</tr>
<tr>
<td>R1: it_is_a(cheetah) ←</td>
<td></td>
<td></td>
<td>From F2 has(dark_stripes).</td>
</tr>
<tr>
<td>it_is(carnivore),has(dark_spots).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2: it_is_a(tiger) ←</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>it_is(carnivore),has(dark_stripes).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3: it_is(carnivore)←</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>does(eat_meat)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Representation of knowledge as rules thus facilitates explanation. Machine learning methods that use rules to represent relationships in the data will facilitate the development of explanatory systems such as expert systems. These include one of the first machine learning systems for developing rule sets, called INDUCE, for diagnosing diseases in soybean plants (Dietterich and Michalski, 1981). Another, BEAGLE, automated acquisition of knowledge from data bases (Forsyth 1981). These machine learning systems learn to explain in the sense that they discover knowledge about a system that can be used for explanations.
One of the meanings of explanation is to make clear the true structure of a system, i.e. to discover the underlying system that produces a set of data. Sometimes this involves deriving hidden or latent variables. For example, background knowledge or statistical analysis may suggest a graph structure between variables shown in Figure 9. Hidden variables can be postulated between groups of variables that appear to have common causes. The relations between the hidden variables are determined, and the process repeated uncovering the entire hidden structure of the system. The result is the graph structure shown in Figure 9. A machine learning system to perform structure recovery learns to explain in the sense that it discovers novel variables not included in the original data set which explain the data better than the observed variables.

Conclusions

Machine learning is a general method for acquiring the range of representations used in conventional modelling and in artificial intelligence. Its applications cover a wide range of problem domains in ecology: databases, geographic information systems, forming descriptions, predicting, explaining, classifying, and modelling of learning processes, such as scientific discovery and evolution. While modelling is partially an art in which human involvement is essential, it is also a time-consuming and frequently a subjective practice that could profitably be streamlined. As the successes of artificial intelligence have shown, some human cognitive activities can be represented as algorithms and performed by machines. Human learning can similarly be modelled. As modelling is an algorithmic process, it seems perfectly amenable to analysis and algorithmic formalisation. A wide range of algorithms for modelling human learning can potentially be applied to model development in ecology.

Author: David R.B. Stockwell
Publication status: Submitted to Ecological Modelling

Abstract: The history of modelling in ecology suggests a conflict between prediction and explanation that continues to undermine effort to establish ecology as a quantitative field of science. What would constitute a research program to investigate this problem? There are many reasons for models not achieving the goals of prediction and explanation separately, what factors would be varied, and what factors controlled? Machine learning offers the potential to investigate the development of models while explicitly addressing prediction and explanation. A research program based on machine learning would develop understanding of the problem and improve the quality and usefulness of modelling in ecology.

Introduction

Any reader of ecological reviews notices a history of conflict in ecology over the role and abilities of ecological modelling. The most frequent and fervent calls are for ecological models that predict e.g.,

"Mathematical modelling in ecology for its own sake does not make sense. We make the models to generate predictions which have to be tested." (Lomnicki 1988)

and explain. e.g.

"A good theory should do more than just predict, it should explain." (Loehle 1983).

Prediction and explanation are the main aims of ecological modelling. Their application to biological systems has a long history. In 1798, Malthus (1978) published a controversial essay on the principles of population growth and decline. He provided a formal explanation of the inevitable limitation of the growth of populations by resources - prediction that a population is doomed if resources are finite, and a mathematical characterisation that explained the prediction. The field of population dynamics has continued to be one of the major subdisciplines within biological modelling. The theory has influenced the management of exploited natural populations in industries such as fisheries and agriculture. Yet, in a very telling review of the most influential models in use in ecology, including fisheries management, Hall (1988) concluded,

"Again and again evidence to the contrary has been ignored by the advocates of the theory, and very weak fits of model to data have been offered as strong support for the theory."
Such finding would have grave implications to the management of resources and impacts on those resources. Decision makers trained in engineering or physical sciences often expect models of ecosystems to be highly mathematical with predictive accuracy and explanatory ability. Such decision makers often believe that resource and impact statements can be improved by better prediction in models of the natural environment. Where the impact of decisions is not apparent for many years, a predictive model can be invaluable for examining the consequences of various options. Yet a dissatisfaction with practical applications of ecological models used in impact assessments is apparent (Bradbury, et al. 1984). While recognition of the limitations of models is necessary, in the long term a greater understanding of the reasons for the difficulties of modelling in ecology would help in the development of more predictive and explanatory models, and hence 'better' models.

This paper puts forward the view that recent advances in the field of computer science aimed at modelling human intelligence can be applied to the problem. Machine learning is the sub-field of AI dedicated to modelling the acquisition of human knowledge with computer programs. As the development of models is a form of acquiring knowledge, machine learning has obvious applications. The aim of the paper is to show how machine learning can address directly the lack of predictive and explanatory models in ecology in a coherent research program.

**Machine Learning**

Machine learning methods are primarily methods of developing representations through supplying an algorithm with data or examples of a concept to be learned. The main method used is induction, drawing conclusions through generalising from specific instances (Michalski 1983a). A more familiar inductive method is statistical curve-fitting techniques, that use data to find a curve providing a 'good fit' to the given data. Machine learning methods are distinguished from a method such curve-fitting by the use of heuristic, or rule of thumb, methods and controlled searching of a set of solutions, as opposed to explicit mathematical solutions. Another feature of machine learning is that the representations are usually 'high-level' languages such as logical statements, or procedural code, representing knowledge about a system. In comparison mathematical structures tend to represent the observable aspects of the system being modelled.

Machine learning explicitly models the methodology for the development of a representation. This is because incorporation of the methodology into a computer program forces a great deal of the assumptions and normally human tasks performed during model development into computer code. The final program is a tool for experimentation in model building, allowing the generation of numerous models under a range of initial conditions, such as varying data sets.
Failure to predict or explain

We need a clear idea of the nature of prediction and explanation in modelling. Explanation and prediction are tasks that models perform, rather than attributes of models such as generality, realism and precision. In explanation, the model is used to communicate some information about an event. For example, the presence of an organism in a region is justified by a model of the causal relationships between and within the biotic and abiotic components of the environment. Prediction in this case is the anticipation of the of biotic response to certain phenomena given value of the causal variables.

Given a requirement for combined prediction and explanation, a model can fail by failing to achieve either prediction or explanation. Thus there are three problems: why models don't predict, and why models don't explain, and the relationship between prediction and explanation. Following are typical examples of claims concerning models that either predict or explain.

Failure to predict

It is claimed that explanatory models often fail to predict. Complex mathematical models composed of differential equations take the role of theory in the field of plant physiology. Their performance can be compared to statistical linear regression based on fitting a straight line to field data.

"Even the most elaborate crop growth models are not yet as accurate in yield prediction as relatively simple regression models" (Gross 1989).

Inclusion of a multitude of factors leads inevitably to a complex mathematical model or simulation. With an increase in the number of components there is an increase in the number of parameters to be determined. The simulation or model can be highly sensitive to the values of these parameters. For example, in natural populations most parameters such as birth and death rates or degrees of competition are difficult to measure and may themselves be dependent on other factors. Thus there is an inevitable decrease in the confidence in predictions of models of increasing numbers of parameters.

Failure to explain

It is often claimed that simple predictive models fail to explain. Simple curve-fitting models are claimed to predict in limited situations. But despite the predictive success of statistical curve-fitting techniques such as regression, multivariate, and logistic analysis, there is a perceived absence of explanatory ability with these models. There is also an absence of a clear idea of what constitutes an explanatory
model. Attempts have been made to make the requirements for biological explanation more explicit.

"Statistical inference may be possible, but, as with univariate analysis, without experiments even the most insightful applications can only hint at roles, processes, causes, influences, and strategies." (James and McCulloch, 1990).

Simplification of models leads inevitably to greater abstraction. With greater abstraction comes averaging of disparate quantities, omission of variables with a minor effect, and concentration on direct interactions at the expense of more subtle 'higher order' effects. Because biological components are absent from simple models the models may be accused of being 'unrealistic' even though the model might predict well within its intended limited domain of application. Two conflicting views of explanation are apparent. One view is that explanations should be simple, abstracting the details into essential features. The alternative view is that explanations should be complex, including all the details of the real world.

**Failure to predict and explain**

The relationship between prediction and explanation is the most interesting and unclear problem. Are they dependent or independent? Are they co-achievable in the same model or do they compete? There is a perception that the two are independently achievable with different types of model. What could prevent the achievement of both prediction and explanation? Many authors have offered their assessment of the problems of prediction and explanation in ecology, with implicit or explicit solutions. These views have been organised into categories, each of which could constitute a component of a research program.

**Inadequate mathematical tools**

Some have blamed inadequate modelling methodologies, suggesting that new forms of analysis need to be developed. In 1940 this may have been the case.

Haskell 1940. "Mathematization is to be regarded as a normal phenomenon in the course of the development of any science" ... " though we do not yet have the mathematics necessary for this." (McIntosh 1980)

Is prediction and explanation simply a matter of increasing the degree of mathematical sophistication? We can compare the situation in ecology with the best scientific models of physics and chemistry. For example:

"Watt (1962) reviewed the rather distinct traditions of mathematics in population ecology and noted the lack of impact of sophisticated mathematical theories in advancing population ecology, as contrasted with their utility in physics." (McIntosh 1980).
Given the burgeoning of mathematics in the last fifty years it is unlikely that the mathematical tools available are inadequate. Models of predator-prey relationships have used advanced mathematics extensively. An example of sophisticated mathematical treatment the modelling of the population dynamics of the crown of thorns starfish, *Acanthaster planci*, in a multi-species system (Antonelli et al. 1989). This treatment explains starfish outbreaks as limit cycle phenomena and has shown agreement with data. It appears to be an example of a successful application of recently developed mathematical tool, although it is not clear if the model can predict starfish outbreaks.

An alternative approach to developing existing mathematical tools is to find and apply new tools. Recently modellers have applied the methods of computing and artificial intelligence to ecology. Examples of this strategy include application of formal languages to landscape modelling (Green et al. 1990), and self-adaptive algorithms, fore-runners of machine learning systems, to predictability of coral reef communities (Green et al. 1987). The usefulness of novel modelling methodologies is always unknown, and there may be a long delay between preliminary applications and useful implementations.

There is a danger that research is driven by novel methodologies and not by the phenomena to be modelled will remain a series of interesting applications, without general application. Even though linear systems theory is now well entrenched into areas of ecology such as nutrient cycling models, at the time the adoption of linear systems theory, ecology to be seen driven by available methods of analysis. According to Watt (1975),

"... present ecological theory has been swept away by a tidal wave of linear systems theory simply because of the great body of standard methods available for linear analysis". "... linear systems theory does not describe ecological systems." (McIntosh 1980)

A major assumption in the use of linear systems theory for modelling is that ecological interactions approximate linearity. The practice of using linear approximations to nonlinear systems can lead to accurate predictions in limited situations but be highly inaccurate outside that range. As a consequence many researchers are now using non-linear models. However, adoption of non-linear models will not necessarily solve the problem. Often the non-linear behaviour of most interest in strategic decisions is catastrophic, such as under what conditions a bushfire becomes a wildfire or when fishery stocks collapse. Chaotic behaviour is a special case of nonlinear systems, systems that change in sudden jumps at a threshold value, or have varying relationships between the variables throughout their range. In this case more sophisticated tools would not lead to better prediction when the system is essentially unpredictable.

To correct the tendency for applications to be driven by new mathematical methods, a research program to explore the potential of a new modelling
methodology, such as machine learning, must focus on the clear and important goals, i.e. the prediction and explanation of biological phenomena.

**General methodology**

Another view is that the application of existing methods is incorrect. This criticism raises a number of issues concerning general methodology for producing adequate models in a reliable way. Current modelling methodology involves a cycle of processing at least consisting of the stages of selection of the general form of the model, model development and testing. Methodology has been criticised at each of these stages.

Selection of the general form of the model usually involves a choice of the general mathematical function to use. One example is the curve-fitting problem, the problem of determining the minimum number of parameters to use in fitting a curve to a set of data points.

"The problem of finding the best member of a family for given data is relatively well understood. The standard solution is the least squares solution. The problem of finding the best family for given data is much less well understood." (Turney 1990a).

The standard solution is to minimise the number of parameters in the family. This solution is usually justified by reference to the well-known principle of parsimony, that we should advance no more complicated hypothesis to explain data than is necessary. An analysis in a precise model of scientific inquiry has shown that the adherence to simplicity can interfere with solving scientific problems (Osherson and Weinstein, 1990). Thus the justification for the practise is unclear, although a justification in the case of linear regression has been provided using the concept of stability (Turney 1990a). Stability is desirable because a stable model is less likely to be falsified by repeated experiments. These analyses demonstrate that precise models of scientific methodology can lead to exciting insights into the problems of modelling.

The other component of modelling methodology that has been criticised is testing, or validation. Validation is believed to be a means of establishing the adequacy of models, and therefore improving models.

"We must, I believe, tighten the link between the development of theory and the testing of theory." (Hall 1988).

What seems to be suggested is the widespread adoption of the modelling cycle for model development, where every modelling project explicitly contains a component of model validation, or testing on real data. These principles have been adopted in software engineering and are essential for maintaining control over quality yet do not seem to be as widely or explicitly adopted in modelling. It is well known however
that validation can falsify and hence, allow definitive rejection of a model. One cannot prove a model to be valid with finite data when there are an infinite number of data available. Thus the distinct separation of model development and testing can lead to wastage of the development effort if the data subsequently prove the model false.

An alternative methodology is to use data while developing the model, i.e. to use an inductive method. Unlike the previous method it ensures that the model is at least consistent with the given data. It is still possible to test models developed in this way by validation on an independent data set. Inductive methodology is a way of tightening the link between the development of theory and the testing of theory to a single stage.

Machine learning methods can be regarded as formalisation of human leaning, one form of learning being the acquisition of scientific knowledge about the world. Formal learning theory has been used as a precise model of scientific practise, and usefully applied to many machine learning problems. In addition, most machine learning methods rely on an inductive methodology where the system generalises from examples or cases of the concept to be learned. Thus application of machine learning could do a great deal to clarify methodological problems besetting ecology.

**Special characteristics of ecology**

Another idea is that some special property of the field of ecology are responsible for the prediction explanation problem. One such property is the prevalence of history-dependant effects. The ancient history of events due to evolution, and the immediate consequences of human actions, affects the present diversity and integrity of natural systems. History dependant events lead to a fragmentation of the problems in ecology, and block the formation of general laws.

"... the ecological sciences may remain a recipe book for handling 'unique anecdotal management problems'." (Wiegleb 1989).

The idea has been put forward that general laws are necessary for theory and that these laws are largely absent from ecology. The argument is that the complexity and heterogeneity of the units of ecology, organisms, blocks the formation of laws of the form found in the physical sciences. This view has been rebutted as a over-simplification of the issues concerning the relationship of laws to the physical world (Macklin and Macklin, 1969).

Another difficulty is the modelling of the sheer size and complexity of the object of study, i.e. global systems.

"I believe that much of the problem has to do with the complexity of the real processes of nature and our desire to find simple patterns." (Hall 1988).
One well known example was the International Biological Program (IBP) whose aim was to monitor the effects of pollution in the biosphere (van Dyne 1972). To do this a model of the energy flows in all the major ecosystems (terrestrial and aquatic) was to be developed. The project was too ambitious and achieved only limited success, probably because ecosystems are too complex to model in this detailed way, and energy flows are too crude a measure to represent the complexity of many species interactions and their behaviours. The development of global models is again becoming topical with the realisation of the significance of elevated CO₂ levels in the atmosphere. The warnings to prospective global climate modellers have been sounded.

"... no matter how much effort we exert to understand physiological responses to enhanced carbon dioxide, we cannot possibly answer the questions posed by policy makers because of the complexity of applying even our limited knowledge at the population and community levels." (Gross 1989).

Another special characteristic of ecology is the amount of random variation or "noise" in the environment. It is generally believed that noise can to some extent be accounted for by increasing the amount of information in the model, either by increasing the amount of data analysed, or the number of variables taken into account. In fact, as stated earlier, complex models have a poor record of prediction.

Noise can be taken into account with stochastic or probabilistic models. Unfortunately stochastic models are difficult and often impossible to solve (intractable), inhibiting their general application. While stochastic equations are more difficult to solve, the advantages of incorporating a realistic estimate of uncertainty in the accuracy of the predictions are great. Monte-Carlo simulation methods can be used to study the behaviour of the models, although this only allows us to quantify the noise in the data, not to reduce it.

Clearly the characteristics of ecology such as the history dependent effects, complexity and noise need to taken into account in any modelling system applied to ecology. While modelling methods may ignore some of these aspects in particular situations, any general theory of modelling in ecology requires a clear idea of the effect of these general characteristics of the domain. Other general characteristics of ecology relevant to ecology could be pursued in a modelling-based research program into prediction and explanation. For example, it would be useful to establish the relative magnitude of single versus multiple interactions, or general effects versus specific effects.

Knowledge of the characteristics of an application area is a form of 'high-level' knowledge, called background knowledge in machine learning, used to assist in learning. Certain forms of high-level knowledge are necessary for learning to occur. Take a simple example of the knowledge that data is noisy, or contains errors. An algorithm without that knowledge that attempts to find hypotheses consistent with all
the data will fail on noisy data as the data will inevitably contain contradictions. A machine learning algorithm with knowledge that the data is noisy may adopt a strategy of examining hypotheses with a certain allowable number of contradictions, and arrive at a set of hypotheses that includes both true and false hypotheses. Although this is a simple example, it illustrates that background knowledge is necessary to perform some learning tasks. Explicit discovery and utilisation of background knowledge could greatly assist in the development of efficient and capable modelling systems.

**Philosophical naivety**

In times of conceptual crisis, people tend to turn to philosophers for a way out, to show perhaps that the dilemma is the result of a simple misunderstanding, or to justify a compromise situation. One could redefine the purposes of theory and maintain that theory does not have to be predictive.

"It is not the purpose of theory to describe what has happened in a particular instance." (Lewontin 1968)

The philosophy of science has developed a number of approaches; many have been influential to science in general. However the issues are far from resolved, and the development of a philosophy of science based on modelling is an active area of inquiry (Da Costa 1990). While it may be true that,...

"...many of the problems and controversies in ecology stem from philosophical considerations often only dimly perceived." (Caswell 1988),

... it is necessary for ecological modellers to throw light on the problems in their particular area rather than wait for philosophers to solve them. As the terms 'prediction' and 'explanation' are abstract concepts a considerable effort must be expended in understanding and defining them.

Artificial intelligence and machine learning can be seen as philosophical explication turned into computer programs (Glymour 1988b). This follows from the fundamental assertion of AI, that cognition is symbol processing and symbol processors are (or can be) as cognisant as humans are. Whether we ascribe to this view or not the writing of programs can be a powerful method of testing our philosophical intuitions.

**Cultural behaviours**

A range of human behavioural problems perceived may contribute to the problems of prediction and explanation in ecology. It is generally believed that communication between mathematical and field ecologists can be difficult. A
modeller benefits from practical experience by knowing the questions of interest to empirical ecologists or land managers. Similarly empirical ecologists may be assisted by concerns beyond results of immediate relevance by speculative investigations. Communicating mathematically based theories to non-technical people is challenging, and particularly useful for convincing others of the relevance of our work. Yet there are reasons why a scientist may choose to pursue purely theoretical concerns.

"... most of us realise there is a prestige ladder in most disciplines and departments where there often is a premium placed on the ability to do mathematical analysis." (Hall 1988).

Mathematical ability is not sufficient for an ecological modeller however, as many empirical scientists would agree that modellers should ...

"... stop developing biological nonsense with mathematical certainty." (Slobodkin 1974).

The difficulties of developing explanatory models that are mathematically tractable, that predict in a noisy, complex and sometimes chaotic system, and of communicating this technical material to nontechnical people while integrating many fields, scales and methods, can seem too great. The magnitude of the task can lead to a retreat into isolated existences in specialised fields. Yet mathematical ecologists must be ultimately concerned with developing models that both predict and explain in ways that the interested empirical researchers can understand and test. Machine learning presents opportunities for assisting in the communication of theories by developing models of reality that are more comprehensible to non-technical audiences, e.g. qualitative models and formal language representations.

**Theoretical constraints on satisfying multiple goals**

Finally, a reason for the problems of achieving prediction and explanation in the same model is theoretical constraints on achieving the two goals simultaneously. It has been demonstrated theoretically that in a sufficiently complex system no single model can simultaneously achieve generality, realism and precision (Levins 1970). Differing models emphasise different aspects of the three attributes. For example, the Lotka-Volterra equations of population growth are general models, spatial simulations of populations are developed with realism the main priority while a polynomial function fit to data on population fluctuations maximises precision. Since there are no universally adopted models the theory of population dynamics must be expressed as a cluster of models with a complex interrelationship.

Do the trade-offs that occur between generality, realism and precision, also occur between prediction and explanation? Each of prediction and explanation represent goals. Hence a model is required to satisfy a number of different goals. In ecology we do not know and cannot represent the nature of the underlying reality that could potentially satisfy all our goals of modelling. This is often because computational
and mathematical tractability, and observational limitations are so limiting as to block the development of realistic models. Thus a large number of acceptable models are possible, each satisfying the goals of the model to varying degrees. Forms of machine learning that can generate a number of possible models and make the goals of the modelling process explicit are more likely to give insight into the problems of modelling with multiple goals.

**A research program using machine learning**

While the prediction/explanation problem could be attributed to simple human limitations, the impression that prediction and explanation are exclusive goals remains. That the two should be inconsistent seems a strange property of the field of ecology. One would think that the existence of a reality would guarantee the existence of a model of that reality that both predicts and explains, by virtue of being a correct model, and the consistency of that reality with the rest of the world as we know it. From a scientific point of view, the existence of surprising contradictions is a promising prospect for inquiry, perhaps being the surface expression of an as yet undiscovered body of knowledge.

The preceding section has indicated potential ways that machine learning could ameliorate possible causes of inadequate modelling performance. The following section sets out the basis of a research program for investigating the problem based on machine learning. As a first step we need to gather empirical information on the aspects of the problem as they occur in a number of current machine learning techniques. Reliance on the literature is a poor substitute for empirical data in an area as abstract as modelling methodologies.

Given this preliminary understanding the problem can be formulated as one of methodology. Does there exist a system $x$ with property $y$ in application area $z$? The existence of a system would be shown by demonstration of a system $x$ that developed predictive and explanatory models. While the demonstration of such a system is a solution that yields a single bit of information, i.e. that prediction and explanation is possible in the application area, the system also presents a method for the achievement of the goal. The existence of a solution does not imply that it is the only solution, the best solution, or that the solution that should be generally adopted by modellers.

One criticism that could be made of any system that fails to predict and explain is that the failure is simply due to inadequate modelling methodology. Factors affecting the efficacy of modelling need to be controlled. These factors include inadequate validation of models, lack of communication between the modeller and empirical researchers, and philosophical naivety. These factors may be controlled as follows.
1. Models fail because of lack of adequate validation on real world data. Machine learning systems can ensure a close link to the real world by developing models directly from data. It would however still be necessary to adopt the modelling cycle: alternating development and testing of systems, with each new system addressing the limitations discovered during testing.

2. Problems in mathematical modelling are a reflection of the rift between mathematical and field ecologist, contributing to difficulties in integrating the mathematics of the model with ecological theory. This factor could be controlled by as far as possible, developing and testing systems in association with practising field scientists, providing valuable guidance into the development of appropriate representations.

3. Problems in ecology are due to misunderstandings, or philosophical naivety. This criticism could be avoided by conducting a thorough analysis of the concepts of prediction and explanation, and formalising appropriate definitions.

The possible reasons for the prediction/explanation problem could be addressed by experimentation on the predictive and explanatory system. Such a system would help to address the following hypotheses.

1. Ecological phenomena are difficult to model because of the lack of knowledge of laws and processes, and range of possible behaviours. The implicit conclusion of this statement is that we need more research before we can construct good models. This view could be refuted by developing predictive models in domains where laws play little role and processes are poorly understood, such as animal response to the environment, and the diagnosis of diseases.

2. Models fail because they are not complex enough, or do not incorporate enough factors. This view could be refuted by adopting an approach using simple qualitative models. For example, whereas differential equations model smooth changes in the system, qualitative models represent processes as state transitions. Qualitative models also have the advantage of providing a means of analysing categorical data not well handled by quantitative models.

3. Are there fundamental theoretical difficulties in developing models that both predict and explain in ecology? The existence of demonstration of the theoretical restrictions on a general modelling methodology for developing models would support the hypothesis above. More generally the discovery of interactions between prediction and explanation would contribute to an understanding of problems of achieving both predictive and explanatory models.
Conclusions

Explanation and prediction are central qualities of models that appear to co-occur infrequently in ecological applications. A number of reasons for this have been suggested: inadequate mathematical tools, problems with modelling methodology, special characteristics of the field of ecology, philosophical naivety, cultural behaviour and theoretical limitations on the co-achievement of goals. Machine learning presents a possible development in modelling methodology that could address the problem. A research program is proposed based on experimentation with existing machine learning systems, philosophical inquiry and the development of a machine learning system that supports prediction and explanation.
Section 2. Evaluation of existing methods

One of the main advantages of applying machine learning methods in ecology is their ability to use qualitative data in the input and output variables (Manuscript 3). A form of qualitative model frequently used in machine learning using noisy data is the decision tree. The prediction and explanation ability of a number of decision tree machine learning methods were examined (Manuscript 4). While predictive capacity was better than existing methods, two major explanatory problems were identified, difficulty in interpretation of the decision trees, and instability, the tendency of the algorithms to produce a number of alternative trees under a resampling model development regime.

A Bayesian classifier system was developed and tested in order to address the instability problem (Manuscript 5). This system is based on a simplification of Bayes' rule relating prior expectations and evidence to expected outcomes. The model has a vector structure composed of conditional probabilities relating input to output values. LBS learns through modification of the individual probabilities, either through data or alteration by experts. The structure is fixed by the predefined variables and their values and not determined by the data, ensuring stability. The system performed as well as decision trees on the prediction of wildlife density. However it failed to predict when predicting outcomes where more than one outcome was present. Another limitation was the capacity to explain, being limited to tracing the effect of inclusion of evidence. Both of these limitations were due to the rigid structure of the system. On the other hand the rigid structure supported rapid, organised development of predictive expert systems in practical applications.
MODELLING OF QUALITATIVE DATA VIA MACHINE LEARNING

D.R.B. Stockwell
Ecosystem Dynamics, Australian National University, Canberra, ACT.

1. INTRODUCTION

The use of qualitative data is necessary in many disciplines - and is particularly the case in ecology where taxa are measured qualitatively and natural assemblages (e.g., vegetation types) are often pre-classified. Assemblages are stable states of an otherwise complex and dynamic system and pre-classification of quantitative data provides an enormous degree of data compression that greatly assists subsequent analysis and interpretation.

Unfortunately, qualitative variables present difficulties for modellers. Most models have continuous, arithmetic relationships between the independent variables - assumptions not necessarily consistent with the use of qualitative variables. An alternative approach to statistical modelling is to develop high-level descriptions or knowledge-based models. This study demonstrates that useful, predictive models of a complex system can be developed using a particular class of knowledge-based model, the decision tree.

High-level models based on production systems, or expert systems, have been widely used to model complex systems [1]. But the power of these systems is greatly increased when they can be developed rigourously from data sets. As yet, few applications of machine learning to the analysis of data from complex biological systems have been described. Rule induction algorithms have been applied to a few biological problems: INDUCE [2] for cancer cell discrimination, CART [3] for diagnosis in cardiology, and the diagnosis of diseases in soybeans [4]. In this study, machine induction of decision trees by two different algorithms ID3 [5] and k-DL [6] is used to predict and model the density of a tree dwelling species, the Greater Glider, Petauroides volans from forest inventory data.

2. MACHINE LEARNING

Induction of decision trees is a form of concept learning. A concept is an implicit definition of set membership. For instance on examining the study area we might say site x has high glider density if x has a high stand condition and a low slope. Any site is defined as belonging to the set of sites with high glider density if it has high stand condition and low slope. Implicit definition of set membership (as opposed to explicitly listing all sites with high glider density) provides a way of predicting if the glider density is high at another nearby site.

Machine learning tasks require the objects of learning and the concepts being learnt to be defined. The task in inductive learning is to find a concept in the class of decision trees, which classifies pre-classified objects into their correct categories.
2.1 Objects of learning

One common form of representing objects is as a vector or ordered list of values, i.e. \((a_1, a_2, ..., a_n)\) where \(a_i \in A_i\), \(a_2 \in A_2\), ..., \(a_n \in A_n\) and each \(A_i\) is an attribute, or a variable which takes qualitative values.

Both measurement (independent) and outcome (dependent) variables in the data set (Table 1) take qualitative values. In this study objects of learning or training examples are a random subset of the total data set in order to test machine learning as a tool for analysis of data. While the examples are often specially prepared to ensure accuracy, there is evidence that decision trees induced from noisy sets of training examples predict better on the original data than training sets of high fidelity [5].

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>GD</th>
<th>D</th>
<th>ST</th>
<th>SC</th>
<th>SQ</th>
<th>FN</th>
<th>S</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>36</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>27</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>45</td>
<td>39</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>52</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>53</td>
<td>21</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400.</td>
<td>42</td>
<td>28</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

X - x co-ordinate; Y - y co-ordinate; GD - Glider density: 0 outside study area, 1 low (0.5-1.0 G.G./ha), 2 medium (1.0 G.G./ha), 3 high (1.1 G.G./ha); D - Development: 1 not a road corridor, 2 road corridor, 3 pine plantation; ST - Stream corridor: 0 not a stream corridor, 1 stream corridor; SC - Stand Condition: 0 outside study area, 1 rock, 2 nil or regeneration, 3 low (0-50 m3/ha sawlogs or 50-150 m3/ha pulp), 4 medium (0-20 m3/ha sawlogs or 50-100 m3/ha pulp), 5 high (0-40 m3/ha sawlogs or 100-150 m3/ha pulp); SQ - Site Quality: 0 outside study area, 1 low (attain sawlog size in 50-60 years), 2 medium (attain sawlog size in 40 years), 3 high (attain sawlog size in 30 years); FN - Floristic Nutrients: 0 outside study area, 1 low (Veg. types 112, 114, 113), 2 medium (Veg. types 152, 154, 155), 3 high (Veg. types 111, 152, 156, 158, 159); S - Slope: 1 flat (0-10%), 2 moderate (10-30%), 3 steep (30%); E - Erosion: 1 stable, 2 medium, 3 highly.

TABLE 1. The contents of the data set, composed of 400 records (20x20) with 10 fields.

2.2 Concepts to be learned

A decision tree is an ordered sequence of decisions about the values of measurement variables which determine the value of the outcome variable, e.g. does \(A_1 = a_1\)? Each decision leads to a partitioning of the data set into non-intersecting subsets, one containing affirmative examples and the other containing negative examples.

Decision trees can be rewritten as sets of if-then conditionals where the precondition of each rule (if...) is a path through the tree and the action (then...) is the inferred value of the outcome. Production rules are an easy to interpret and widely used in knowledge-based systems. The decision trees of the ID3 algorithm are equivalent to nested case statements found in the Pascal language whereas decision lists \((k-DL)\) are equivalent to nested if-then-else statements.

Typically ID3's trees are bushy (also see Figure 1), with many branches (possible decisions) leading from each node (or measurement variable) depending on the number of possible values of the measurement variable. In contrast, \(k-DL\) trees are linear with a branch for the examples for which the hypothesis is true
EXISTING METHODS

(if-then) and another branch for the remaining examples (else ...); e.g.

**ID3** - Case SC of

0: GD=1;
1: Case FN of
   1: GD=2;
   2: GD=3;
2: Case S of

**k-DL** - if SC=0 then GD=0 else
   if SQ=2 and FN=2 then GD=1 else
   if S=1 then GD=3 else
   if SC=3 and S=2 then GB=3 else

The task of the learner, or induction algorithm, is to find decisions concerning the measurements which lead to an efficient and correct prediction of the outcome. The algorithms examined in this study achieve this with a generate-and-test algorithm: an hypotheses is generated and tested against the training set. In ID3 an hypothesis is a decision to partition on a particular measurement variable. In k-DL an hypothesis is an implication from a conjunction of measurements, \( A_1=a_1 \) and \( A_2=a_2 \) \( \ldots \) \( A_k=a_k \Rightarrow A_0=a_0 \) where \( k \) is the number of terms in the conjunction. The best hypothesis according to some criterion is then selected, and the training set is partitioned. The process repeated on each of the subsets until all examples in a subset have the same outcome or no hypotheses meet the criteria.

The major confounding factor is conflict - examples with identical measurements and different outcomes (e.g. Table I, examples 2 and 3). It is clearly impossible to differentiate between these examples using the information available which leads to a maximum level of predictive accuracy by any means.

Accepting that the performance, or percentage of correctly classified examples on the training set, will not be 100% due to conflicts, another confounding factor is the size of the training set - is it an adequate sample? Decision trees can be produced from randomly selected training sets of different size and then applied to an independent test set. If the test set is adequate the performance should approach a maximum constant value when the size of the training set is less than the test set.

Computational speed is a limiting factor with generate-and-test algorithms. While reducing the number of objects of learning (the training set) decreases computation time, another simplification is to reduce the size of the concept class. Decision lists produce an concept class which increases as the number of conjunctions in the decision increases, i.e. \( 1\text{-DL} < 2\text{-DL} < \ldots < 7\text{-DL} \). As decision in class \( 1\text{-DL} \) is the value of a single variable, \( A_1=a_1 \), the number of hypotheses is equal to the number of values of measurement variables multiplied by the number of outcome values, i.e. \( 4(3+2+5+4+3+3) = 92 \) possible decisions at each partition. In class \( 7\text{-DL} \), a decision includes all combinations of values of variables; \( A_1=a_1,A_2=a_2,\ldots,A_7=a_7 \). The number of possible decisions at each partition of the \( 7\text{-DL} \) tree (\( k=7 \)), assuming three values per measurement variable (\( n=3 \)), is then; \( 4\sum_{i=1}^{7}n^i \cdot \binom{3}{i} = 4.7!(3/1!6/2!5!+3^2/2!5!+\ldots+3^7/7!0!) = 65,532 \).

### 2.3 Acceptance of hypotheses

A criterion for accepting hypotheses is needed for two reasons: to choose between alternative hypotheses and to determine whether an hypothesis would improve the classification. A statistic based on the Chi squared test has been suggested [5] and was modified for use in \( k\text{-DL} \). The examples which affirm a decision are regarded as a sample of a population, namely the training set. The statistic;
MODELLING QUALITATIVE DATA

\[ \chi^2 = \frac{(p-p')^2}{p'} - \frac{(n-n')^2}{n'} \]

where;  
\( p' \) is the expected frequency of the value of the outcome variable in the training set  
\( p \) is the frequency of the value of the outcome variable, or the number correctly classified in the sample  
\( n' \) is the expected frequency of the remaining values in the test set and,  
\( n \) is the frequency of the remaining values of the outcome variable, or conflicts, in the sample.

approximates a Chi squared distribution with one degree of freedom. The difference of the terms is used to provide a two-sided test, eliminating hypotheses of the form, if \( A_1 = a_1 \) then \( A_2 = a_2 \), as explicit negation was not used in this implementation. If the sample chosen by a particular hypothesis differs significantly from the population (the training set) \( \chi^2 \) will be large. The structure of decision trees was not found to be very sensitive to the actual \( \chi^2 \) value, and \( \chi^2 > 2 \) (or \( \alpha < 0.2 \)) was used as the criterion for accepting hypotheses.

3. THE EXAMPLE

The decision trees were used to predict the density of the Greater Glider from forest inventory data (Table 1). The examples for learning were 4 hectare quadrats of forest with measurements listed in Table 1. The distribution of the Greater Glider as predicted by the decision trees was compared with the known distribution in the well-surveyed Waratah Creek section of Coolangubra State Forest, New South Wales, Australia. As the database was based on grid-cells, a grid-cell map was prepared from a polygon map for this species, based on surveys carried out between 1981 and 1987 (S.M. Davey and R.P. Kavanagh unpublished data, also [7]).

The measure of the performance of the decision trees was the overall classification accuracy (OCA) - the percentage of grid cells in which population density was correctly estimated. The null hypothesis, that the decision tree does no better than a random assignment (RA) of levels to the matrix, was tested against the OCA. The value of RA was assumed to have a binomial distribution after Thomas and Allcock [8] and was calculated by the equation;

\[ \sum_{i=0}^{3} \frac{E_i O_i}{N} + 1.96 s \sqrt{N} \]

where;  
\( E_i \) is the number of cells predicted by decision tree to be at level i,  
\( O_i \) is the actual number of cells at level i (from survey data)  
\( N \) total number of cells and  
\( s \) is the standard deviation, \( \sqrt{NE_i(1-E_i/100))/100.} \)

4. RESULTS AND DISCUSSION

All algorithms stabilised into a tree with similar structure and performance with training sets greater than 50 examples. Representative trees and their performance is shown in Figure 1 and Table 2. All trees classified the test set significantly better than random assignment (\( \alpha < 0.01 \)).
EXISTING METHODS

<table>
<thead>
<tr>
<th></th>
<th>ID3</th>
<th>1-DL</th>
<th>7-DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCA</td>
<td>60%</td>
<td>52%</td>
<td>57%</td>
</tr>
<tr>
<td>RA</td>
<td>30±1%</td>
<td>31±1%</td>
<td>31±1%</td>
</tr>
<tr>
<td>α</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>ntrain</td>
<td>60</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>ntest</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
</tbody>
</table>

OCA: Overall Classification Accuracy (exact)
RA: Percentage of correct classifications expected in random assignment
α: Probability of OCA equalling RA
ntrain: number of examples used to generate tree
ntest: number of examples used to test tree

TABLE 2. Comparison of the performance at predicting the population density of the Greater Glider.

The ID3 algorithm developed a complex hierarchical model of the system. Stand Condition was chosen as the most important initial factor in determining the Arboreal Density, while Slope, Site Quality and Floristic Nutrients were of importance at higher levels of Stand Condition.

FIGURE 1. Examples of the trees with the greatest predictive ability as generated each of the algorithms. The ID3 algorithm produces trees which branch on each value of a variable, whereas the decision lists produce trees which branch on a value (1-DL) or a conjunction of values (7-DL) of variables.

The decision lists 1-DL and 7-DL present a list of statistically valid hypotheses about the data set. The 7-DL predicted with greater accuracy, probably due to the more detailed conceptual basis, although 1-DL performed adequately despite the greater simplicity of the tree. In general the decisions accepted in 7-DL
contained no more than three conjunctions, i.e. 3-DL would have been sufficient. A clearer interpretation of the 1-DL list is achieved by rewriting the ordered hypotheses into its alternative conceptual form.

If Stand Condition is nil then Glider Density is nil
else if Slope is low then Glider Density is high
else if Site Quality is medium then Glider Density is low
else if Floristic Nutrients is medium then Glider Density is medium
else if Development is a road corridor then Glider Density is low

In comparison, rewriting the tree produced by the ID3 algorithm would take 17 rules. The complexity of the ID3 algorithm with higher values stand condition may reflect an increase in complexity of the biological system with greater forest development. However the k-DL models suggest a simpler interpretation based on statistically acceptable hypotheses about the data, which could be related to the biological system.

Good performance (49%) was often achieved with small training sets of 30. The algorithms converged smoothly to a stable value of performance, although trees developed from differing sized training sets were sometimes very different in structure. The class 3-DL would have been adequate for this data, a factor 260 times less computationally expensive. These results indicate useful simplifications to decrease the computation time without decreasing the predictive ability or interpretive clarity.

5. CONCLUSIONS

Decision trees developed on a subset of the data set with each machine learning system gave predictions of the population density of Greater Gliders with high-level concepts which could be interpreted biologically. Although ID3 is fast and generates trees with good predictive ability the forms of hypothesis allowable are restricted. Induction of decision lists gave accurate, and simpler trees which were easily converted to production rules. The results support the view that high-level representations of qualitative data from complex systems can be predictive and informative.

Acknowledgement: The data and assistance supplied by S.M. Davey, R.P. Kavanagh, Department of Forestry, Australian National University and the Forestry Commission of New South Wales is gratefully acknowledged.

References

58
Abstract. Estimation of wildlife density is a major part of forest management. Because ecosystems are complex and frequently poorly understood, few predictive models exist. It would be useful to develop models directly from data sets from particular forests and use these models for prediction and explanation. However, qualitative variables and non-linearities in the data make standard curve-fitting techniques difficult or impossible. This study evaluates machine induction of decision trees to predict density of the greater glider (Petauroides volans) in forests of the southeastern region of New South Wales, Australia. Three different algorithms for the induction of decision trees, CN2, ID3 and CART, were compared with three other models: a linear statistical model, a decision table developed by principal components analysis, and an expert system developed by knowledge acquisition. The machine-learning assisted methods were quicker than knowledge acquisition, gave better predictions than other models, and produced concise instructions for discriminating areas of high and low wildlife density. The study demonstrates the advantages of using machine learning algorithms to assist the development of classification rules for predictive modelling.

Using Induction of Decision Trees to Predict Greater Glider Density

David R.B. Stockwell
Ecosystem Dynamics
Australian National University
Canberra, ACT 2601, Australia

Stuart M. Davey
Department of Forestry
Australian National University
Canberra, ACT 2601, Australia

J. Richard Davis
Division of Water Resources, CSIRO
Canberra, ACT 2601, Australia

Ian R. Noble
Ecosystem Dynamics
Australian National University
Canberra, ACT 2601, Australia

Although expert systems are useful in wildlife management (Starfield and Bleloch 1983, Marcot 1984, Rauscher and Cooney 1986), they are also time consuming to develop by interviews with experts. Machine learning methods may substantially reduce the time taken to develop expert systems by inducing the rules from cases or pre-classified data items. One well-studied method, induction of decision trees (or classification rules), has been shown to be useful in modelling complex data sets (Breiman et al. 1984). A number of algorithms have been published. The ID3 algorithm (Quinlan 1986) is frequently included in expert system development shells to aid rule formation. What are the strengths and weaknesses of these algorithms in a practical application and how do they compare with conventional methods of model development?

Until now, most machine learning algorithms have been designed to produce models in error-free domains. Empirical studies investigating the behavior of ID3 in noisy, real-world domains have shown high predictive accuracy (Mingers 1989, Stockwell 1989, Williams 1987). A recent comparative study has shown that models developed by decision tree
induction systems are faster at classifying data than are arithmetic operators (Goodman 1990). This study presents an empirical comparison of the predictive and explanatory capacity of models produced by three practical decision tree induction algorithms: CN2 (Clarke 1989), ID3 (Quinlan 1986) and CART (Breiman et al. 1984). Models developed by more well-known techniques are included for comparison: multiple regression (MR), a decision table (Wyatt 1986) developed with the results of principal components analysis (PCA) (Braithwaite et al. 1983), and an expert system developed by knowledge acquisition (KA) (Stockwell 1987).

History
Forest ecologists believe that the animals most sensitive to forest disturbances, such as timber harvesting, are those dependent on extensive areas of mature, complex forest for their food and shelter. Planning and monitoring of logging and fauna conservation areas requires either accurate census information or a reliable model of the distribution and abundance of these animals. The presence of arboreal marsupials (possums and gliders) is widely recognized as an indicator of healthy, mature forest.

Determining the distribution of a population of possums or gliders is a difficult task. Surveying the population directly is time consuming. The complexity of the forest ecosystem prevents the development of simple and accurate quantitative models. For instance, the population density of Australian possums and gliders has been found to be non-linearly related to forestry inventory variables by both curvilinear or threshold functions (Braithwaite et al. 1983, Braithwaite 1984). Often independent variables are of qualitative form, as ordinal classes (e.g., slope class), types (e.g., vegetation) or indices (e.g., diversity) and unsuitable for use in quantitative models.

Figure 1. The study area was located on the coastal ranges and tableland of southeastern New South Wales, Australia (a,b). The area has been extensively surveyed for arboreal marsupials over seven years, resulting in a detailed knowledge of their distribution and abundance, particularly the greater glider (c).
Stockwell et al.: Decision Trees to Predict Greater Glider Density

This study was based in Coolangubra State Forest, located in southeastern New South Wales of Australia (Fig. 1a, b). The 22,000-ha forest contains both tableland and foothill eucalypt forest and is important to wildlife, especially arboreal marsupials (Kavanagh 1984). The study area, 1600 ha in extent, was located at Waratah Creek in tableland forest on the western portion of the forest (Fig. 1c). It includes the intensively studied wildlife research reserve used by Kavanagh in his studies of arboreal marsupials (Kavanagh 1984, 1987a).

The wildlife species found within the study area have been surveyed by both the Forestry Commission and personnel from the Australian National University. A number of habitat studies have been undertaken (reviewed in Davey and Norton 1989) within or near the Waratah Creek Wildlife Reserve and have shown the presence and density of arboreal marsupials to be related to foliages and soil nutrient levels (Braithwaite et al. 1983, 1984, Braithwaite 1984), structural parameters (Davey 1984) and seasonal tree phenology (Kavanagh 1984, 1987a). These studies show that the correlation between quality of arboreal marsupial habitat and environmental parameters is very complex and dependent upon site potential (as measured by floristic nutrients), stand productivity, stratal distribution, quantity and quality of foliage biomass, stand history, tree species diversity, the availability of winter food resources, and the availability of tree hollows for shelter.

Most of the information on the distribution of arboreal marsupials used for developing and testing the models were the result of regular surveys (Kavanagh 1984, 1987a, 1987b) undertaken in forests surrounding (and within) the wildlife reserve (Fig. 1). Accurate density estimates of greater gliders were available for the areas covered by these regular surveys. Less frequent surveys were undertaken in the northern, eastern, and southeastern portions of the study area, and along Wog Way and the fire trail to Pheasants Peak. Density estimates for these portions would be less accurate than those for areas that had been regularly surveyed.

A forest rating scheme was developed by Wyatt (1986) and referred to as PCA. It is based on tree species and their related floristic nutrient value; the higher the composite nutrient value, the more likely the presence of arboreal marsupials (data after Braithwaite et al. 1983, 1984). Because each vegetation type has a constant composition of species, levels of greater gliders were assigned to each of the 11 vegetation types occurring in the area. The levels used by Wyatt were adopted for the present study: nil (0 animals/ha), low (less than 0.5 animals/ha), medium (0.5 to 1.0 animals/ha) and high (greater than 1.0 animals/ha) arboreal density.

Predictor variables were extracted from a forest inventory database for Coolangubra State Forest and the surrounding land. The database was developed for research and teaching purposes by members of the Forestry Department at the Australian National University (Turner 1986, Wyatt 1986). It covers a land area of 20 km (north-south) and 28 km (east-west), and is composed of 14000 grid-cells, each 4 ha in area with one record per grid-cell. Each record contains typical inventory data for forest management, such as the cadastre, topography, roads, streams, slope, erosion class, soil type, geology, forest type, stand growth potential, and merchantability. The mapped area encompasses the Coolangubra State Forest (22,000 ha). This database is regarded as very reliable on the western side where the study area is situated, becoming gradually less reliable to the east due to fewer surveys being undertaken in that area.

The expert system, referred to as KA, was developed through a series of interviews with a wildlife expert (Stockwell 1987), aided by an expert system shell (GEM) specifically designed for geographic applications (Davis and Nanninga 1984). The structure of the expert system KA is shown in Figure 2. A thematic map of the distribution of the greater glider over a 800-ha area (Fig. 1c) was prepared to test the expert system. Estimates of greater glider density were based on the intensive surveys

Figure 2. The dependency diagram for the expert's model. The value of a variable depends on the value of variables below. Each arrow represents a number of rules.
carried out between 1981 and 1987 (Davey and Kavanagh, unpublished data). The surveys were conducted by spotlighting along transects, either on foot or by vehicle, and correction for weather effects were applied where necessary. The polygon map was converted into a grid-cell map to conform with the forestry inventory database.

Although Stockwell (1987) developed an expert system for numerous marsupials occurring in southeastern Australia, this paper will confine itself to modelling one species, the greater glider, Petauroides volans, regarded as an indicator of the suite of arboreal marsupials present in the area. The polygon map of the density of the greater glider developed to test the expert system in the above study was used for the present study. The models PCA and KA are included for comparison with the machine learning models.

Methods

A brief description of the methods used for developing two of the models (PCA and MR) has been given above. A full description is given in Wyatt (PCA, 1986) and Stockwell (KA, 1987). Because methods used to develop the decision tree models (ID3, CN2, and CART) have been described elsewhere, only a brief outline is given here. The methods of comparison are described below.

Because a distinct model is produced by the application of each algorithm, the abbreviations of each method are used to identify the algorithm and the resulting model. The methods of model development were:

- multiple regression to produce a linear model (MR),
- principal components analysis leading to a decision table (POA),
- knowledge acquisition leading to a rule base describing the expert's model of the problem (KA),
- machine induction of decision trees by three different algorithms (CN2, ID3, and CART).

Two additional models (MGC and MSC) are included for validating the use of decision tree models. They provide upper and lower points of reference for the accuracy and complexity of the decision trees and are not proposed as algorithms applicable to general modelling tasks.

All models can be thought of as rules for classifying data into classes. The main difference between the models is the method of development; i.e., hybrid statistical approach (MR and PCA), interviews with an expert (KA), and variations on heuristic search algorithms (ID3, CN2, CART).

The models developed by conventional methods (MR, PCA, and KA) are compared with the decision trees produced by the induction algorithms. Comparison between the induction algorithms indicates the strengths and weaknesses of each algorithm.

Decision Trees. A decision tree can be described as a classification rule for determining the class of any object from the values of its attributes (Quinlan 1986). For example, IF (ST=1 AND FN=3) OR S=1 THEN GD=3 is a classification rule for determining sites of high greater glider density. The attributes may be represented by real-values, integers, or symbols. As applied to our problem, an object is the grid-cell represented by a record in the database, classes represent the greater glider density (levels 0 to 3) as determined by the survey data, and attributes are the forest inventory variables (Table 1).

Multiple regression was used to create a simple classification rule. A linear model was fitted directly to the numeric coding of the data using the stepwise multiple-regression technique. In order to obtain a qualitative

### Table 1. The contents of the forestry inventory database used to predict greater glider density. Composed of 14000 records (140 x 100), each recording the value of an variable in a quadrant 200 m square.

<table>
<thead>
<tr>
<th>D</th>
<th>ST</th>
<th>SC</th>
<th>SQ</th>
<th>FN</th>
<th>S</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **D** Development: 1 not a road corridor, 2 road corridor, 3 pine plantation.
- **ST** Stream corridor: 0 not a stream corridor, 1 stream corridor.
- **SC** Stand condition: 0 outside study area, 1 rock, 2 nil or regeneration, 3 low (0 m²/ha of sawlogs or 0-50 m²/ha pulp), 4 medium (0-20 m²/ha of sawlogs or 50-100 m²/ha pulp), 5 high (0-40 m²/ha of sawlogs or 100-150 m²/ha pulp).
- **SQ** Site quality: 0 outside study area, 1 low (attain sawlog size in 50-60 years), 2 medium (attain sawlog size in 40 years), 3 high (attain sawlog size in 30 years).
- **FN** Floristic nutrients: 0 outside study area, 1 low (Veg. types 112, 114, 113), 2 medium (Veg. types 152, 154, 155), 3 high (Veg. types 111, 152, 156, 158, 159).
- **S** Slope: 1 flat (0-10%), 2 moderate(10-30%), 3 steep (30%).
- **E** Erosion: 1 stable, 2 medium, 3 high.
result for the density, the result of the linear model was rounded to the nearest integer and used as the estimate of density class. The resulting model can be expressed as:

$$MR \approx \text{round} (1.10SC + 0.31FN - 0.64S + 1.10), r = 0.62,$$

where \text{round} is the function rounding real numbers to their nearest integer. This model is a classification rule where the decision about the value of the class is based on a calculation with the variables.

The general strategy of the ID3 algorithm is to find a variable that best divides the set of objects into homogeneous classes. The choice is usually based on a classic measure of information or entropy, and is an indication of the discriminatory power of that variable. The set of data objects is then divided into subsets with the same value relative to the chosen variable. Each subset is either divided again on the next most discriminatory variable or assigned a predicted class value. The number of choices in each stage of the search strategy is limited to the number of variables. Typically, ID3 produces a bushy tree (many branches) as shown in Figure 3.

Stopping criteria can be based on:
1) homogeneity - when all objects in the subset belong to the same class or are relatively homogeneous according to a test based on the frequency of particular values, or 2) structure - there are no variables left to divide the subset on or a pre-determined maximum number of variables has been used (fixed tree depth). The algorithm then forms a leaf by assigning the most frequent class to the final subset.

Most recent applications of the algorithm use a form of pruning to control over-fitting. After growing a tree as large as possible, branches are removed that do not contribute to or diminish the accuracy of the tree. The accuracy of a leaf can be assessed on the training set or on an other independent test set not used for building the initial tree. In this study, pruning was applied by growing a number of trees to a specified depth and testing them on an independent test set. Model accuracy as inferred from the test set reaches a maximum at a depth of two or three. To enable comparison with CN2, trees of, at most, depth two were used.

The CN2 algorithm (Clark 1989) produces decision lists, a linear form of decision tree (Fig. 4). The algorithm is divisive, like ID3, but instead of dividing on all the values of a single variable, it searches for a fixed maximum number (in this case two) of conjunctions of values of variables (e.g., $SQ = 2$ AND $FN = 2$) to use as a predictor of a class. It has been shown with a similar algorithm on this data set that two conjunctions are sufficient for maximum accuracy (Stockwell 1989). A limit to the number of conjunctions is necessary to enable the search to be conducted in a reasonable time. The approximate number of choices at each stage of the search strategy is $(nm)^2$, where $n$ is the number of variables and $m$ is the average number of values in each variable.

The algorithm conducts a general to specific search for the best conjunction by finding best single conjunctions, forming double conjunctions from those singles, and so on. The basic algorithm can be viewed as

**Figure 3.** A decision tree produced by ID3. The set of objects is divided into subsets with the same value with respect to a chosen variable.

**Figure 4.** A decision tree produced by CN2. A conjunction of values of variables is used to discriminate a class from the rest of the set of objects.
several hill-climbing searches in parallel. An information measure guides the search, and a likelihood ratio statistic establishes the significance of the chosen conjunction.

After finding the best conjunction, a leaf is formed, the objects forming the leaf are removed from the total set of objects, and the algorithm is applied to the objects remaining. Eventually no statistically significant complex can be found and a leaf is formed with the remaining objects. The algorithm seems to produce simple trees with a moderate degree of searching (Fig. 4).

The CART algorithm (Breiman et al. 1984) uses a probability criterion based on Bayes’ Rule for splitting the set of objects. It first constructs a number of extensive bifurcating (two-branched) trees. A number of values of a single variable can be used to divide the set of objects. The trees are then pruned by testing on an independent test set. CART can be seen as an extensive search procedure for internally disjunctive concepts (e.g., $SC=3 \text{ OR } 4$). Most of the trees produced were linear but more arbitrary branching is possible (Fig. 5a, b, c). The approximate number of choices at each stage of the strategy is $n^m$, where $n$ is the number of variables and $m$ is the average number of values for each variable.

All decision tree models have a complexity of structure somewhere between two extreme models. The maximally general classification (MGC) attributes all data to a single most commonly occurring class (e.g., IF (null) THEN $GD=2$). The MGC is the simplest model that could be proposed. The maximally specific classification (MSC) attributes the most commonly occurring class to each distinct configuration of values of variables in the data (e.g., IF $D=1$ AND $ST=0$ AND $SC=4$ AND $SO=3$ AND $FN=2$ AND $S=2$ AND $E=3$ THEN $GD=1$), ...). The MSC is the most complex model that could be proposed. These models are a reference for determining how much simplification of the data set (data compression) has been achieved by the decision tree algorithms, and the inherent classification error in the data set.

The CART algorithm was written in FORTRAN and supplied compiled by the author of the algorithm. All other modelling (implementations of ID3 and CN2, MSC, and MGC) and testing of the algorithms was performed by the senior author using Turbo Prolog.

Application and Comparison of the Results
Conflicts, or errors due to noise or misclassification of the data objects, prevent decision trees from classifying all cases correctly. Conflicts occur
if objects of different classes have identical values for each variable. It is impossible to differentiate between such cases using the information available.

One cannot reliably estimate the accuracy of prediction for any model developed and tested on the same data set, because over-fitting can result in poor accuracy when the model is applied to independent data sets. To compare decision tree induction methods, an estimate of the variance of the predictive accuracy is needed. The data set was randomly split six times into two sets of 200 cases, a training set used for developing the models and a test set for estimating the accuracy of the model. All measures of accuracy and complexity were averaged over the six random splits.

The models PCA and KA were developed independently of the data used in this study. They were tested on the whole data set to obtain the most accurate estimate of their predictive capacity. The linear regression model, MR, following standard statistical practice, was developed on all the available data. Its accuracy was determined by re-classifying the original data with the model.

Three main requirements of a model useful for management of wildlife populations are:

- **Accuracy.** The model should predict the correct outcome on independent data.
- **Comprehensibility.** The model should be a simple method of classifying objects.
- **Efficiency.** The model should be quick to develop and apply to new data sets.

The main measure of the accuracy of the prediction of the models was the percentage of cases correctly classified. The null hypothesis that the decision tree does no better than a random assignment of levels to each grid-cell in the survey map at the proportions of levels in the model was adopted. The classification accuracy of random assignment was, assumed to have a binomial distribution (Thomas and Allcock 1984). The expected classification accuracy in random assignment and the standard error of the estimate, \( e \), can then be determined from \( p \):

\[
p = \frac{\sum p_i N_i}{N}
\]

\[
e = \sqrt{\frac{pq}{N}}
\]

where \( N \) is the total number of cases, \( N_i \) is the number of cases at a level \( i \), \( p_i \) is the fraction of cases at a level \( i \), and \( q = 1 - p \).

A Student's \( t \) test for significance of the difference between means was used to compare accuracy among the models. Estimates of the standard deviation for each model could be obtained from the results of six replicated random samples. Inherent classification inaccuracy in the data limits the maximum predictive accuracy that could be achieved. The MSC has the maximum possible predictive accuracy on the training set because it is the most specific classification of the data. Therefore, an indication of the inherent classification inaccuracy is given by the MSC. The MGC indicates the accuracy of a null model where no variable predicts the outcome.

Comprehensibility is often related to measures of information complexity. Classification rules can be decomposed into a number of simpler rules, where a single rule is a leaf in the decision tree. The results can be expressed as a list of rules in a natural language where each branch in the decision tree is a rule. Below is the decomposed natural language rule for the tree in Figure 4.

**IF** Stand Condition is zero then Glider Density is zero, ELSE
**IF** Slope is medium and Floristic Nutrients is low than Glider Density is low, ELSE
**IF** Site Quality is medium and Stand Condition is medium than Glider Density is low, ELSE
Glider Density is medium.

Because the MSC is the most specific model of the data, it is composed of the maximum number of rules. The degree of data compression due to a particular decision tree model is indicated by comparison with the MSC information complexity. In comparison, the MGC, composed of one rule, has a complexity of one.

The similarity of the model to the way the expert thinks about the system is also related to comprehensibility. This notion is difficult to define exactly. In an application, the expert would choose between possible models, based on experience and the results of previous studies. In this way models are differentiated and related to previous work.

When large amounts of data are being processed regularly, efficiency becomes an important issue. An algorithm using the model should be fast and integrate easily with systems used in management planning, such as geographic information systems (GIS) and linear programming. Normally two operations are performed: 1) the induction of the decision tree, and 2) the classification of data with the learned tree. Because implementations differ between algorithms, an exact measure of time is not meaningful; however, the time complexity of induction and classification can be estimated and the algorithms ranked.

A measure of the induction time complexity (TC) for the decision tree algorithms can be estimated from the degree of search required to produce the model. The number of choices at each stage of the search strategy was estimated previously: \( n \) for ID3, \( (n-m)^2 \) for CN2 and \( n^2 \) for CART. Substitution of the values for the number of variables (\( n=7 \)) and the average number of values per variable (\( m=3.4 \)) gives the approximate size of the search space. The time complexity of the non-search algorithms is based on development time so that all the algorithms can be ranked.

A computational measure of classification time complexity (TC) is the maximum number of computations...
required to classify a data item. In regression models, this measure equals the number of arithmetic operations plus one for the final rounding operation. In a decision tree, the maximum number of computations is equal to the depth of the tree.

Results

Accuracy. The classification accuracy on the training set, independent test set, information complexity, and relative levels of induction and classification complexity are given in Table 2. All models except MGC were significantly better at classifying the test set than the expected accuracy due to random assignment (30%) and could be useful for predicting greater glider density.

The consistent decrease in accuracy from the training set (p<sub>train</sub>) to the test set (p<sub>test</sub>) demonstrates that classification accuracy of the decision tree on the training set was not a good indication of the predictive accuracy on the independent test sets. The MGC has a large decrease because the model is too specific and does not classify well on independent data sets.

Table 2. Comparison of the accuracy of the models of population density of the greater glider over the Warataah Creek study area, Coolangubra State Forest.

<table>
<thead>
<tr>
<th>Model</th>
<th>P&lt;sub&gt;train&lt;/sub&gt;</th>
<th>P&lt;sub&gt;test&lt;/sub&gt;</th>
<th>I.C.</th>
<th>TC&lt;sub&gt;i&lt;/sub&gt;</th>
<th>TC&lt;sub&gt;r&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>45%</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>41%</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>KA</td>
<td>48%</td>
<td>29</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>MGC</td>
<td>36.7±0.8%</td>
<td>33.2±1.6%</td>
<td>1.0±0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID3</td>
<td>60.7±1.3%</td>
<td>57.3±2.2%</td>
<td>11.2±0.2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CN2</td>
<td>50.8±2.6%</td>
<td>45.2±2.2%</td>
<td>5.2±0.5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>CART</td>
<td>61.2±3.5%</td>
<td>54.8±3.9%</td>
<td>5.0±0.9</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>MSC</td>
<td>75.8±1.2%</td>
<td>48.3±1.6%</td>
<td>55.7±1.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The predictive accuracy of models was compared to determine the significance of any difference in accuracy. The qualitative result in increasing order was MGC, PCA < MR, CN2, KA, MSC < CART, ID3 < where the symbol α represents a significant increase in accuracy with a probability of type one error of less than α. The decision trees ID3 and CART gave the most accurate predictions of greater glider density (57% and 55%). In contrast, MR and PCA generated the poorest predictions (45% and 41%). The accuracy of the expert system KA and the decision list CN2 were intermediate.

Comparability. The order of increasing information complexity was MGC < MR, CART, CN2 < ID3, PCA < where KA < MSC. The decision tree induction algorithms CART and CN2 produced the simplest of the useful models, having five rules as opposed to 11 in ID3 and 29 in the expert system. All the models compressed the data considerably as the information complexity of MGC (66 rules) shows.

Different models gave different explanations of the dependencies within the data set. The correlative method MR illustrated the major trends in the data; the major variables correlated with greater glider density (SC, S, and FN) and whether the correlations were positive or negative. PCA also found major trends in the form of transformed linear correlations of all the variables, the most significant being interpreted as floristic nutrient levels. The variables favored in the correlative methods were prominent in the decision trees.

Efficiency. Upon substitution for n and m into the search space equations, the order of increasing time complexity of the decision tree induction (TC<sub>i</sub>) was ID3 (7) < CART (74) < CN2 (566). The machine learning algorithms ID3 and CN2 took from one to five minutes, respectively, on an IBM AT. The trees produced by CART were completed in a few seconds on a substantially more powerful machine.

The development time of the algorithms was smallest in the non-search algorithms (MR and PCA) and largest in the expert system. The knowledge...
acquisition bottleneck is well known, and much time was spent organizing the rules into logical groups and checking for repetitions and omissions. Difficulties were experienced with the expert stating his knowledge in the form of rules. Furthermore, much of the published literature relating to the management of arboreal marsupials during forestry operations was not readily transferable into rules. Tables from documents seemed easier to formulate into rules than information presented within the discussion of a paper. For example, Specht's (1981) classification table for forest structure translates directly into a series of rules, whereas Davey's (1984) and Kavanagh's (1984) discussions require the direct involvement of an expert to interpret their material.

The order of classification time complexity was $\text{PCA} < \text{KA}, \text{ID3} < \text{CN2}, \text{CART} < \text{MR}$. The time complexity of the algorithms was smallest in PCA because only one operation—assigning a class to a value of a variable—is needed. The regression-based model MR had six operations—three multiplications, two additions, and a rounding operation. The time complexity of MGC and MSC is not relevant as they are not being proposed as applicable algorithms.

Discussion

The best predictors of wildlife density were the decision trees ID3 and CART. Of these trees, CART was the simplest (information complexity of five against 11). While both these methods could be useful for building models for predicting wildlife density, CART gave the most parsimonious model and ID3 had less induction and classification time complexity. The decision tree CN2 performed no better than the statistical model MR and was also among the algorithms with the highest time complexity.

A number of factors contributed to the lower predictive accuracy of the models MR and PCA. The linear, continuous model is inconsistent with the qualitative nature of the data, leading to errors such as predictions of minus one when the lowest value was zero. The relationship of one variable, slope (S), appeared to be non-linear with respect to glider density.

The model PCA is based on only one of the three main predictive variables (FN of SC and S). Constraining the model to a single variable leads to less discriminatory capacity than models with more variables. The models PCA and MR were easy to compute (low TC) and yielded information on the vegetation types associated with glider densities (PCA) and the overall response of greater glider density to stand condition, and floristic nutrients slope (MR). The conventional models express general trends (e.g., glider density is positively related to stand condition), whereas the decision trees are more specific (e.g., if the stand condition is zero then glider density is zero).

Although the expert system showed intermediate performance, the development time and number of rules were greater than the decision trees. If the primary aim of developing a model is prediction, the extended length of time (months) for development of an expert system by knowledge acquisition through interviews with the expert could only be justified by absence of data. Even then the time could be well spent gathering a training set of cases for the machine learning system to analyze. The decision tree generated by a system could then be used in rule-based expert systems for predictions or incorporated into efficient code in a GIS.

Simpler models would be preferred by managers because simple statements are easier to apply and communicate to people. By this criterion, CART would be preferred for predicting wildlife density. However, there was a great deal of variability in the structure of the six trees produced by CART, with trees ranging from two leaves based only on slope to a non-linear tree with eight leaves (Fig. 5a, b, c). Structural variability presents the problem of choosing between trees that are all valid on the grounds of predictive accuracy.

The decision tree generated by ID3 had low computational time complexity and produced consistent results. Stand condition (SC) was always selected as the best initial discriminator of greater glider density in all trees. Slope (S) was the next best discriminator at levels of medium to high forest development.

These results are consistent with the findings of Braithwaite et al. (1983), who emphasized the importance of retaining areas of low slope in forested areas for preserving arboreal marsupial populations, largely because vegetation types that have a high nutrient index are related to high soil nutrient levels, which is, in turn, a function of slope. The importance of low slope to glider density is a feature of the Eden region due to the accumulation of nutrients through run-off into areas of low slope.

Conclusions

Induction of decision trees is an accurate, informative and efficient method for predicting wildlife density when basic forest inventory data are available. While we cannot generalize from a single applied study such as this, the findings are consistent with the conclusions of previous studies in other applications. The method could be useful in domains where linear statistical models perform poorly due to non-linearity and qualitative independent variables. Decision trees may be quicker to develop than expert systems and quicker to apply to the classification of large amounts of data than arithmetic models.

A number of directions for research are suggested. Reliable induction methods are important, and a greater
theoretical understanding of search strategy and significance tests for noisy domains is needed. More empirical studies would allow generalization from specific applications and lead to strategies, tests, and methodologies that succeed in all situations. The algorithms need to be further developed for domains with many continuous predictor variables and integrated with user interfaces.

Quinlan (1986) concluded that ID3 leaves much to be desired in terms of representation, and the results of this study suggest that CART or CN2 may lead to simpler representations. While a number of possible decision trees with similar accuracy were produced, both within models and between models there does not seem to be any method of choosing between them. A method that produces all valid models, such as the version space approach (Mitchell 1982), might help to resolve ambiguity.

Scientific work involves more than prediction. We also need to explain the functioning of the system. While measuring the accuracy is straightforward, measures of the explanatory capability of a model are problematic. Explanatory or "deep" models incorporating the mechanisms at work in a system would have more lasting value to scientists than the present "shallow" probability-based methods.

Acknowledgments

This work was accomplished during David Stockwell's studies towards his doctoral degree in the Ecosystem Dynamics Group, Australian National University. Numerous discussions with David Green, Ian Noble, Peter Cochrane, and Charles Zammit were very helpful. The assistance of Brian Turner, Sue Cuddy, and the anonymous reviewers in helping to clarify this manuscript is acknowledged. Rod Kavanagh and the N.S.W. Forestry Commission provided information and data necessary for the development and/or testing of the expert systems. We thank Rob Pearson and the Department of Computer Science, Australian Defence Force Academy, for the use of CART on the Convex computer.

References


Stockwell et al.: Decision Trees to Predict Greater Glider Density

EXISTING METHODS

Approach to problems in ecological management that are difficult to quantify. Journal of Environmental Management 16(3): 261-269.


Wyatt, S.W. 1986. The role of a geographic information system in forest level management planning. Honors thesis, Department of Forestry, Australian National University.

David R. B. Stockwell obtained a degree in science (honors) from the Australian National University in 1987 and is now working towards a Ph.D. in machine learning in the Ecosystem Dynamics Group at the Research School of Biological Science, Australian National University. He has been involved in the development of an expert system for diagnosing foodborne diseases in the Western Pacific Region. His current research project is an evaluation of the applicability of machine learning and AI representations to ecological theories.

J. Richard Davis obtained his degree in physics from Otago University, New Zealand, in 1969, and graduated from the Australian National University with a Ph.D. in nuclear physics in 1973. In 1974 he commenced research in land management in urban areas initially, and then in extensive rural areas. He joined the (then) CSIRO Division of Land Use Research in 1977 in order to work on the development of computerized techniques to assist land use planners. Since 1984 he has been particularly interested in applying knowledge-based techniques to complement traditional mathematical approaches. His recent projects have included constructing expert systems for fire management and tick control in cattle, and writing a knowledge-based program for drawing up local government zoning schemes.

Ian R. Noble obtained a Ph.D. degree from the University of Adelaide in 1975 on models of and zone grazing systems. He now leads the Ecosystem Dynamics Group at the Australian National University. His main research interest is in predicting the long-term dynamics of plant communities, especially where they are subject to disturbances such as fire, grazing, or climatic change. He has long been interested in using modelling and expert systems to make the results of research more readily available to land managers.

Stuart M. Davey has been lecturing, tutoring, and supervising students in the fields of wildlife management and forest ecology for the past nine years. He is presently employed in the Department of Forestry, Australian National University, as a senior tutor and has completed research towards a doctorate on the subject of arboreal marsupial habitat. His research and academic interests include forest ecology, management and planning, wildlife habitat and management.

AI Applications
5. LBS: Bayesian Learning System for rapid expert system development

Author: David R.B. Stockwell
Publication status: Submitted Expert Systems with Applications

Abstract. This paper presents the results of experience with a novel expert system shell called Learning Base System (LBS). The primary goal in developing LBS was to enable the rapid development of testable expert systems. The strategy adopted was to use a Bayesian classifier system as the form of knowledge representation, and adapt it to allow incremental acquisition of knowledge from both data and experts, and prediction and validation procedures. The advantages and limitations of the system are described in three applications. The first application is the diagnosis of diseases in crops, illustrating knowledge acquisition by an expert in a data-poor domain. The second illustrates how LBS could be used in a geographic information system. The third is the development and testing of models for predicting wildlife density solely from data. The Bayesian classifier is shown to be a flexible formalism for implementing a wide variety of knowledge-based tasks.

Introduction

In a recent survey of knowledge engineers, rapid development of efficient, cost-effective predictive expert systems was regarded as the most essential factor in further penetration of expert systems into business, manufacturing and science (Yue and Cox, 1991). Rapid development increases the number and range of applications of expert systems - particularly in crisis situations, such as diagnosis of disease epidemics, and environmental disasters such as oil spills. In this paper I describe and evaluate an expert system development tool or shell using an approach based on Bayesian probability theory for facilitating rapid development. The system is called Learning Base System (LBS). Rapid development follows from facilitation and integration of three knowledge acquisition tasks. Firstly, data is used to develop expert systems through machine learning methods. Secondly, a representation with structure intermediate in complexity between sets of single rules and a full matrix of probabilities facilitates the acquisition of knowledge from experts. Thirdly, statistical validation methods for assessing the predictive accuracy of the expert system are incorporated into the expert system development environment.

Bayesian methods have a long and controversial history as they have developed in parallel with more conventional statistical methods (Press 1989). More recently the approach has been debated in the context of appropriate AI systems for reasoning under uncertainty (Saffiotti 1987). Bayesian methods are arguably the most suitable model of the diagnostic process using inconclusive evidence (Spiegelhalter 1985).
appears that a Bayesian theoretical framework is useful for analysing a wide variety of machine learning systems (Buntine 1991).

The computational system used in LBS is based on simplifications introduced in the Bayesian network approach (Pearl 1986, Morawski 1989). The main representational structure is a matrix containing the frequencies of the occurrence of values of variables. However the matrix is decomposed into vectors as appropriate to facilitate knowledge acquisition, prediction and machine learning. For example, vectors of probabilities with the same outcome are grouped so that a single vector can be interpreted as a group of related rules. Like the Bayesian network, the basic Bayesian classifier is a simplification to ameliorate the complexity of the pure Bayesian approach.

Firstly, the major functions of LBS: prediction, knowledge acquisition, learning and knowledge base validation are described. Then, application of LBS in three different domains are illustrated: diagnosis of diseases by acquisition of knowledge from an expert, automatic mapping in a geographic information system by examples, and the automatic prediction of animal density from a data set is illustrated. The strengths and limitations of LBS are compared with alternative approaches to predictive modelling in these domains.

Goals of LBS

As LBS was developed as a general purpose expert system shell, or expert system development environment, all necessary functions for knowledge acquisition were provided within the shell. LBS provides life cycle support: gathering and organising knowledge, transferring that knowledge to the shell, modifying and testing the expert system, and running applications. There are a number of areas where this cycle could run more smoothly and rapidly. These are the sub-goals of LBS.

Knowledge can be gathered in various forms, both as deterministic and probabilistic rules, and as examples with a range of data types. Integration facilitates the full use of available knowledge. Machine learning assists knowledge acquisition by using forms of information that are either more accessible (e.g. observational data) or easier for the expert to specify (e.g. examples) to develop usable models. Acquisition of knowledge from examples is called inductive methodology, and is an aid to the rapid development of knowledge-based systems. LBS learns through modification of the conditional probabilities of the Bayesian classifier matrix.

Supporting revisions is essential for avoiding the 'knowledge base maintenance' bottleneck, because the consequences of changes are recognisable immediately providing instant feedback to the user (Morik 1991). A shell that acquires knowledge incrementally is able to elicit, test and apply knowledge smoothly at each stage. Non-incremental development can lead to the need for complete revision of the
expert system when new information becomes available. In LBS the probabilities
can be modified at will, either through the presentation of actual data, or by an
expert. The output of the expert system responds smoothly and appropriately to these
changes.

Rapid development requires that the form of representation of the knowledge
presented to the user should be easy for the developer to understand and modify.
Representations convenient for people may not be efficient representations for the
computer. LBS organises probabilities into an intermediate structure roughly
equivalent to groups of related rules in a rule-based expert system. A different
organisation is used by the system.

Formal testing of expert systems helps to maintain modelling standards, increase
user confidence in the expert system and guide development. Data is essential to this
stage in the cycle. LBS provides a number of techniques for testing expert systems,
including estimates of inherent noise in the data set, and unbiased estimates of
predictive accuracy on independent data.

As LBS was to be available to a wide range of users it was developed for the PC.
LBS is written in PDC-PROLOG. The tasks and abilities of the LBS shell listed
below represents a fairly complete set of functions that a diagnostic shell and expert
system needs.

*Prediction* - entering information and receiving a list of possible predictions or diagnoses rated
in order of their likelihood.
*Updating* - modifying incrementally with new information.
*Testing* - testing predictive performance.
*Explanation or tracing* - showing the relationship between predictions and evidence by tracing
the changes in the probability of the outcome as evidence is incorporated.
*Display* - examining the conditional probabilities to see how important particular evidence is for
indicating a diagnosis.
*Set-up - Expert system creation* - creating new expert system databases. The user enters the
variables and values, and the data structures containing the conditional probabilities are
generated by the expert system shell.
*Modification* - altering the knowledge base, either by initializing the whole learning base, or
modifying single values by hand.
*Data-base handling and file editing* - facilitating insertion, deletion, search or modification of
data base entries in a data set of examples.
*User-friendly* - Menus, forms, context-sensitive help, status lines and a user defined glossary of
terms are all available.
The functions of LBS.

The integration of prediction, knowledge acquisition, machine learning and validation is achieved through an ambiguity in the definition of probability. First there is the set theoretic definition of probability, i.e. given a randomly selected set of instances with a class frequency of 50%, we would expect 50% to be members of that class. Then there is the interpretation of probability as a belief. For instance, the partial belief \( P \), that a crop has a disease be expressed as a 50% chance of being present or, \( P(d) = 0.5 \). This ambiguity in the interpretation of probability can be exploited to integrate knowledge acquisition through data composed of class frequencies and knowledge acquisition from an expert's beliefs. The final predictions of the expert system are the outcomes with the highest probability, which can also be interpreted as an expected accuracy or as a degree of belief in a conclusion.

<table>
<thead>
<tr>
<th>Leaf pattern</th>
<th>chlorosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf structure/shape</td>
<td>drooping</td>
</tr>
<tr>
<td>Whole plant</td>
<td>wilting</td>
</tr>
<tr>
<td>Leaf colour</td>
<td>dark green</td>
</tr>
<tr>
<td>Particle shape</td>
<td></td>
</tr>
<tr>
<td>Geographic region</td>
<td></td>
</tr>
</tbody>
</table>

Growing Region

<table>
<thead>
<tr>
<th>none</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
</tr>
<tr>
<td>South</td>
</tr>
<tr>
<td>North-East</td>
</tr>
<tr>
<td>East</td>
</tr>
<tr>
<td>West</td>
</tr>
<tr>
<td>Central Plain</td>
</tr>
<tr>
<td>Every growing region</td>
</tr>
</tbody>
</table>

b.

<table>
<thead>
<tr>
<th>%</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.9</td>
<td>rice gall dwarf virus</td>
</tr>
<tr>
<td>10.1</td>
<td>rice grassy stunt virus</td>
</tr>
<tr>
<td>0.02</td>
<td>rice ragged stunt virus</td>
</tr>
<tr>
<td>0.01</td>
<td>rice transitory yellowing virus</td>
</tr>
<tr>
<td>0.00</td>
<td>rice yellow orange leaf virus</td>
</tr>
<tr>
<td>0.00</td>
<td>none</td>
</tr>
</tbody>
</table>

Figure 1. An example of diagnosis of a crop disease from symptoms. (a) The user chooses values from a menu. Fields may be left blank (e.g. Particle shape). (b) The predictions are listed in descending probabilities.

Prediction

Figure 1 illustrates a typical prediction in LBS. First the user selects possible values for the evidence variables from a menu (Figure 1a). The user may choose no value (none) or enter a blank. LBS then uses the selected values to calculate the probabilities of each outcome, and outputs the predictions in descending order (Figure 1b). The final probability is arrived at through incremental incorporation of evidence where evidence refers to values of variables relevant to the outcome. Evidence is relevant if, for example, previous observations have shown a particular
EXISTING METHODS

symptom was present when the disease was present. This tendency is then partial
evidence for the presence of the disease. Hence we have three given probabilities:

1. the **prior probability** of disease \( P(d) \),
2. the probability of the symptom \( P(e) \), and
3. the **conditional probability** of the symptom appearing given the presence of the disease \( P(eld). \)

and one that can be calculated:

4. the **posterior probability** of the disease being present, given the symptom, \( P(dle) \).

If \( P(dle) \) is high then we may be justified in asserting that the disease is present. \( P(dle) \) is called the posterior probability and is the quantity we are interested in for making predictions. Providing \( P(e) > 0 \), it is calculated by the equation:

\[
P(dle) = \frac{P(eld)P(d)}{P(e)}
\]

Equation 1 supports the inference that a disease occurs when the outcome \( P(dle) \) is high. It can be thought of as a rule of inference that links evidence to outcomes. Consider the following example of disease diagnosis. The incidence of a disease is one in ten cases, therefore the prior probability is 0.1. The probability of the occurrence of a symptom given that disease is 0.8. The probability of the occurrence of the symptom is 0.2. Therefore by equation 1 the probability of the disease given the symptom is:

\[
P(dle) = \frac{0.8 \times 0.1}{0.2} = 0.4
\]

A simplification is introduced by the property that one and only one of a variable's values may be true at a time. Thus the sum of probabilities over all its values is one. For example among the number of possible diagnoses, only one may be correct. The incidence for all diagnoses may then be grouped into a vector where the probabilities sum to one. The vector of probabilities prior to the incorporation of a piece of evidence is called the prior probability vector \( P(d) = \langle P(d_1), P(d_2), ..., P(d_n) \rangle \). Given the prior vector \( P(d) \) and the appropriate evidence conditional \( P(eld) \) the calculation used to obtain the posterior vector is:

\[
P(dle) = \alpha P(eld) \ast P(d)
\]

where \( \ast \) denotes the term by term multiplication of vectors and \( \alpha \) is a normalising vector to ensure the posterior probabilities sum to one.

The normalising factor circumvents the need to keep note of the probabilities of the evidence \( P(e) \). It is simply the average value of all the \( P(d_1, e) \). Evidence can be perceived by the user to affect the outcome vector as a whole, rather than contributing to many separate inferences, simplifying the presentation of the inference process.
The inference proceeds by incorporating successive pieces of evidence into the prior vector. The calculated posterior probability vector then becomes a prior probability for the next piece of evidence. This process continues until no evidence remains. The final diagnoses may then be sorted into decreasing order of probability and output, or the highest probability chosen as a prediction. For example, the starting point for the complete diagnosis is a record of the incidence of a suspected disease (prior probability). This probability is modified by evidence supplied by the user (e.g. leaf symptom, vein symptom, leaf colour, etc.) and the conditional probability stored in the database. For example given the incidence of bean mosaic virus relative to other virus infections, and the proportion of plants with bean mosaic virus that exhibit mosaic (conditional probability), the probability that a given plant with mosaic symptom is infected with bean mosaic will be calculated. The corresponding probabilities for all other viruses in the database will also be calculated. These results (posterior probabilities) become the prior probabilities for the next calculation. The process is repeated until all the symptoms have been incorporated. The list of diagnoses is displayed in order of decreasing probability.

The main limitations of the Bayesian classifier is the need for categorical variables, and the independence between variables. While Bayesian statistics in their pure form can be applied to continuous, ordinal and categorical variables, real-valued variables must be converted into categories. This pre-manipulation of data has not been found to be a great hindrance in practise.

Independence assumptions are necessary for decomposing the diagnostic process into operations on single variables. Under the important assumption of independence in the evidence, any two pieces of evidence presented singly are as significant as the two pieces presented in conjunction. A Bayesian system operating under this assumption is known as 'naive Bayes.' The consequences on dependence between variables are reported in a forthcoming study (Stockwell, in preparation). Evidence from two correlated variables over-estimates the probability of a disease as the same piece of information is effectively counted twice. Although the estimated posterior probabilities are then incorrect, this may not unduly affect the relative probabilities and hence the ranking of diseases. Therefore a greater degree of correlation can be tolerated when the user is only interested in finding which disease or diseases is most likely.

Knowledge Acquisition

It has been suggested that an expert has a better understanding or can estimate the values of $P(eld)$ better than $P(dle)$. For example a wildlife expert is more familiar with the probability of environmental factors associated with an animal than the probability the animal will be present given a particular combination of environmental factors. A doctor is more familiar with the probabilities of symptoms given a particular disease, than with the probability of diseases given a particular
EXISTING METHODS

symptom. This preference for presentation of information in a particular form is a
form of human bias where estimates of probability are conditioned on 'interesting'
outcomes. There is an advantage in presenting information in a format that is
consistent with this human bias.

The expert enters knowledge into LBS by modifying a diagnosis vector. Each
diagnosis variable has a vector of conditional probabilities associated with a single
outcome d: P(£ld) = <P(e1|d),P(e2|d),...,P(en|d)>. An alternative arrangement of
conditional probabilities is called an evidence conditional. The evidence conditional
is a vector of conditional probabilities associated with single values of variables. The
purpose of these vectors is to provide an intermediate level of organisation, both
aiding the user in understanding the values derived by the LBS and assisting in
explanation of the calculations performed. The three types of vectors are listed in
Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Vector</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>prior and posterior vector</td>
<td>P(d)</td>
<td>&lt;P(d1),P(d2),...,P(dm)&gt;</td>
<td>the vector listing the current probability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>values of the outcomes d1 ... dm.</td>
</tr>
<tr>
<td>diagnosis conditional</td>
<td>P(£ld)</td>
<td>&lt;P(e1</td>
<td>ld),P(e2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>variable e given a certain outcome d.</td>
</tr>
<tr>
<td>evidence conditional</td>
<td>P(eld)</td>
<td>&lt;P(eld1),P(eld2),...,P(eldm)&gt;</td>
<td>the probabilities of a value e of a variable</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>given each of the outcome values d1 ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>dm.</td>
</tr>
</tbody>
</table>

Table 1. The data structures of LBS are represented as three types of probability vector: the prior
probability vector, the diagnosis conditional vector and the evidence conditional vector.

Note that the diagnosis conditional is equivalent to a row of a full conditional
probability matrix and the column is equivalent to a column in the full matrix. The
two conditional vectors are used for two different purposes. The diagnosis
conditional vector is used in learning while the evidence conditional vector is used to
incorporate evidence e.g.

Learning: Given example with class d : P1(£ld) → P2(£ld)
Prediction: Given evidence e : P(eld) * P1(d) → aP2(d)

The probabilities in the conditional probability matrix can be modified by the
expert and stored in the expert system database. As an example of encoding
knowledge by modifying conditional probabilities, consider a disease, 'rice gall leaf
virus' and a variable 'leaf pattern.' A screen similar to Figure 2a would be presented
to the expert. Given that galls always occur with rice gall leaf virus and chlorosis is
sometimes present, the expert might modify the probabilities to those shown in
Figure 2b. The expert would continue to input estimates of the probabilities for all
combinations of diseases and diagnostic variables. The expert system database would
then be ready for testing.
a. Disease
rice gall leaf virus

<table>
<thead>
<tr>
<th>%</th>
<th>Leaf Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>none</td>
</tr>
<tr>
<td>20</td>
<td>chlorosis</td>
</tr>
<tr>
<td>20</td>
<td>rusty flecks</td>
</tr>
<tr>
<td>20</td>
<td>gall</td>
</tr>
<tr>
<td>20</td>
<td>vein swelling</td>
</tr>
</tbody>
</table>

b. Disease
rice gall leaf virus

<table>
<thead>
<tr>
<th>%</th>
<th>Leaf Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>none</td>
</tr>
<tr>
<td>20</td>
<td>chlorosis</td>
</tr>
<tr>
<td>0</td>
<td>rusty flecks</td>
</tr>
<tr>
<td>100</td>
<td>gall</td>
</tr>
<tr>
<td>0</td>
<td>vein swelling</td>
</tr>
</tbody>
</table>

Figure 2. An example of inputting knowledge about a crop disease, rice gall leaf virus. (a) A conditional probability vector set to initial probabilities - no knowledge is incorporated at this stage. (b) The conditional probability after the incorporation of knowledge.

Deterministic evidence can be represented by probabilities of zero or one in the diagnosis vector. By equation 1, if \( P(e|d_i) = 0 \) then \( P(d_i|e) = 0 \) for all subsequent inferences. Thus setting \( P(e|d_i) = 0 \) is a way of stating a deterministic rule:

'if e then d_i' has a probability of zero (is false irrespective of any other evidence).

Similarly when an evidence vector has one value set at 1 and the rest are set to zero, this has the effect of a deterministic rule:

'if e then d_i' has a probability of one (is true irrespective of any other evidence).

The user can incorporate deterministic rules by modifying the evidence vectors directly, irrespective of the values that might be suggested by the examples and set by the learning algorithm.

Learning.

A wide variety of machine learning methods have been developed. Most do not support general knowledge elicitation through incremental acquisition of knowledge. For instance, incremental change to decision trees may require complete structural change to the tree. An algorithm may also produce a number of different decision trees during resampling validation. Maintaining these alternative representations of the knowledge is unsatisfactory and difficult for the users to comprehend (Stockwell et al. 1990 and Manuscript 4). By comparison, the Bayesian classifier has a stable structure that supports incremental modification through alteration of probability values.
EXISTING METHODS

LBS is 'taught' with examples by altering the value of the prior and conditional probability vectors. An example is composed of an outcome and a list of values of variables. The conditional and prior probabilities are modified according to the frequencies of values and outcomes and elementary probability theory. The equations for updating the elements of the diagnosis conditional vector are:

\[ P_{\text{new}}(e_i|d) = \begin{cases} 
\frac{P(e_i|d) \cdot n+1}{n+1} & \text{where the value of } e \text{ is } e_i \text{ and} \\
\frac{P(e_i|d) \cdot n-1}{n+1} & \text{where } e \text{ doesn't have the value } e_i.
\end{cases} \]

Where evidence \( e \) has outcome \( d \), and \( n \) is the number of cases seen so far.

Similarly the updating of the elements of the prior vector for a piece of evidence is given by:

\[ P_{\text{new}}(d_i) = \begin{cases} 
\frac{P(d_i) \cdot n+1}{n+1} & \text{where } e \text{ has diagnosis } d_i \text{, and} \\
\frac{P(d_i) \cdot n-1}{n+1} & \text{where } e \text{ doesn't have diagnosis } d_i.
\end{cases} \]

Validation of LBS

The use of unvalidated scientific models must have negative consequences, particularly the management of economically important resources (Hall 1988). Recognition of the need for more scientific methods of managing limited resources, and possible legal accountability of modellers, means automated systems need to be rigorously tested. Systems based on the human expert as a source of knowledge merely enshrine expert opinion - or ignorance. LBS contains a number of measures of accuracy, each with some informative value to a predictive model developer:

1. the model estimate of the accuracy of its prediction,
2. resubstitution, or estimates based on the data used to develop the model,
3. cross-validation, where an entirely different data set is used for the testing as was used for development of the model, and
4. data resampling, a method of independent validation with a single data set.

The model estimate is determined by finding the average posterior probabilities for each diagnosis. It is useful for estimating the approximate accuracy of a single prediction. In resubstitution the training set is used as the test set, i.e. the data used to develop the model is substituted back into the model. It is useful to determine the inherent noise in the data relative to other similar sets, and to determine the upper limit on the predictive accuracy. In cross-validation, the model is developed on one example set and tested on another. The test is completely independent of the data used to develop the model. The data sets may be divided randomly or on some other criterion such as temporal or spatial separation.

79
Data resampling methods develop a measure of the expected accuracy for predictions. The data is repeatedly separated into two sets with one used for developing the model, and the other for determining the accuracy. While resubstitution invariably leads to an optimistic estimate of accuracy of the model when predictions are made on independent data, resampling eliminates this bias (Verbyla and Litvaitis, 1989). LBS uses a resampling method called ten-fold validation. The procedure is: (1) The data set is partitioned into ten randomly selected, equally sized subsamples, (2) the first subsample is excluded, and (3) the model is developed with the remaining subsamples. (4) The model is then tested on the excluded subsample. (5) The steps two to five are repeated, excluding a different subsample each time. (6) The estimate of the model accuracy is calculated from the combination of the tested samples.

An example of the display of results in LBS is shown in figure 3. The overall accuracy for the output variable (GG or Greater Glider) is shown in the first column. The remaining columns report the results for the outcomes: absent, present and missing value, respectively.

<table>
<thead>
<tr>
<th>Total</th>
<th>0</th>
<th>0.9999</th>
<th>MV</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 Right</td>
<td>12</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>5 Wrong</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>90.00 % Estimated</td>
<td>85.01 % Estimated</td>
<td>50.40 % Estimated</td>
<td>0.00 % Estimated</td>
</tr>
</tbody>
</table>

Figure 3. An example of output from validation of a knowledge base for predicting the occurrence of a species. Shown are the overall accuracy for the output variable (GG or Greater Glider), absent, present and missing values respectively.

Limitations due to inconsistencies, missing values and memory

The handling of inconsistencies and missing values has proved troublesome. An inconsistency can occur when a deterministic prior vector, one that is composed of all zeros except one, is multiplied by an evidence vector that is deterministic in a different variable. The resulting vector is composed completely of zeros, suggesting that no outcome is the case. This violates the assumption that there must be one and only one outcome.

Several interpretations of contradictions are possible (Poole 1989). We could discard the prediction, on the basis that that part of the knowledge base is inconsistent. We could output a 'none' value indicating that no outcome is satisfactory. Or the contradictions could be interpreted as saying that on the basis of the evidence presented we cannot distinguish between d and not d. This approach is taken is LBS. In case of contradiction, the prior vector is reset to an uninformative prior (equal probabilities for each value).

Missing values are treated as an additional value automatically added to the list of values for a variable. A contradiction in the other values of a variable can result in
the missing value becoming an outcome of a prediction. This result is undesirable as a contradiction doesn't imply a missing value. One can eliminate this eventuality by setting the prior expectation of a missing value to zero.

As all Bayesian systems make intensive use of storage and computation, some tradeoffs in systems with limited working memory are inevitable. A database on hard disk or in extended memory allows the data to be accessed while the program is running, overcoming the limitations of PC RAM, but can slow down some operations. LBS can handle tasks of intermediate size with a typical application having 12 variables with five to 30 values in each variable.Scrolling windows allow more values to be displayed on the screen than the number of rows on the screen.

Applications of LBS

The following section describes three typical applications of LBS. Each application represents a different way of acquiring knowledge: subjective assessment and modification of probabilities by the expert, input of training examples created by the expert to reflect typical situations, and automated model development using raw data. Learning systems have been ordered along a knowledge dimension according to their background assumptions. Weak domain knowledge, or few assumptions is at one end and strong domain knowledge, or many assumptions at the other (Cheeseman 1990). A useful conceptual aid is to plot knowledge based systems along two dimensions of information, one of expert knowledge and the other of data (Figure 4). In this classification, LBS lies in between the extremes of neural nets (knowledge-poor and data-rich) and rule-based systems (knowledge-rich and data-poor). The Bayesian classifier uses weak domain knowledge when learning from examples, and strong domain knowledge when a deterministic rule is set by an expert and can be applied to domains that are rich or poor in knowledge or data.

<table>
<thead>
<tr>
<th>Rich</th>
<th>Data</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Nets</td>
<td>Learning Base System</td>
<td>Rule-based expert systems</td>
</tr>
</tbody>
</table>

Figure 4. Knowledge based systems can be classified along two dimensions according to their requirements for data and knowledge. Neural nets need many data, and no inputs of knowledge from an expert. Rule-based systems can be developed from the knowledge of an expert alone with no reference to data. The Bayesian classifier system is intermediate to the two extremes.
Diagnosis of crop viruses - knowledge-rich and data-poor

Two expert systems for diagnosing viruses of crops in Thailand were developed with the LBS shell (Stockwell and Dilokkunanan, in preparation). The first diagnosed five viral diseases of rice from seven evidence variables. The second diagnosed 18 legume diseases from 13 variables. The evidence variables were mainly symptoms easily recognisable from examination of the plant such as abnormal leaf pattern, vein colour, leaf, flower, fruit stem and whole plant characteristics. Geographic region, transmission vector and particle shape were the remaining evidence variables. Only one variable, particle shape, would require laboratory facilities to determine. Thus the diagnostic systems could be cost effective alternatives to laboratory diagnosis when absolute confirmation is not required.

Diagnosis can be difficult when there are large number of possible diseases, and no simple diagnostic indicators for each disease. Often a symptom is associated with a number of diseases. For instance, the presence of mosaic symptom on leaves is consistent with a large number of common plant viruses. It was found using LBS to combine partial evidence, that a number of inconclusive symptoms was usually sufficient to reach a conclusive diagnosis. The power of combining partial evidence was surprising to the expert. Often only a few symptoms that would normally have been regarded as inadequate for conclusive results were sufficient to produce a correct diagnosis with high probability.

After becoming familiar with the system the expert developed the second expert system in a single working day. This demonstrates the rapidity of developing expert systems even when no basic data is available. LBS was thought to satisfy the criteria of 'appropriate technology' for developing countries: scientific validity, adaptability to local needs, user-friendliness, and cost effective acquisition, maintenance and utilization with the resources the community can afford (King and Beck, 1990). As well as improving decision making, productivity and diagnostics, expert systems also have a role in transferring expertise to users in biology and agriculture (Olson et al. 1989).

Limitations in disease diagnosis

The main problems to emerge in the application of LBS to the diagnosis of viruses in crops was the possible mis-match between the assumptions of the naive Bayes algorithm and the problem domain. The two main assumptions of naive Bayes in the context of disease diagnosis are: (1) only one disease is responsible for all the observed symptoms, and (2) there is little interaction between diagnostic variables. When more than one virus infects a plant, novel symptoms may be produced which are rarely the sum of those of the constituent viruses. LBS makes no allowance for these symptom complexes. Another problem occurs even with additive symptoms of
Existing methods

Multiple infections. As an atypical symptom is evidence against a disease, the probability of one of the infecting diseases will be lowered despite other evidence for the disease. If both diseases have incompatible symptoms they will reduce the probability of each to zero.

LBS provides a strategy that can be applied in the case of multiple diseases. By setting the probability of a symptom given a disease to zero the occurrence of that symptom will not count against the diagnosis. While this approach prevents cancellation when multiple diseases are present, it utilises only evidence for a disease and discards evidence against it. If the conditional probability is set at a very low number, e.g. 0.01 the probability of a disease will be strongly reduced if the symptom is not consistent with the disease. This approach incorporates both evidence for and against a disease into the diagnosis, providing maximum discrimination between diseases.

An example of a serious violation of the independence assumption is when the symptoms of a virus are completely different on different host plants. In this case it is possible to decompose the problem into a number of smaller problems based on diseases that have similar effects on a group of host plants (e.g. such as rice diseases and legume diseases).

A geographic information system - partial knowledge and data

The following example is an illustration of a possible application of LBS in a geographic information system (GIS). A GIS is a tool for managing large numbers of spatially related data. It normally contains large files of attribute vectors - variables values referenced by location. This information is often displayed as layers or thematic maps, where each value of a single variable is displayed as a coloured picture element, or pixel, on the screen. GIS layers may contain information on roads, vegetation, land usage, and topographic features such as slope or aspect. This information itself is often useful to store in computerised form. The real advantage of GISs however is the ability to derive information through combining or overlaying layers. The user can then go beyond information already stored in the data base and explore relationships between layers. Typical combinations include logical ANDs and ORs, exclusive ORs, additions and so on. This method is efficient providing the way to combine the layers is already known, e.g. when the derived layer is a simple conjunction of two layers. In many cases there is no such prior information and the user must then develop a predictive model for the purpose.

An LBS type system could be used to develop a predictive model automatically. The expert system would be trained by selecting both positive and negative picture elements as examples of the type of information to be found, and specifying the possible input variables (Figure 5a). LBS would then develop an expert system from these examples (Figure 5b). The user could refine and test the expert system before
using it to predict the occurrence of the predicted class across the whole picture area (Figure 5c). After development, the knowledge base could be stored and re-used with any subset of the variables used in the original training session.

![Diagram of Selection, Development, and Application stages](image)

**Figure 5.** Stages in the application of LBS to modelling in a geographic information system. (a) The user selects positive and negative examples of the type of area to be recognised. (b) A model is developed from these examples. (c) The accuracy of the model is checked and reapplied to the picture area.

### Comparison with other methods

Statistical modelling techniques or decision trees could, and have, been used in GISs. However, a particular feature of the geographic domain can be difficult to handle with these methods. Due to sampling bias, the information is often incomplete or of variable reliability. While decision tree induction has proved practical for edge detection, the induced tree structure is not flexible enough for an expert system environment (Goodman and Smyth, 1990). For example if data for a branch in the tree is unavailable during application of the tree, the classification halts inconclusively. In principle a new tree needs to be developed for every new combination of variables. A Bayesian classifier on the other hand can be developed using a large number of variables, and perform predictions using a subset of those variables. This illustrates another advantage of the value the ability to diagnose using partial evidence.

The statistical methods available for classification and prediction tasks as above are multiple linear regression, logistic regression and linear discriminant function analysis (James and McCulloch, 1990). Each make assumptions about the form of the underlying function or structure in the data. Linear regression and logistic regression assumes the relationship of input with output variables is linear. Discriminant analysis is very sensitive to outliers, assuming the data is non-disjunctive. These forms of analysis use global functions, functions that describe a particular type of relationship between the input and output variables over the whole of the domain. It is very important that the data analyses by these methods conform to these functions. In contrast the Bayesian classifier makes no such assumptions about the form of the function. LBS is therefore applicable to a greater range of data, and can be applied with less concern about whether the data meets these restrictive assumptions.
Prediction of wildlife occurrence - knowledge-poor and data-rich.

The accurate prediction of the distribution of plants and animals is a crucial task in the management of natural resources. Unfortunately little is usually known about the factors that determine the distribution of species. This is due to the large number of species that could be studied, their often cryptic or nocturnal habits, and shortage of funding for basic distributional studies. Hence the prediction of the distribution of a species takes place in a knowledge poor domain. The data sets of environmental variables are usually composed of numerous variables of unknown applicability to the species of interest.

The Waratah Creek greater glider data set consists of seven predictor variables for 400 sites in the Coolangubra State Forest region of south-eastern New South Wales, Australia. The outcome variable is the greater glider density at four levels, nil (0 animals/ha), low (less than 0.5 animals/ha), medium (0.5 to 1.0 animal/ha), and high (greater than 1.0 animals/ha). The predictor variables were categorical forest inventory variables: land development, proximity to a stream, stand condition, site quality, floristic nutrients, slope class and erosion potential.

Table 2 compares the predictive accuracy of application of LBS with the results of accuracy on the same data set, in a previously published multi-model comparison (Stockwell et al. 1990 and Manuscript 4). The models compared were: a rule based expert system developed entirely through interviews with an expert (Stockwell 1987), and three decision tree induction algorithms, ID3 (Quinlan 1986), CN2 (Clark and Niblet, 1989) and CART (Breiman et al. 1984). LBS predicted as accurately as the best of the decision tree methods, ID3 and CART and much better than the expert system developed through conventional methods of knowledge acquisition (KA) (Table 2). In all models an estimate of predictive accuracy given by resubstitution was better than the resampling estimate. This illustrates the importance of resampling procedures for validation of all models. LBS showed less of an overestimate than other methods.
LBS: BAYESIAN LEARNING SYSTEM

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy resubstitution</th>
<th>Accuracy resampling</th>
<th>Selection at 99%</th>
<th>Number selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBS</td>
<td>58.2±0.5%</td>
<td>56.4±0.5%</td>
<td>80.1±1.0</td>
<td>27±1%</td>
</tr>
<tr>
<td>KA</td>
<td>-</td>
<td>48%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ID3</td>
<td>60.7±1.3%</td>
<td>57.3±2.2%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CN2</td>
<td>50.8±2.6%</td>
<td>45.2±2.2%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CART</td>
<td>61.2±3.5%</td>
<td>54.8±3.9%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Comparison of a range of predictive methods on the Waratah Creek Greater Glider data set. The models compared are LBS - Learning Base System, KA - a rule based expert system developed by an expert, and decision tree induction algorithms, ID3, CN2 and CART. Resubstitution indicates the fit of the model to data. Resampling methods give the unbiased predictive accuracy. The final two columns gives the accuracy of predictions using a decision rule that predicts only when the posterior probability estimated by the model is greater than a certain level, and the average percentage of data items selected for prediction. Predictions with a posterior probability greater than 99% are significantly better than predictions with no cutoff. All errors are quoted as standard errors of the mean.

Because LBS predicts by determining the posterior probability of outcomes, this posterior probability can be used as a basis for boosting the predictive accuracy. By only predicting in the cases where the probability was greater than 99% the predictive accuracy was increased from 56% to 80%. The increase in predictive accuracy was traded off against coverage however, as the percentage of data selected decreased to 27%. Selective prediction would be useful for developing maps with varying degrees of confidence, and identifying areas where additional data needed to be gathered to increase confidence.

Comparison with knowledge-rich modelling.

One form of knowledge-rich modelling is to develop a mathematical description of the processes in the modelled system and then to parametise these equations with careful and time-consuming experimentation. Such research for modelling wildlife distribution would be prohibitively expensive and limited to only a few species. The alternative approach adopted by the Bayesian classifier minimises the need for experimentation by simply using the statistical relationships between observed variables. Thus LBS requires no knowledge of the underlying processes. Models of this form have been called 'calculation tools' to emphasise their purely instrumental function in prediction (Loehle 1983).

Machine learning researchers on the other hand would classify this type of modelling empirical, as knowledge is acquired purely from empirical data. While empirical knowledge acquisition is rapid, it also has limitations. Even though LBS provides validation procedures for determining the expected predictive accuracy of the model, care must be taken to ensure the data is adequate and randomly sampled so that the predictions are not over-generalised.
Bayesian classifier as knowledge representation

In the introduction to this paper the main goal of rapid development of expert systems for accurate diagnosis was stated. This required an integrated and structured approach to knowledge acquisition that supported learning from data. The experience has been that the Bayesian classifier achieves these goals. The ability to explain the reasoning that led to a diagnosis was not one of the goals. The explanation capacity of LBS is limited to tracing the effect of inclusion of evidence on the diagnosis vector. Explanation of reasoning has come to be a facility expected of expert systems, being one of the major innovations of the practical medical diagnostic systems using a rule-based architecture (Davis et al. 1977). The form of knowledge representation used in LBS was chosen to facilitate knowledge acquisition by the expert. It is worth briefly discussing the issues surrounding the form of knowledge representation and why LBS seems to have limited explanatory reasoning in relation to rule-based systems.

The knowledge representation in a rule-based system is a large number of independent rules, each representing a singular 'chunk' of knowledge. This knowledge might encapsulate a statement that a certain symptom suggests a certain disease with a certain probability. The role of a rule is taken in the Bayesian classifier by a single value in the conditional probability matrix. LBS structures these single values as vectors of probabilistic rules, with groups of similar rules in the same vector. Vectors are intermediate in organisational complexity to rules and matrices. This provides a useful organisation to the knowledge and unlike rule-based expert systems, the knowledge base cannot grow without bounds as the size of the expert system knowledge base is fixed on the initial specification of the database. Groups of rules in a rule-based expert system form complex but ill-defined structures. Representation of collections of rules in a matrix, 'structures' all the rules and facilitates the acquisition of knowledge from experts (Thompson and Thompson, 1987). LBS is easier to develop and maintain than a rule-based system due to this intermediate structure.

However, the structure used in LBS limits the representation in three main ways. Firstly, in rule-based expert systems, rules may have any number of expressions in their precondition. A single value of conditional probability in LBS is equivalent to a single rule in a rule-based expert system. Therefore rules in LBS only have a single value in their precondition. This restriction means 'higher order' interactions between variables cannot be used for prediction. Secondly, in LBS only one diagnosis of one variable can be predicted. These restrictions are incorporated into LBS to simplify it and keep the computational time down. Thirdly, the conditional probabilities in LBS give no clear indication of the importance of relationships of evidence variables to the outcome. Thus unlike rule-based systems where significant causal relationships can be expressed using a single rule, the knowledge is distributed throughout many values in the probability matrix. The Bayesian classifier system is similar to a neural
net in this respect, where distributed representation facilitates incremental acquisition of knowledge. The consequence of structuring and a distributed representation however is a negative effect on the ability of the system to explain it's reasoning.

Conclusions

LBS has many features necessary for achieving the goals of rapid development: knowledge integration, incremental acquisition through machine learning and knowledge elicitation, and rigorous testing facilities. The success of LBS on a number of applications illustrates the flexibility of Bayesian classifiers as a means to efficient development of predictive and integrated expert systems. Independence and single outcome assumptions do present limitations in disease diagnosis. Further research could be directed to understanding the behaviour of the system in automated model development and creating a more general system. LBS is available from the author.

Acknowledgements

Thank you Stuart Davey and the Forestry Commission of New South Wales for the use of the Waratah Creek greater glider data. Adrian Gibbs and Uriwan Dilokkunanant provided an opportunity to develop LBS further through the ACIAR funded Thailand crop diagnosis project. David Green contributed valuable comments on an early draft of this manuscript.
Section 3. Theoretical development of prediction and explanation

Induction of decision trees and the LBS system illustrated features of the prediction/explanation problem. The main problems seemed to be related to model structure, as shown by structural instability and difficulties in interpreting the components of the model. The model development systems produced models that predicted well, despite difficulties in using these models for explanations. The main prediction problem was unexpected failure in certain situations. Problems of structure are concerned with choices made about the form of model to be applied to a particular problem. Structural problems concern broad issues in modelling methodology, rather than problems of the development and application of specific types of models.

It was obvious that a greater understanding of the nature of prediction and explanation was needed to address these problems. Prediction was found to be a paradigmatic problem, i.e. a problem with a number of components that could be posed in a number of different and incomparable ways (Manuscript 6). For example, a number of distinct predictive problems correspond to particular relationships between the data set used to develop a model and the data set used to test a model. These distinct predictive problems were defined as: resubstitution, resampling, cross-validation, forecasting, extrapolation, interpolation, and transmission.

A view of explanation used in machine learning was applied to ecological model development (Manuscript 7). On this view, developing structural models entails structure on three levels: events, characterised by the structure of data; concepts, characterised by the structure of logic; and bias, the preferences for general types of structure. Similarly a number of paradigms are found in explanation: causal, mechanistic, general laws, probabilistic, analogical and teleological. Model development is defined and constrained by these forms of structure. The explanatory paradigm that was most applicable to ecology needed to be identified. Using a survey of students and scientists, the causal paradigm was found to be the most applicable (Manuscript 8). Logical if-then statements also support causal interpretations. Thus logical if-then statements, embodied by sets of rules in a rule-based expert system, represent a form of model that could be both predictive and explanatory.
6. Learning to predict: an empirical evaluation of the Bayesian classifier as a general predictive system

Author: David R.B. Stockwell
Publication Status: Submitted to Machine Learning

Abstract The predictive accuracy of machine learning methods is often evaluated in specific situations, i.e. with noise free and domain specific data, or with specific predictive protocols such as data fitting or resampling. Consequently the predictive tests depend on the situation, and comparisons between models are difficult. This paper examines the general problem of learning to predict in the context of one particular learning system, the Bayesian classifier. The effects of varying components of the predictive task are determined on an ecological data set. The components of the predictive task varied were: object of prediction, type of data, predictive protocol, decision system, and measures of predictive accuracy. The results show how success or failure of prediction depends on the exact specification of the predictive paradigm.

Introduction.

Studies of empirical machine learning have used a range of methods of evaluating predictive adequacy. For example, in tests of the ID3 algorithm, models were developed on a training set, the set was corrupted in a controlled way, and then the predictive accuracy was tested (Quinlan 1986). This procedure simulates the development of a model in a noise-free environment, such as a laboratory, and a 'real-world' noisy situation. In another evaluation of ID3, decision trees were trained on 70% of the data and tested on the remaining 30% (Mingers 1989a). This procedure simulates an independent test, such as repetition of an experiment in a different laboratory. These procedures illustrate two different predictive situations in which models are tested.

The success of one machine learning method at one task does not suggest that it will be successful at another. An example of this phenomena is 'overfitting'. In overfitting the model developed is very specific to the training set and performs poorly on the test set. Overfitting is known to be a problem with multi-variate statistical methods when the data set contains many predictor variables (Verbyla 1986). Overfitting is an example of a model succeeding at one type of prediction (resubstitution) and failing in another (resampling). The growing awareness of the problem of overfitting of models has led to a trend in machine learning to unbiased evaluation of prediction, estimated by resampling methods such as jackknife or bootstrap resampling (Mingers 1989a, 1989b).

Prediction is often assumed to mean that a model must 'fit the facts' (Loehle 1983). The variety of forms of prediction used in model evaluations shows that the
notion of fitting the given facts is inadequate. The facts the model is designed to fit may not be available when the model is being developed. For example, models are generally designed to predict in situations where no data is available, e.g. future situations (forecasting), situations distant in space, or situations not covered by the range of the given facts (extrapolation and interpolation).

One reaction to the plurality of predictive tasks is to adopt a specific method as the norm. Resampling predictive tests are proposed as standard methods of evaluating empirically developed models in many fields (Verbyla and Litvaitis, 1989). Another approach is to solve the general problem of finding the best predictive model. This problem is solved in theory by the Bayesian inductive procedure - simply enumerate all possible models, select a set of those with the highest posterior probability, and use a weighted average of their predictions for predicting (Cheeseman 1990). This algorithm is impractical to implement as the number of possible models is usually very large. In practice, concessions to computational tractability are made that do not guarantee the success of this process: a restricted class of possible models is chosen, and only one final model is selected.

A third approach taken in this paper is to generalise the notion of prediction into a predictive paradigm, incorporating the range of alternatives, and develop more general predictive systems. This study takes an empirical approach to evaluating the effect of alternatives. The main reason for an empirical study is that the formalisation of both for prediction by resampling and the Bayesian classifier system is complex. An empirical study is sufficient for exploration of the problems and difficulties of prediction. In order to address specific predictive situations, a predictive paradigm is defined, following the inductive paradigm defined by Kelly and Glymour (1990). The components of a predictive paradigm are:

- a language for expressing predictions,
- a formal predictive protocol,
- a space of theoretically possible situations,
- a decision system for asserting a prediction given a situation,
- a criterion of predictive adequacy, and
- a method for converging on a decision system.

There are a variety of ways of filling each of the components of the predictive paradigm, depending on the given problem, and methods available for providing predictions. In this paper a Bayesian classifier system was used - part of a system for rapid development of predictive systems (Manuscript 5). Experiments are performed on a data set on the occurrence of waterbirds, illustrating the components of the
paradigm and the effect of variations on the classifier system. The components of the paradigm examined in this paper are as follows.

- The language for expressing predictions is restricted to presence and absence of water birds.
- A number of formal predictive protocols are compared: resubstitution, resampling and forecasting.
- The space of theoretically possible situations includes predicting a range of species of birds from the physical, chemical and biological parameters of water bodies in a range of disused quarry pits in Gloucestershire, England.
- The decision system, the Bayesian classifier, uses a decision rule based on the estimated posterior probability, which is varied according to the expected probability of successful prediction.
- The measure of predictive adequacy is determined from comparison of the percentage of correct predictions with a range of null models.
- The effect of increasing data and numbers of variables on convergence of the decision system is determined.

It is shown that there are many choices in each component of the paradigm that effect the outcome of a predictive trial. A predictive system may fail to predict due to inappropriate choices in the components of the test situation. A greater understanding of the components of the predictive paradigm and the way it affects predictive accuracy will allow more reliable modelling and model comparisons and assist in understanding of the issues underpinning the automated learning of acceptable predictive models.

**Prediction with a Bayesian system**

A number of AI systems for reasoning under uncertainty have adopted a Bayesian classifier approach (Saffiotti 1987). One of the first was PROSPECTOR, an expert system for mineral exploration (Duda *et al.* 1976). Medical diagnosis is also particularly suited to Bayesian methods (Spiegelhalter 1985). Bayesian belief networks combine a number of Bayesian classifiers into a directed acyclic graph (Pearl 1986, Morawski 1989, Olson *et al.* 1990). A wide variety of inductive systems, including decision tree induction, can be analysed theoretically from a Bayesian perspective (Buntine 1991).

The Bayesian classifier is based on simple calculations with probabilities connecting states of input variables with states of output variable, represented as a matrix of probabilities. As a simple example of the operation of a Bayesian classifier from a medical diagnostic system. First, evidence of a symptom is provided. The
A high $P(dle)$ justifies predicting the disease to be present. Equation 1 can be thought of as a predictive inference that links evidence to outcomes. For example suppose that the incidence of a disease is one in ten cases, so that the prior probability of a person having the disease is 0.1. Suppose also that the probability of the occurrence of a symptom given that disease is 0.99, and the probability of the symptom occurring in the general population is 0.2. Therefore by equation 1 the probability of the disease given the symptom is:

$$P(dle) = \frac{0.99 \times 0.1}{0.2} = 0.5$$

The result of incorporation of a single piece of evidence, $P(dle)$ can in turn be the input to equation 1 if more evidence is available.

Prediction with a Bayesian classifier occurs in two stages. The first stage is the determination of the posterior probabilities of the output classes. In a general Bayesian system prediction requires us to simply calculate the probabilities of values of an output variable $d$, based on the combinations of input values or 'evidence' $e_{11}, \ldots, e_{mm}$:

$$P(d_1|e_{11}, \ldots, e_{mm}), P(d_2|e_{11}, \ldots, e_{mm}), \ldots, P(d_n|e_{11}, \ldots, e_{mm})$$

In the second stage the probabilities for the categories $d_1, \ldots, d_n$ are examined and a decision rule selects the predicted value based on the calculated probability. The one with the highest probability is chosen, $d_j$ say. Thus $d_j$ is inferred from the evidence, $e_{11}, \ldots, e_{mm} \rightarrow d_j$ if

$$\max_i(P(d_i|e_{11}, \ldots, e_{mn})) = P(d_j|e_{11}, \ldots, e_{mn})$$

This straightforward statistical approach belies one fundamental limitation of computation mechanisms. The storage needed for the joint distribution, $P(dle_1, \ldots, e_n)$ increases exponentially as the number of variables increases, rapidly surpassing the limits of any computer system. If we have $q$ variables $e_1, e_2, \ldots, e_q$ with $m$ values and $n$ possible outcomes, the total number of possible rules forming simple conjunctions of the variables: $e_1 \& e_2 \& \ldots \& e_q \rightarrow d$ is $m^{qn}$.

One solution is to limit the dimensionality of the basic units by considering only interactions between single pieces of evidence and single diagnoses. The total
number of components then increases linearly. The conditional probabilities needed for the inference may be represented as an \( n \times m \times q \) matrix of probabilities.

A modification of Bayes' theorem (equation 1) is used to calculate the result of combinations of evidence. This equation can be derived from the set definition of conditional probability:

\[
P(e|d) = \frac{P(e \& d)}{P(e)} = \frac{P(e \& d)}{P(d)} \cdot \frac{P(d)}{P(e)} = P(e) \cdot \frac{P(d|e)}{P(d)}
\]

In general Bayesian prediction evidence may point to more than one value being true at a time - i.e. more than one disease being diagnosed or a set of values being predicted. Another simplification is the property that one and only one of a variable's values may be true at a time. When only one of the diagnosis variables values is assumed to be correct the incidence for all diagnoses may then be grouped into a vector where the probabilities sum to one. The vector of probabilities prior to the incorporation of a piece of evidence is called the prior probability vector

\[
P(d) = <P(d_1), P(d_2), ..., P(d_n)>
\]

Bayes rule shows that in the case of identical and independently distributed variables the posterior probability is proportional to the product of the conditional probabilities of single variables (Press 1989):

\[
P(d|e_1, ..., e_n) \propto P(e_1|d_i)P(e_2|d_i) \ldots P(e_n|d_i)
\]

Given the prior vector \( P(d) \) and the appropriate evidence conditional \( P(e|d) \) the calculation used to obtain the posterior vector is (Pearl 1986):

\[
P(d|e) = \beta P(e|d) \ast P(d)
\]

* denotes the term by term multiplication of vectors and 

\( \beta \) is a normalising vector to ensure that the posterior probabilities sum to one, i.e.

\[
\beta = \frac{1}{n} \sum_{i=1}^{n} P(d_i|e)
\]

Under the important assumption of independence in the evidence, any two pieces of evidence presented singly are as significant as the two pieces presented in conjunction. Evidence is incorporated into the prior vector. The calculated posterior probability vector then becomes a prior probability for the next piece of evidence. This process continues until no evidence remains. The final posterior probabilities are then be sorted into order of probability and the highest probability chosen as a prediction.

The Bayesian classifier is implemented here on an expert system development shell called Learning Base System LBS (Manuscript 5). The system has been
successfully used to diagnose viral diseases in crops in Thailand (Stockwell and Dilokkunanant, in preparation) and gastrointestinal diseases in the Western Pacific Region.

**Specification of the paradigm**

The following section presents the details of the components of the predictive paradigm evaluated in the Bayesian classifier. A number of sections contain sensitivity analysis to demonstrate the effect on predictive accuracy. The results are evaluated at the end of each section.

**Languages for expressing predictions**

Models generally predict by asserting the expected value of a variable. The variables can be categorical (e.g. which horse is going to win a race), or continuous (e.g. the maximum temperature tomorrow). Prediction can also refer to the predictions of the range of a function. For example, in a soil water simulation model, soil moisture levels are predicted over the whole year (Davis, et al. 1991). The models predict a function with respect to time, rather than a single value. Hence there are two types of prediction:

- Values or states of variable in models
- Parts of the models themselves, or their relationships.

The second type of prediction is called structure discovery. This view of prediction would also include inductive processes of theory discovery, as studied by Kelly and Glymour (1990). They have examined a range of structures of increasing complexity and shown the increasing difficulty with increasing complexity of the structure being predicted. This means that the predictive success of models at different tasks cannot be directly compared, the predictive task itself has an inherent difficulty. As a simple example, it is more difficult to predict the outcome of a categorical variable with many states, than one with only two states. We cannot directly compare the predictive accuracy of say, a quantitative model that outputs the expected density of a species of animal in animals per hectare, and a qualitative model that outputs presence/absence categories.

**The Bayesian classifier system**

The Bayesian classifier only performs prediction of states of variables. It can predict any number of categories, although in this study only two were used. The data set used in experimental testing comprises seven outcome variables and eleven predictor variables. Called the Gloucestershire breeding water bird data set, the data
was collected from 33 abandoned gravel pits in Gloucestershire, England, over two years, 1987-88. The outcome variables had two categories: either presence or absence. The predictor variables were a mixture of real values and categories. The real-valued variables were grouped by an expert on the data into a small number of classes (2 to 5). Three of the outcome variables are grouped data on the total numbers of birds, total juveniles and total number of broods (roughly the number of breeding pairs). The remaining outcome variables were categories giving single species of bird: Greater Crested Grebe, Tufted Duck, Canada Geese, and Mute Swans. The input variables were conductivity, acidity, calcium content, water depth, age since the site was last used for gravel extraction, area, number of aquatic plants, number of riparian plants, water-based disturbance (categorical), bank-based disturbance (categorical), and number of islands.

A formal predictive protocol

Can we can determine the accuracy of a model on predictions of the future? In many predictive applications the data is gathered in a different place or at a different time to the evidence used to predict. We need to know how well the predictions hold across two independent data sets that are collected at different times. A predictive protocol is a relationship between the situations that provide the data used to develop the model and the situations on which predictions occur. Two sets of data are needed to develop and test a model. The training set is used to develop the model. The test set is used to establish the predictive accuracy of the model. A number of relationships between the test set and training set can be defined:

- resubstitution - the test set may be the same as the training set, also called goodness-of-fit,
- resampling - the test set and training set are based on random draws from the same data,
- cross-validation - a different data set is used for the testing as was used for development of the model,
- forecasting - the training set is collected prior to the test set,
- extrapolation - the test set contains data with values for the independent variables that are outside the range of the same variables in the training set, also called a 'hard test of a model',
- interpolation - the test set contains situations with values for the independent variables that are between values of the same variables in the training set.
- transmission - the test set is the training set corrupted with random noise.

One of the most useful predictive protocols for model development is resampling as it detects 'overfitting' of a model to data. A model is overfit when it predicts well on the data set it is developed on, but poorly on the test set. In the case of decision
trees, an overfit tree is too large, containing branches formed by data in the training set that are not found in the test set. Specifying too many parameters is responsible for overfitting of linear regression functions, and can be controlled by keeping the number of variables to a minimum (Turney 1990a). Pruning procedures help to control overfitting in decision tree induction (Mingers 1989b). A wide range of resampling methods yield similar estimates of model accuracy, although the variance can vary depending on the method (Verbyla and Litvaitis, 1989).

Comparison of resubstitution, resampling and forecasting

The Bayesian classifier was used to forecast the bird population in one year from the population in a previous year. The model was developed on one data set and tested on another using the resampling method, ten-fold validation, as a guide to the predictive accuracy. Other resampling methods provide an estimate of accuracy with less variance, ten-fold validation is a compromise between variance and computer processing time. The steps in ten-fold validation are: (1) partition data randomly into ten equally sized subsets, (2) exclude the first subset, (3) develop the model on the remaining subsamples, (4) test the model on the excluded subsample, (5) repeat steps two to five, nine times, excluding a different subsample each time, (6) calculate the predictive accuracy from the results of the tested samples through dividing the total number of correct predictions, by the total number of predictions.

The first column of Table 1 is the predictive accuracy on 1987 data as determined by resubstitution. The next column shows the predictive accuracy as determined by resampling. Resubstitution overestimates predictive accuracy in all cases, demonstrating the overfitting phenomenon. The predictive accuracy of the model developed with the 1987 data was then determined on the 1988 data set. These results are listed in the third data column. In most cases the resampling estimates are not significantly different from the predictive accuracy on the 1988 data. In one case prediction decreased, suggesting some aspect of the Great Crested Grebes changed from 1987 to 1988. On average, resampling estimates of predictive accuracy provide a good estimate of the predictive accuracy on the successive year's data set.
Table 1. The accuracy of predictions for 1988 based on resampling predictions from 1987. Column 1 is determined by resubstitution on 1987 data, column 2 by 10-fold validation on 1987 data, and column 3 by cross-validation of a model developed on 1987 data with the data from 1988. The comparison between column 1 and column 2 is the null hypothesis $H_0:1=2$ compared to the alternative hypothesis, $H_a:1>2$. The comparison between column 2 and column 3 is the null hypothesis $H_0:2=3$ compared to the alternative hypothesis $H_a: 2\neq 3$. The phenomena of 'overfitting' is shown by the decrease in predictive accuracy from column 1 to column 2. Resampling however is in almost all cases a good estimate of the predictive accuracy of the model on the next years data.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>1. 1987 resubstitution</th>
<th>Change</th>
<th>2. 1987 resampling</th>
<th>Change (α)</th>
<th>3. 1988 cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloucestershire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total adults</td>
<td>97.0</td>
<td>&gt;</td>
<td>84.3±1.2</td>
<td>= (0.05)</td>
<td>86.7</td>
</tr>
<tr>
<td>Breeding birds</td>
<td>100</td>
<td>&gt;</td>
<td>65.1±2.6</td>
<td>= (0.48)</td>
<td>63.3</td>
</tr>
<tr>
<td>Juvenile birds</td>
<td>87.9</td>
<td>&gt;</td>
<td>68.7±3.1</td>
<td>= (0.52)</td>
<td>66.7</td>
</tr>
<tr>
<td>Great Crested Grebes</td>
<td>81.8</td>
<td>&gt;</td>
<td>64.7±2.5</td>
<td>&gt; (0.004)</td>
<td>56.7</td>
</tr>
<tr>
<td>Tufted Ducks</td>
<td>90.9</td>
<td>&gt;</td>
<td>72.1±3.4</td>
<td>= (0.2)</td>
<td>76.7</td>
</tr>
<tr>
<td>Mute Swans</td>
<td>72.1</td>
<td>&gt;</td>
<td>46.5±7.6</td>
<td>= (0.38)</td>
<td>53.3</td>
</tr>
<tr>
<td>Canada Geese</td>
<td>81.8</td>
<td>&gt;</td>
<td>53.5±4.1</td>
<td>= (0.12)</td>
<td>60.0</td>
</tr>
</tbody>
</table>

As shown by the results for the Greater Crested Grebe, the predictive accuracy cannot be reliably estimated on temporally distant data. As illustrated by these results, we can ceritus paribus, or provided all things remain the same, predict the probability of a future event using a resampling predictive protocol. Of course, there are problems in showing all things remain the same. As a potentially infinite number of things could change in the natural world, application of a model relies on the necessarily subjective evaluation of whether the ceritus paribus condition holds.

Theoretically possible situations

A 'situation' is a particular combination of input values. Possible situations are sets of situations, as represented for example, by a data set of attributes of sites. All model development methods make assumptions about possible situations that can occur. If situations occur that are not anticipated, a form of violation of assumptions occurs. Multi-variate systems for example make assumptions about the form of the relationship between the variables, that the relationship is global, and that the relationship of the independent variables to the dependent variable is linear. Other functional relationship are theoretically possible, (e.g. non-linear) and the effect of violation of the assumptions can be unpredictable.
Data set used in the study.

Because the conditional probabilities in the Bayesian system relate input states to output states the system makes no global assumptions about the form of the relationship between the input and output variables. Thus a wide range of functional responses including discontinuities and non-linearity are included in the possible situations. However, Bayesian classifiers make two assumptions: the variables are independent, and only one output is possible (so-called 'naive Bayes'). A Bayesian classifier could be thought of as a diagnostic node (an output variable) surrounded by evidence nodes (or input variables) where links between nodes represent statistical correlations. Only links between the input variables and the output variable are permitted.

Some combinations of variables are consistent with the assumptions with the model and others are not. Given knowledge of the assumptions of the modelling system, we can modify the model space to meet those assumptions. In the case of the assumption that the variables are independent, covariance analysis can be used to determine the correlations between groups of variables. Covariance analysis of the data revealed a number of variables with a Pearson correlation coefficient of about 0.4. Variables were removed to give a data set with less correlation between variables, by taking into account the wildlife experts knowledge of the biological system and the information from the covariance matrix. A final data set composed of six variables contained no pair of variables with a Pearson correlation coefficient value greater than 0.22. This procedure produced two data sets with different degrees of independence between the variables. When variables are more independent the estimates of the model are more consistent with the resampling estimates (Table 2). The model estimate is the average estimated accuracy of the highest posterior probability after incorporating all evidence.

The reduction in variables, though necessary for the data set to meet the assumptions of the predictive system, removes information that may be useful. Another predictive modelling system, with different assumptions would require that different variables be removed to meet assumptions. For instance, a linear regression model would require removal of variables with a non-linear relationship to the output variable. If different predictive systems with different assumptions are used in the same situation then failure to predict could be due to interaction between the features of the data set and the assumptions of the model. Given that algorithms should be tested in situations that are consistent with their assumptions, the requirement for different variables and different information as input to the algorithms means the outputs cannot be compared.
The difference between the model estimate and the resampling estimate before and after removal of correlated variables. The degree of overestimation of predictive accuracy by the model is calculated from the difference between the resubstitution and resampling estimates of predictive accuracy. Column 1 lists the overestimates by the 11 variable model. Column 2 lists the overestimates of the 6 variable model. In most cases the 6 variable model overestimates less than the 11 variable model (Ho:1=2, Ha:1≠2).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>11 variables model estimate</th>
<th>6 variables model estimate</th>
<th>1. 11 variables over-estimate</th>
<th>2. 6 variables over-estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloucestershire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total adults</td>
<td>98.3±0.2</td>
<td>95.8±0.5</td>
<td>15.5</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Breeding birds</td>
<td>90.0±0.6</td>
<td>83.8±0.5</td>
<td>17.3</td>
<td>= (0.48)</td>
</tr>
<tr>
<td>Juvenile birds</td>
<td>91.8±0.3</td>
<td>83.7±0.5</td>
<td>19.1</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Great Crested Grebes</td>
<td>86.4±0.3</td>
<td>78.7±0.9</td>
<td>20.2</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Tufted Ducks</td>
<td>88.7±0.6</td>
<td>85.3±0.5</td>
<td>21.5</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Mute Swans</td>
<td>89.3±0.5</td>
<td>78.3±0.6</td>
<td>37.8</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Canada Geese</td>
<td>85.0±0.5</td>
<td>81.0±0.3</td>
<td>26.9</td>
<td>= (0.46)</td>
</tr>
</tbody>
</table>

A predictive decision system

In the Bayesian classifier the final posterior probabilities of the output values are sorted into order of probability and the highest probability chosen as a prediction. The selection of the maximum probability constitutes a predictive decision system. This decision to choose the maximum is only one of a number of possible strategies. An alternative strategy is, after incorporation of all evidence and a number of the output values have the close to maximum posterior probability, to output the set of values. In order to assure reasonable certainty in the predicted outcome we may decide to only predict if the posterior probability is greater than a pre-determined value, representing an acceptable level of confidence. When a number of models of the system are being used to predict, a decision rule based on a weighted average of all the models is the most reliable (Cheeseman 1990).

Variation in the decision rule

We can choose when to predict and when to refrain from prediction. For example, a gambler may choose to bet or not to bet. A rational gambler chooses to refrain from betting when the odds of winning are not justified by the returns. This approach may be used to increases the accuracy of prediction, by using posterior probability calculated by the Bayesian system, to estimate the expected probability of predicting correctly. The posterior probability can be set at any level of accuracy.
predictive decision rule with no cutoff value is compared with a decision rule with a 95% cutoff level (Table 3). The results show an increase in accuracy, but not to the level of the cutoff value.

Table 3. The accuracy of predictions selected by a model-estimated cutoff level. All predictive accuracies are determined by 10-fold validation. Column 1 is the accuracy when the model predicts on all data (cutoff 0%). Column 2 is the accuracy on data where the posterior probability estimated by the Bayesian classifier exceeds 95%. Column 3 shows the average number of data items selected. The accuracy increases in most cases and the number of selected items decreases (Ho: 1=2, Ha:1≠2).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>1. 0% cutoff</th>
<th>Change (α)</th>
<th>2. 95% cutoff</th>
<th>Total test data</th>
<th>3. Selected test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloucestershire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total adults</td>
<td>84.3±0.4</td>
<td>&lt; (0.001)</td>
<td>90.5±0.4</td>
<td>33</td>
<td>29.8±0.4</td>
</tr>
<tr>
<td>Breeding birds</td>
<td>65.1±0.8</td>
<td>&lt; (0.001)</td>
<td>84.3±1.0</td>
<td>33</td>
<td>15.7±0.7</td>
</tr>
<tr>
<td>Juvenile birds</td>
<td>68.7±1.0</td>
<td>&lt; (0.001)</td>
<td>81.5±1.6</td>
<td>33</td>
<td>16.5±0.4</td>
</tr>
<tr>
<td>Great Crested Grebes</td>
<td>64.7±0.8</td>
<td>&lt; (0.001)</td>
<td>88.4±1.2</td>
<td>33</td>
<td>10.2±0.5</td>
</tr>
<tr>
<td>Tufted Ducks</td>
<td>72.1±1.1</td>
<td>= (0.01)</td>
<td>68.4±2.0</td>
<td>33</td>
<td>14.3±0.8</td>
</tr>
<tr>
<td>Mute Swans</td>
<td>46.5±2.4</td>
<td>&lt; (0.001)</td>
<td>66.6±2.4</td>
<td>33</td>
<td>15±0.7</td>
</tr>
<tr>
<td>Canada Geese</td>
<td>53.5±1.3</td>
<td>&lt; (0.001)</td>
<td>62.5±2.7</td>
<td>33</td>
<td>10.5±0.3</td>
</tr>
</tbody>
</table>

Accuracy of prediction is not only a matter of providing a high posterior probability, although that is necessary to estimate the posterior probability of each prediction. The outcome is improved at the expense of a diminished coverage of the rule. The predictive accuracy of models from different domains cannot be compared as accuracy can be increased arbitrarily by using model estimates of the accuracy of prediction. The Bayesian classifier system provides a way of performing this more general form of prediction due to the ability to estimate the accuracy of predictions in each situation.

A criterion of predictive adequacy

The measure of predictive accuracy must be a reliable indicator of the usefulness of the model. In the case of categorical predictions, the predictive accuracy on the test set is the number of correctly predicted situations divided by the total number of situations. This measure of prediction cannot be used in a quantitative model that predicts values for a continuous variable. In this case the sum of squares of residuals, or correlation coefficients are used. In order to have a criterion, the measure of predictive accuracy needs to be compared to something. In most cases the object of comparison is the null model. Normally the value of the measure of predictive
accuracy of the null model is compared with the corresponding value for the model. It is shown below that the null model is not unique.

Another feature of criteria of predictive accuracy is that measures compress the performance of the model into a single value. A more detailed analysis might be given by a Chi-squared contingency test where each possible combination of observed to expected classes is evaluated. The results expressed in a confusion table provide a detailed analysis of the performance across all observed, and predicted classes. The accuracy can vary greatly across the classes. Similarly the correlation coefficient is a coarse measure of prediction over the whole range of predictions that assumes the error is constant over the whole range of the model. Thus criteria of predictive accuracy can be misleading when predictive accuracy is not homogeneous.

Bayesian prior probabilities

Where the primary aim of empirical modelling is prediction, acceptance of a model requires that the model predicts better than a null model. There are a number of alternatives for the null model. Bayesian theory provides alternative null models in the form of alternative prior probabilities. Below are three possible null strategies for prediction, where \( n \) and \( p \) are the expected probability of a randomly drawn datum being in the absent and present classes respectively. Below, \( n \) is the fraction of correct predictions, and \( p \) the number of incorrect predictions.

- **Uninformed strategy.** No knowledge at all leads to an equal expectation over all values (e.g. with two outcomes, 0.5). The uninformed strategy would get half of each of the alternatives correct. The expected predictive accuracy is \( \frac{n}{2} + \frac{p}{2} = 0.5 \) for all \( n \) and \( p \).

- **A statistical model.** A model of the system where the outcomes are generated at random gives the statistical null model, a strategy based on predicting the frequency of the outcome values in the proportion of the outcome values. The expected predictive accuracy is then equal to \( n^2 + p^2 \).

- **Most frequent strategy.** Another strategy that turns out to give the highest expected predictive accuracy is to predict the most frequent outcome, maximum(\( n, p \)).

The strategy of choosing the most frequent outcome in all cases leads to better prediction than the null statistical model (Table 4). Note that the most frequent strategy is superior to the model in one case where the model was significantly better at prediction than the null statistical model (Tufted Ducks). Thus a statistically significant model of a situation is not necessarily better than a null model. This demonstrates that statistical significance is not a reliable guide to prediction, particularly when the distributions of the outcomes are highly uneven. To illustrate this with a simple example, if the distribution of outcomes is 0.9 to 0.1, the statistical
null model gives an expected accuracy of 0.81 and the maximum strategy gives a predictive accuracy of 0.9. A model that predicts with an accuracy of 0.85 predicts significantly better than the statistical null model at 0.81 but would predict less accurately than the maximum strategy. The acceptance of a model is dependent on the alternatives available rather than a preconceived view of the null model.

Table 4. The predictive accuracy of the model compared with three null hypotheses. Column 1 lists the proportions of outcome values in the data set (n=absent, p=present). Column 2 is the expected predictive accuracy given an uninformed prior (n+p)/2. Column 3 is the expected predictive accuracy given a statistical prior (n^2+p^2)/2. Column 4 is the expected predictive accuracy given a strategy of predicting only the outcome that is most frequent. Column 5 is the predictive accuracy determined by 10-fold validation. The number of successful models is determined by the number of models where the predictive accuracy in column 5 is significantly greater than the expected predictive accuracy in the null model. (H0:null=resampling, Hα: null≠resampling).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloucestershire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total adults</td>
<td>0.09, 0.91</td>
<td>50</td>
<td>84</td>
<td>91</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Breeding birds</td>
<td>0.33, 0.67</td>
<td>50</td>
<td>56</td>
<td>67</td>
<td>&lt; (0.001)</td>
</tr>
<tr>
<td>Juvenile birds</td>
<td>0.30, 0.70</td>
<td>50</td>
<td>58</td>
<td>70</td>
<td>&lt; (0.02)</td>
</tr>
<tr>
<td>Great Crested Grebes</td>
<td>0.45, 0.55</td>
<td>50</td>
<td>51</td>
<td>55</td>
<td>&lt; (0.001)</td>
</tr>
<tr>
<td>Tufted Ducks</td>
<td>0.73, 0.27</td>
<td>50</td>
<td>61</td>
<td>73</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Mute Swans</td>
<td>0.67, 0.33</td>
<td>50</td>
<td>56</td>
<td>67</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Canada Geese</td>
<td>0.70, 0.30</td>
<td>50</td>
<td>58</td>
<td>70</td>
<td>&gt; (0.001)</td>
</tr>
<tr>
<td>Successful models</td>
<td>7</td>
<td>5</td>
<td></td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

A criterion for converging on a predictive system

A convergent process is one that tends towards a fixed point. A convergent modelling process tends through selection or elimination or revision of models to a fixed single model. In statistical model development such as curve fitting, after specification of the potential class of curves, the convergent process might consist in removing variables until the residual sum of squares is minimised. In machine learning of a concept, the theorist narrows down on a possible solution from a large set of possibilities (Osherson 1991). The notion of successful scientific discovery is of an investigative process that converges on a structure that is a true model of reality being studied.

In real situations, e.g. scientific discovery, the true structure of reality is unknown and so we do not know if our model is correct. We rely on predictive accuracy as a
guide to successful convergence. Predictive accuracy is only a guide because noisy data makes exact prediction impossible. Given predictive accuracy as the only guide to discovery, and that exact prediction is unlikely, an obvious criterion of convergence is that the predictive accuracy is not increasing with increasing evidence. Two forms of increasing evidence are increasing numbers of data, and increasing number of variables. A range of forms of convergence are also possible (e.g. convergence in the limit, or monotonic convergence). In monotonic convergence, used as a criterion in the following section, the predictive accuracy is non-decreasing with an increase in the evidence presented.

*Presentation of data*

Data are presented incrementally in many learning and model development systems. There are two factors to consider: the maximum predictive accuracy possible and the rate at which the model approaches the maximum with increasing data. In the case of curve fitting, more data decreases the confidence limits of estimates of parameters i.e. leads to better estimate of equation parameters. However it does not necessarily follow that more data leads to better prediction. Table 5 compares the predictive accuracy of resampling with one year to that of two years combined. In most of the cases there is no difference in the accuracy with increased data but in two cases the predictive accuracy increases with the addition of two years data.

It appears that the Bayesian classifier needs only a certain number of data for maximum predictive accuracy, and the predictive accuracy becomes optimised at a value dependent on the inherent noise in the data. If models are compared, it should be establishing that the amount of data has optimised predictive accuracy. If models are compared with a fixed set of data then two factors are affecting prediction: the inherent predictive accuracy of the model, and the amount of data needed by the model for prediction.
Table 5. The predictive accuracy with one and two year's data. Column 1 is the predictive accuracy as determined by 10-fold validation on one years data \( (n=33) \). Column 2 shows the predictive accuracy by 10-fold validation on two years data \( (n=66) \). The predictive accuracy is either the same or significantly greater in the larger data set \( (H_0: 1=2, \ H_a: 1 \neq 2) \).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>1. 1987</th>
<th>Change (( \alpha ))</th>
<th>2. 1987/88</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloucestershire</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total adults</td>
<td>84.3±0.4</td>
<td>&lt; (0.04)</td>
<td>86.5±0.4</td>
</tr>
<tr>
<td>Breeding birds</td>
<td>65.1±0.8</td>
<td>= (0.24)</td>
<td>64.3±0.9</td>
</tr>
<tr>
<td>Juvenile birds</td>
<td>68.7±1.0</td>
<td>= (0.62)</td>
<td>68.0±1.0</td>
</tr>
<tr>
<td>Great Crested Grebes</td>
<td>64.7±0.8</td>
<td>= (0.06)</td>
<td>66.7±0.7</td>
</tr>
<tr>
<td>Tufted Ducks</td>
<td>72.1±1.0</td>
<td>= (0.64)</td>
<td>72.7±0.6</td>
</tr>
<tr>
<td>Mute Swans</td>
<td>46.5±2.4</td>
<td>= (0.64)</td>
<td>47.7±0.6</td>
</tr>
<tr>
<td>Canada Geese</td>
<td>53.5±1.3</td>
<td>&lt; (0.001)</td>
<td>66.9±1.3</td>
</tr>
</tbody>
</table>

**Number of variables**

The unpredictable effect of varying the number of variables is a well known problem in simplification of linear regression models (James and McCulloch, 1990). In curve fitting the sum of squares always decreases with an increase in the number of variables or parameters. Modellers have relied on a bias towards simplification, to control the increase in numbers of parameters. The justification for this practice is questionable (Glymour 1981).
Table 6. The effect on predictive accuracy with the removal of variables. Column 1 is the predictive accuracy by 10-fold validation on an 11 variable model. Column 2 is the predictive accuracy by 10-fold validation on a 6 variable model. The difference between the two (2-1) is shown in column 3. The accuracy increased, decreased or remained the same (H₀: 1=2, Hₐ: 1≠2).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>1. 11 variables</th>
<th>Change (α)</th>
<th>2. 6 variables</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloucestershire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total adults</td>
<td>82.8±0.5</td>
<td>&lt; (0.001)</td>
<td>84.3±0.4</td>
<td>+1.5</td>
</tr>
<tr>
<td>Breeding birds</td>
<td>72.7±0.6</td>
<td>&gt; (0.001)</td>
<td>65.1±0.8</td>
<td>-7.6</td>
</tr>
<tr>
<td>Juvenile birds</td>
<td>72.7±1.2</td>
<td>&gt; (0.01)</td>
<td>68.7±1.0</td>
<td>-4.0</td>
</tr>
<tr>
<td>Great Crested Grebes</td>
<td>66.2±1.6</td>
<td>= (0.39)</td>
<td>64.7±0.8</td>
<td>-1.5</td>
</tr>
<tr>
<td>Tufted Ducks</td>
<td>67.2±1.1</td>
<td>&lt; (0.001)</td>
<td>72.1±1.1</td>
<td>+4.9</td>
</tr>
<tr>
<td>Mute Swans</td>
<td>51.5±1.9</td>
<td>= (0.1)</td>
<td>46.5±2.4</td>
<td>-5.0</td>
</tr>
<tr>
<td>Canada Geese</td>
<td>58.1±2.1</td>
<td>= (0.06)</td>
<td>53.5±1.3</td>
<td>-4.6</td>
</tr>
</tbody>
</table>

Monotonic convergence requires that the predictive accuracy increases with increasing numbers of variables. The data set composed of 6 variables was compared with the data set composed of 11 variables. The five variables removed were based on statistical analysis to find the most correlated variables and an assessment of the least useful for biological explanation (Table 6). The result was an increase, decrease or constant effect on accuracy, i.e. anything can happen. The variability of predictive accuracy shows that monotonic convergence does not occur with removal of variables.

Conclusions

This paper has examined a number of factors that can affect the predictive accuracy of a model. The empirical results imply the following effects summarised on table 7. The main findings are as follows:
A number of predictive protocols were defined. Resampling can simulate prediction of temporally distant events (i.e. forecasting). Resampling is probably the most useful general purpose predictive protocol, due to its ability to detect a common problem in modelling, overfitting.

Partial prediction, i.e. predicting only when the probability of success is high, allows higher predictive accuracy. The Bayesian classifier allows this more general form of prediction as it estimates the probability of success through the posterior probability.

Acceptance of the model is relative to the assumed null model. A statistically significant model may not be the best predictor. A Bayesian analysis based on prior probabilities of successful prediction allows evaluation of the alternative null strategies for prediction. The best alternative strategy should be compared with the model being tested, rather that a prior conceived null model of the system.

Given a monotonic criterion for convergence, the Bayesian classifier system appears to converge onto a constant value for predictive accuracy with increasing data. Evaluations of models should show that sufficient data has been provided to ensure convergence, else comparison of the model compares sensitivity of the model to data as well as inherent predictive accuracy.

Increasing the number of variables affects predictive accuracy non-monotonically. This factor is the most problematic for the purposes of simplification and adherence to assumptions of the model. While the other factors seem to behave monotonically with respect to predictive accuracy, simplicity and the choice of model structure seem particularly problematic.
Adherence to the assumption of independence of the variables increases the accuracy of the Bayesian classifier's determination of expected predictive accuracy. Adherence to independence assumptions is a relative factor as in most cases it is not possible to eliminate the correlation in natural variables entirely.

The study describes a range of factors that must be taken into consideration in evaluation of the performance of any predictive model. It would be of assistance to specify a predictive paradigm, the general form of predictive task that the predictive evaluation is simulating. A statement such as "this models predicts with an accuracy of 80%" is not very useful as an evaluation of the predictiveness of a model due to overfitting. An alternative strategy, such as predicting a fixed value, may predict with as well as or better than 80% accuracy. The saturation of the model with data hasn't been established, and the model may be predicting as well as it possible, or it may predict better given more data. Finally the model may predict better or worse depending on the variables included in the model.

A precise statement of predictive accuracy is:

"This model predicts with an 80% accuracy using ten-fold validation procedure, on 100% of the data from domain x. It predicted 50% better than the best alternative null strategy. The number of data needed to reach maximum predictive accuracy was 150. Substitution of any single variable with any of the other variables resulted in a decrease in the predictive accuracy."

The Bayesian classifier system is a useful basis to experimentally examine these methodological processes. While Bayesian classifiers can function as an efficient learning and classification system they are vulnerable to paradigmatic considerations posed in this paper. LBS is available from the author.

Acknowledgements.

Thank you Habiba Gitay and Tony Fox of the Wildfowl and Wetlands Trust for providing the Gloucestershire breeding water bird data set. Thank you also Gerlese Åkerlind and David Green for valuable suggestions.
7. A structural view of scientific explanation by explanatory modelling systems in ecology

Author: David R.B. Stockwell
Publication status: For Artificial Intelligence

Abstract: The purpose of this paper is to introduce a practical symbolic approach to explanatory modelling in ecology. The approach is based on structuralism - modelling the world through a combination of qualitative structures, and variable components of structure such as objects, and states of objects. Three levels of structural organisation are identified: events, concepts and bias, providing a path from reality, through models and explanations to understanding. A number of forms of explanatory structure are also identified. The final section reviews arguments against using structural approaches for modelling reality.

Introduction

Theories, and their mathematical expressions, models, are used in science to explain observed phenomena. Scientific explanation is possibly the most widely discussed and contentious issue in the philosophy of science. Despite the many theories and reviews that have been produced, many basic questions remain open. The following questions form the basis of this paper:

- What is explanation?
- What purpose does explanation serve?
- How does explanation lead to understanding?
- What are the limits of explanation?

The first assumption in this work is that explanations are composed of symbols arranged in definite patterns, called a symbol structure. This approach to representing the world is basic to modelling. Clearly on this definition, written language, mathematics and religious symbolism come under the definition of modelling. The second assumption is that explanation has a structural component. The idea is that there exists a framework which is 'underlying' the observable phenomena and contributes to its generation. The parsing rules of grammar express the persistent patterns in language forming valid sentences of words. The discovery of the genetic code has been interpreted in the manner of structuralism (Sibatani 1985). The genetic code is a persistent pattern generating a vast number of biological
features. The evolutionary processes that generate varied phenotypic characters are largely unknown and the physical and chemical processes at a smaller scale are largely irrelevant. At the particular level of DNA however there is an invariant structure that is essential to explaining and understanding biological systems.

The term 'structure' in science is used with a certain impreciseness to convey differing concepts such as persistent similarities in the objects of study, solid morphological components of cells, or general patterns of interactions. The meaning of structure being conveyed here is like the use of the word in common language. It is helpful to imagine the shell of a building, within which arbitrary events occur. The structure 'generates' phenomena through allowing certain configurations of events and disallowing others. This view of structure leaves open the question of whether structure actually exists, or is simply a function of human perception. The two viewpoints on this question represent a 'strong' and a 'weak' view of structuralism. A strong view of structuralism would presuppose the existence of a logical structure or principle lying deep-seated and hence invisible under surface phenomena we observe and try to explain (Stent 1978). A weaker more pragmatic view would maintain that the concept of structure is a useful tool for explaining observable, (surface) phenomena.

![Figure 1. A simple graph.](image)

In a more formal vein, structure can be illustrated using graph notation. The theory of graphs is provides an appropriate degree of abstractness and qualitative symbolism for representing features of the structuralist view of explanation. A graph consists of a finite number of vertices and edges. A vertex is a simple object, normally a variable. An edge is a connection between two vertices, expressed as a relation between vertices. An example of a graph is shown on Figure 1. We can define this graph by saying it consists of the pair $<V,E>$, where

$$V = \{A, B, C, D, E\},$$

and

$$E = \{AB, AC, CE, CD, BD\}. $$

This graph is *undirected* meaning that the edges can also be written the other way, i.e. $AB$ is equivalent to $BA$. In *directed* graphs the order of the vertices is important.
A path is a list of the successive vertices with edges between them in the graph, e.g. ACE. A graph is connected if there is a path from every node to every other node in the graph. A simple path is a path in which no vertex is repeated. A cycle is a simple path except that the first and the last vertex are the same, e.g. ACDBA.

The advent of automated explanation systems, expert systems, brings a new question to explanation: "How do we implement explanations on a computer system, and implement them efficiently?". The aim of this paper is to use a structuralist perspective to bring together these issues and present a preliminary framework for discussing and implementing explanatory computer systems in the field of ecology.

Questions of structure arise in artificial intelligence (AI) called knowledge representation. Knowledge-based systems are composed of two parts: a representation language containing the knowledge about the domain of interest, and a system for accessing that knowledge (Levesque 1986). A central concern of knowledge-based research is the services the representation system provides to the accessing system (e.g. Doyle and Patil, 1991). Typically, performance issues such as the completeness, and efficiency of computer systems for explaining knowledge are treated in the context of the structure of general representation languages and accessing systems. The form of representation, used to represent knowledge plays a large role in determining the competence and efficiency of the whole system.

Scientific applications provides other constraints on representation. Firstly, the generality of representation languages is constrained by the perceived nature of reality. One widely held view of models is that the components of the model must represent 'causal' relationships between existent, observable objects. Words such as causal, existent, and objects, need exact definition before they can be incorporated into computer systems, but this definition presents serious problems, particularly with probabilistic causation (Davis 1988). Secondly the accessing system must do more than explain what is known; it must engender understanding of the system of interest. A more comprehensive treatment of these issues follows in the section "What is the purpose of explanation".

The issue of explanation is extensively treated in biology, particularly in the fields of evolutionary theory and ecology. These fields are rich in explanatory structures, combining many disciplines, e.g. animal and plant physiology, molecular and population biology, physics and chemistry. This richness blocks any simple view of explanation in fields such as community, landscape and conservation ecology which lack a coherent body of theory (McIntosh 1980, 1987). Nevertheless there are frequent calls for a unified approach e.g. in vegetation science (van der Maarel 1989) and in geoinformation theory (Molenaar 1991). Many philosophers are documenting a 'biological way of thinking' that contributes original material to the philosophy of science (Matthen and Linsky, 1988). One aspect of biological thinking is a plurality
What is explanation?

Any attempt at excessive formalisation and looking at explanation in isolation from natural uses reveals a field so intertwined that explanation ceases to be valid as a separate research topic (Southwick 1991). A useful preliminary step to formal analysis is to analyse specific examples of a general pattern, or paradigm (Kelly and Glymour, 1990). The components of an explanatory paradigm are:

- the underlying domain being explained, (e.g. ecology)

- an abstraction of the domain (e.g. a model)

- the explanation (e.g. a component of a model),

- the recipient of the explanation (e.g. a user), and

- the explanation system, (e.g. a computer program).

Precise ways of filling out these components include mathematical equations and proofs, AI explanation systems such as rule-based expert systems, theoretical and empirical scientific theories, and informal discourse such as justifications and demonstrations. The components of ecological explanations on AI systems follows.

The underlying domain being explained

Ecological systems are composed of large numbers of interacting components. Consider an ecological hierarchy such as ecosystems, communities, species, individuals, and cells. Each of these levels is composed of a large number of distinct units. In general graph terms, each hierarchical level has a large number of vertices, one for each distinguishable unit in the level. A non-exhaustive list of general types of interactions includes birth, growth, foraging, predation, mating, competition, death, mutualism, and commensalism. In general graph terms, there are a large number of types of edges. For example, at the species level, each species can interact with many other species in its environment, forming an arbitrary network. From general graph theory, the number of edges in a graph, or interactions in a network, increases hyper-exponentially with the number of species. Viewed in this way ecology may be one of the most complex domains of scientific inquiry.
An abstraction of the domain

Abstraction of the domain is concerned with the relationship between the model and the world. There are two main schools of thought in scientific explanation regarding this relationship: the syntactic view and the semantic view. The syntactic view treats explanations as logical deductions, and relates to the world through a principle of coherence, or internal consistency. This view is represented by the deductive-nomological, or D-N, theory that treats scientific explanation as deduction from known general laws. The development, testing and evolution of theories under this view have been widely discussed in the philosophy of science literature for the last 50 years (Kitcher and Salmon, 1989). The primary test of a good D-N theory is that it is consistent with the facts and that it involves one or more logical statements framed as general laws. However the D-N theory is now regarded as inadequate as a theory of scientific explanation. There are major problems with the theory in regard to confirmation of hypotheses by facts (Glymour 1980) and in the account of the connection of the theoretical terms to reality (Da Costa 1990). One of the main difficulties arises from the relationship between the basic terms of a theory, which are true or false, and phenomena in the real world, which are existent or non-existent. Nevertheless in AI systems, explanation is often treated as logical deduction (Poole 1988, 1989, 1990).

Another reason for the demise of the D-N theory of scientific explanation was that it is not widely used in science. The widespread use of mathematical models constitutes support for a different, 'semantic' approach (Suppe 1977, van Frasson 1980, Giere 1988). The criterion of a good model in the semantic sense is correspondence with reality. On the semantic view, a scientific theory consists of the set of possible models, the structures made available for modelling the domain, where a structure has the same meaning as a graph, i.e. a set of objects (or individual events), and a set of relations, or ordered sequences, on those objects (Da Costa 1990). A recent theory of pluralistic explanation in evolutionary theory is based on 'theory clusters', complex and evolving entities formed around a conceptual core (Sintonen 1991). The idea of theory clusters is basically structuralist as it presupposes an underlying abstract structure that unites the domain.

The semantic view of theories as abstractions can be readily represented using graphs, where the vertices are objects and the edges are interactions. The objects represent observable or unobservable entities and the interactions are transfers of matter, energy or information between the vertices. While it might appear that the syntactic and the semantic approach represent two fundamentally different schools of modelling, the two approaches have been shown to be equivalent (Turney 1990b). There may be a tradeoff between the two forms which may make the syntactic approach more appropriate for some situations and the semantic preferable for others.
The explanation system.

The ability to provide explanations allow expert systems for say, diagnosing diseases from symptoms, to produce a rationale for supporting answers. When the consequences of a suggested diagnosis are great, such as a hazardous and/or invasive treatment, providing an explanation plays a role in increasing user confidence in the diagnosis. This ability is essential to teaching systems and helps the user to understand the reasoning of the machine. At least three levels of competence of the explanation system can be formulated:

- a 'one off' explanation, such as a representational model or tutorial diagram,
- a system that responds to a range of queries, such as a database system, and
- a system that interacts with a user, as in a dialogue.

In a dialogue, the recipient negotiates with the explanation giver on the subject and detail of the explanation (Gilbert 1989). This level of competence is at present beyond the ability into computers. While we do not have sophisticated natural language systems, simpler formal languages are a possible tool for providing dialogue explanations, with potential applications in ecology (Green et al. 1986, Green et al. 1988c).

The explanation

As science concerns reality, and models are conceptual constructs, a working definition of explanation is "a specific strategy for connecting observed reality and a conceptual construct" (Picket and Kolasa, 1989). This definition is very general. Two types of specific strategy can be distinguished, based on the syntactic/coherence theory and the semantic/correspondence theory of explanation:

- redescriptions of reality which compress the given observations, and
- arguments, i.e. connected assertions to support a view.

The most direct means of connecting the conceptual construct or structure to reality is to redescribe the parts of reality, one-to-one. Thus explanation comes about through picturable objects, events and their interactions. The structuralist however emphasises the abstract structures that constrain the objects and events. A later section describes in more detail six basic structures used in explanation in ecology: cause, mechanism, general laws, probability, analogy and teleology. A causal structure, for example, represents a temporally ordered chain of interactions between states of objects, represented by a path in a graph.
Logic supports explanations through proofs. Proofs do not represent causal processes, or paths in graphs, for a number of reasons (De Kleer and Brown, 1984). Firstly, an explanation proof can proceed by contradiction, i.e. the value of a variable is explained by absence of any other possible explanation. Secondly, proofs may be non-unique, contrary to the idea of there being a single cause for any particular event. Thirdly, explanation proofs may not be presented in the temporal order expected for the chain of causal effects.

The recipient of the explanation

A distinction relevant to the recipient of an explanation is to consider the goal of the explanation as a response to a question (van Frasson 1980). Questions requiring explanation may be either 'how' and 'why' questions. For example:

Q: How do you get to the Post Office?
A: Go round the corner, then down 50m on your left.

In the 'how' question the explanation is a process of making some information known in detail, e.g. by describing the temporal sequence of connected actions with objects that will result in successful completion of the task. Explanation in response to a 'why' question has a different flavour, e.g.

Q: Why are you going to the Post Office?
A: To get a stamp for my letter.

Explanation in this case is intended to support or justify an action. Clearly there are distinct differences in explanations as answers to how and why questions. An answer to a 'how' question is a description of reality; the semantic/correspondence method of explaining seems more appropriate for providing an explanation. As the answer to a why question is a justification, the syntactic/coherence theory of explanation would seem to be appropriate. Appreciation of the recipient's expectations in asking for an explanation, would result in responses more relevant to their needs and therefore more acceptable explanations.

What is the purpose of explanation?

Very often the distinction between prediction, explanation and understanding is not appreciated. The modelling process is normally perceived as an interaction between two components, data and the model itself. The process of 'fitting models to data' encourages the idea that the model is simply a concise redescription of the data. A good fit of model to data is then put forward as an objective justification of the model. This view glosses over the way that the structure of the model includes
assumptions about the appropriate forms of equation to fit to the data. For example, high significance of a linear regression model does not justify a linear model of the data. The significance simply states that the coefficient for the slope in the linear equation is probably not equal to zero.

The distinction can be made more clear by examining the meaning of the words, explanation and understanding. While a dictionary definition of explanation is "to make known in detail", understanding means "to perceive the meaning of (words, person, language, etc.)" (Concise Oxford Dictionary). Thus, the purpose of explanation is the communication of understanding. Models represent an understanding of the nature of system, and are used for explanation. Thus, good models are required to do more than fit the known facts. It follows there are three aims to a model: constructing a structure that is empirically adequate, providing an explanation, and producing in us a sense of understanding of how the world could possibly be the way it is (Cushing 1991). The following parts of this section analyse how the structure of models might complete these tasks.

This three-tiered structure is sometimes used in machine learning (Rendell et al. 1989). The first level is the event space - the set of all possible objects or data. The second level is composed of a space of models. Each concept characterises a set of events in the event space. The third level is the space of biases. A single bias characterises a set of models. For example, the parsimony bias characterises all those models that are 'simple' with respect to some criterion. The three tasks of explanation can be identified with the relationship of models to three levels: empirical adequacy with event space, providing an explanation with models, and producing a sense of understanding with bias space (Figure 2).
Empirical adequacy and the event space

A model is empirically adequate if it represents the events, and only the events in the event space. One measure of empirical adequacy is prediction in its variety of forms (Manuscript 6). The events are represented through a symbolic structure called data. Data are normally composed of logical atoms representing values of variables. For example, given a variable $A$, the following are equivalent.

"The value of $A$ is a" $\equiv eq(A, a) \equiv Aa$.

One usual structural property of a variable is mutual exclusivity of the values - that it takes one and only one value. The usual structural property of compositions of atoms in data is that they are finite lists of conjunctions, i.e. given data $x \in X$,

$$x = A_1a_1 \wedge A_2a_2 \wedge ... \wedge A_na_n$$

where $n < \infty$.

In terms of a graph structure, a variable is a vertex. Mutual exclusivity means that one and only one value fills the circle in the graph representation (Figure 3). Treating data as a finite set of conjunctions can be interpreted as the states, or values of all the variables in the graph at a given point in time.
The structure of data or data characteristics play a major role in determining the
difficulty of learning concepts (Rendell and Cho, 1990). Examples of data
characteristics are the number of data, number of attributes, scales of the attributes
(i.e. nominal, ordered or interval), error or noise, and the sampling distribution of the
data in the event space.

Providing an explanation

The statistical mathematical model characterises data by fitting of a contoured
surface that as closely as possible approximates all the locations of all the points in
the event space. Empirical adequacy in mathematical models is determined by
minimising objective functions such as the sum of squares of the residuals. A
mathematical model is both a space of events, and an hypothesis. Consider the
equation \( f: \mathbb{R} \to \mathbb{R} \). As a mathematical model the function can have a logic-based
interpretation as an hypothesis \( eq(y,f(x)) \), where \( eq \) is numeric equality. Alternatively
the function can have an analytical-based interpretation as a set of points:

\( \{ <x,y> \} \) in \( \mathbb{R} \times \mathbb{R} \) such that for all \( x \) and \( y, y=f(x) \).

Models may be partial; they fit only data in some region of the event space and
do not range of other area. A model may form an envelope by enclosing a given set
of data by defining a region in the event space, but include other points not in the
data set. A concept has the property that it fits all the given data and only the data in
the event space. A concept has been precisely defined as a sentence in universal bi-
conditional form with an atomic formula (the concept to be learned) on one side of
the sentence (Glymour 1988b). Consider the sentence

\[ \forall x: Cx \leftrightarrow Ax \land Bx \] \(\) where \( x \in X \), a space of events.

The sentence states that given a concept, the predicates A and B are true for all
the events in the event space, and that given the predicates A and B are true for all the
events in the events, the concept is true. Various form of structure for the right hand
side of the sentence are possible. Types of concept widely used are simple
conjunctive concepts \((A \land B)\), simple disjunctive concepts \((A \lor B)\), conjunctive normal forms \((A \land B\) where \(A\) and \(B\) are simple disjunctive concepts), disjunctive normal forms \((A \lor B\) where \(A\) and \(B\) are simple conjunctive concepts). In Boolean graphs, a concept has vertices composed of logical propositions, and each proposition can have the values true or false (Figure 4).

![Figure 4. A graph representing the logical concept \(C = A \land D\).](image)

Many concepts however have a structure that can be decomposed into a number of independent parts, such as sets of rules, or hypotheses. Providing an explanation in this case consists of finding the relevant parts of the model of the system for the events being explained. Explanation of a fact consists in showing those logical statements that are true or could be true in the domain that allow deduction of a particular fact or facts (Poole 1989, 1990).

Many researchers in AI are searching for a general language for knowledge representation (Doyle and Patil, 1991). A range of types of representations are used to make explanation in different domains tractable (Levesque and Brachman, 1987). In practice, task-specific subsets of logic are needed for efficient implementations of explanatory systems. Where the structure of the model allows sets of hypotheses, the interactions between them are a major factor in determining the difficulty of providing explanations algorithmic processes (Bylander et al. 1991). Characteristics that make explanations difficult are: trying to find the best explanation, incompatibility (i.e. mutually exclusive hypotheses such as \(Aa \land \neg Aa\)) and cancellation (i.e. when one element of a composite hypothesis "cancels" a datum that another element would otherwise explain). The high degree of abstraction provided by a general graph structure may assists in representing these details.

**Sense of understanding**

Consider the following requests for explanations:

How are the points arranged?
How did the points come to be arranged in this way?
The first, calls for a description of a model, a compression of the description of the events in the event space. The second also calls for a description of the model, only in this case it may go beyond what is known of the data. Rather it calls for a description of a model with some additional, possibly temporal information. In this case an explanation might consist of another model based on particles moving on a surface. For example, given a description of points in a desert landscape, the temporal model may describe the change in the distribution under some ecological process such as erosion, migration, germination, and some of the physics of particles moving under laminar flows, that could be applied in domains other than ecology. This model is very general, incorporating a set of possible models. While a concept is a useful and concise characterisation of a set of data, it does not go beyond the given data. This sense of going beyond is necessary to provide a sense of understanding of how or why a particular fact should be true. There is a need for more information to satisfy our desire to understand - to know why the world should be that way. This additional information can be supplied by bias, where a bias is defined as a preference for types of concepts.

Bias also has an important computational role in making machine learning more efficient by guiding the search into regions of the hypothesis space that are potentially more rewarding (Gordon and Perlis, 1989). Learning systems with variable bias management have been shown to be particularly effective (Rendell et al. 1989). Biases can be identified with particular types of graph structures. For example completeness (i.e. explaining all the data) could be seen as a graph with all possible edges between vertices present. A parsimonious (i.e. that no proper subset of the model explains as much data or as well as the original) graph is one without extraneous vertices or edges.

How does explanation lead to understanding?

A number of forms of natural bias are described in the following section, with reference to their appearance in ecology and their interaction with models in AI systems. These biases provide acceptable structures for scientific explanations. Thus the forms of explanations may be determined by factors other than the objective phenomena, e.g. cultural factors, convenience, and computational efficiency can be involved also. The view that explanations are not determined objectively by phenomena is shared by recent 'post-modern' views of scientific practise (Rouse 1991). Another aspect of the post-modern view is the plurality of explanation types - alternative, acceptable forms of explanation for the same event. The view proposed in this paper is that explanation constitutes a fragmentation, or set of partial representations of a graph. While the fragments can be used to reconstruct a full graph, it appears that the fragmented form is adequate for communicating understanding and explication of a model of a system. A fragmented graph has
advantages in modelling as the simpler fragments can be handled separately, rather than the complete, complex graph structure.

The following section describes six forms of explanation: causal, general laws, mechanistic, probabilistic, analogical and teleological. The actual forms are derived from three main sources: theory on forms of explanation as methods of transferring belief (Osherson 1990), an informal review of the literature in ecology conducted by the author, and an informal review of the methods used in automated explanation systems.

**Causal modelling**

Causal explanations of phenomena refer explicitly to connections between real world events. The real world events in causal models are particular occurrences - e.g. interesting pieces of the history of individual events. A causal model can be represented as paths in through a graph. The edges have a directional relationship with the general structure, 'cause-effect'. Explanation proceeds by listing the path leading to an event. A causal model would be composed of a set of fragments of paths (Figure 5). The set of paths could potentially be used to reconstruct the original directed graph structure.

![Causal model diagram](image)

Figure 5. A causal model of the graph above might be a set of possible paths through the graph, e.g. {ABD, ACE, CD, CE}.

The same qualitative structure can be seen in quantitative causal relationships. For example, the increase in a subject's response to an increasing dose of a drug, or the increase in plant growth with increasing water, nutrients, and light can be expressed as the 'cause-effect' pair of vertices, 'increasing dose-increasing response'. The qualitative causal relationship edge is interpreted as a monotonic change in the effect variable as a function of the causal variable. The relationship between quantitative equations and their conversion to qualitative causal systems has been extensively studied in AI systems. This relationship has been expressed in qualitative reasoning systems (Kuipers 1986). In these systems, variables are abstracted to a
small number of states (e.g. {below, mean, above}) and types of change (e.g. {decreasing, stationary, increasing}).

General laws

General laws are representations of the patterns in phenomena where all events are an instance of the pattern. A law is related to events by necessity, i.e. to deny the law would contradict the events. For example, to deny the law of gravitation would contradict the data on falling bodies. Laws force an economy of view, compressing a complex analysis into a single statement of fact. An explanation by a law takes the form of a specialisation of a general graph structure. For example, in Figure 6, a part of the graph is specialised into a more complex structure. The events however are still explained by the single graph structure. A general law does not admit the existence of separate graphs as models of phenomena, i.e. the graph is connected.

While unbreakable laws describe the behaviour of numerous physical phenomena, the dominion of laws in ecology is less clear. Complex chance factors in the history and evolution of organisms contributes to idiosyncratic patterns in the data. Most organisms respond partially to laws in the form of global constraints. For example the distribution of birds in North America is constrained by energetic functions of their average population density, geographic range and average adult body mass (Brown and Maurer, 1987). Thus the use of connected graphs as representation of general laws may be contingent on the level of description of events in the world.

Mechanism

Mechanistic explanation alludes to the parts of a machine and their arrangement. Mechanistic models explain events by identifying the effect parts of the entity, e.g., interactions between individuals in a group, or organs in a body or single genes in a genetic complex. A natural representation of this form of explanation as a graph structure is to fragment a full graph into pieces relevant to the event being explained.
The strategy of mechanistic explanation is to explain events by states of objects in the system, i.e. to isolate and identify the action of a single unit within a larger functioning entity (Figure 7). This process is known as reduction. Most simulations are mechanistic models. A large number of smaller but well-understood components can be used to build up a description of the system.

![Figure 7](image-url)

Figure 7. The process of mechanistic explanation consists of fragmenting a given graph into independent parts that can then be attributed to a given event.

Mechanistic explanation is often regarded as the best form of explanation in science. While mechanistic explanation has been a major source of innovation in ecology, there are many examples of central concepts in ecology that are not strictly reducible to components: species as individuals, fitness, Mendelian genes to DNA, organisms and ecosystems (Schoener 1986a). Biologists often treat organisms as distinct and indivisible entities without reference to their physiology. Some macro properties are due to statistical effects of micro properties, such as heat, air pressure and ionic potential. In cases that display cohesion, micro reduction is largely irrelevant to the explanation (Collier 1990). Reduction can also lead to complication and hence intractability (Slobodkin 1988). Consider the gas laws. If a model of an engine was framed at the molecular level of abstraction then the large number of molecules would lead to an impossible simulation problem.

**Probability**

Noise and uncertainty are constant features of biological data. Probability can quantify this uncertainty. The symbolic structure of a probabilistic explanation is composed of many components that interact weakly. This may be represented by a graph with dashed edges connecting the vertices (Figure 8). To explain using probability, the occurrence of the cause must raise the probability of the event to be explained; the event is not determined by the cause.
While ecologists use probabilistic models to explain, it is well known that one cannot argue reliably from probability to causation. The most common objection is that positive statistical relevance is merely a correlative relationship and does not represent an underlying causal relationships of the system. For instance, a regression is the line of best fit in a data set, but gives no insight into way the variables interact with each other. The structure of the explanation may thus differ from the underlying deterministic structure of the system. As shown in Figure 8, probabilistic relationships can include close but non-direct connections, or paths of length two between vertices. The directionality of a causal graph is also not present. In determining the actual structure of a system, probabilistic graphs provide a preliminary version which can later be refined by deleting edges. One of the strengths of a probabilistic theory may be its ability to provide partial explanations (Davis 1988).

**Analogy**

Analogy is an inference method where a similarity between previous and the current situation can be exploited. The bias of the representation is to use structures that are similar to more familiar structures. Analogy can be represented by two similar graph structures. Analogy is also known as 'parallel inference' as the explanation proceeds in parallel but distinct from the events to be explained (Figure 9). Analogy is not a valid form of inference, but can be useful for exploratory modelling (Carbonell et al. 1983).
Figure 9. Analogical explanation relies on the similarity between one explanation and another, represented by similarity between two graph structures, a and b.

**Teleology**

Teleological explanations refer to the purpose or aim of an event as its justification. Teleological arguments are widespread in ecology, particularly as they can be implicitly justified by the theory of evolution. The theory of evolution maintains that species survive through physiological adaptations that increase reproductive fitness. Physiological adaptations are a means of achieving the goal of survival. Thus the goal of survival explains the adaptation. In general, the event is explained by referring to the purpose, goal or function of the event. The bias in a teleological explanation is to prefer units with self-motivation, or self-organising abilities. Goal motivated behaviour may be represented using cyclical graph structure, called an autocatalytic loop, as suggested by Ulanowicz (1988). In an ideal autocatalytic loop each component activates each component in turn (Figure 10). In a model of evolution, for example, the components of the loop could be a physiological adaption (A), and evolutionary fitness (B).

Figure 10. A graph representation of a teleological model, as an autocatalytic loop. Each component activates the other component in turn, leading to autonomous, and apparently goal-directed behaviour.

Teleological explanations are used in informal discussions of the purpose of biological adaptations. A great deal of controversy surrounds teleological arguments as it suggests that individual animals and plants are able to pursue goals. A similar situation arises in the biological treatment of features of organisms such as organs. Did the features develop and then perform functions or did the need for the function lead to their development through evolutionary selection? A teleological bias ignores
the strict details of organic development and allows hypotheses that state that final features develop in order to fill a need.

**What are the limitation to the structuralist view of explanation?**

Despite the generality of the idea of using structures of symbols as an approach to representing reality, the approach has distinct characteristics with consequences for scientific modelling. The rise in popularity of neural nets in artificial intelligence with a distributed approach to representation has led to the development of two distinct camps. While neural nets have performed many of the task performed by symbol structures, it seems that the main limitation of neural nets is the lack ability to communicate their expertise in any way. This is exactly what is required for explanation. Until this ability is demonstrated, symbol structures still have the advantage of being a concise method of representing and communicating knowledge. Nevertheless, the implicit assumptions of a structural approach to modelling need to be appreciated. The objections to the structural approach are in many ways the same as those of a logicist (Kirsh 1991). These views can readily be adapted to counter the views of structuralism. The assumptions are:

1. **Pre-eminence.** Structuralism assumes pre-eminence of models of the system, i.e. the model transcends all other forms of discourse concerning the world. What is the proper role of symbolic structures in human discourse? In the context of models as tools for system management, models are merely a component in a policy debate. It has been suggested their role is largely rhetorical, and participate in the world through claims of authority (Macpherson and Grant, 1989). An unfortunate consequence of this for science is that it can mean the modeller is more important than how well the model performs.

2. **Disembodiment.** Models and the world it embodies can be studied in abstraction from the world it represents and functions in. Disembodiment discounts the importance of perceptions, measurements, and skills needed to relate the world to models. In fact, our development of models may be largely driven by our tools for measuring reality. Examples of the profound effect of technological innovation in biology, DNA sequencing, support a view that methods of analysis, and not conceptual innovation, is the basis of scientific progress (Wegmann 1988). On this view it is the character of the data we can gather that determines the nature of our models of the world.

Another area where the need for embodiment arises is in the man-machine interface. Models of the user help to tailor explanations and aid understanding (Kass and Finin, 1988). An explanation to a knowledgable user may be less detailed than given to a novice. For example, in a system for advising students on courses, the
user's goal influences the justification and the advice given (McKeown 1988). Given a simple yes or no answer to a question from a user about courses to take, the system explains by reference to one the different goals a person could have for taking that course. Thus the explanations given by the system cannot be given in isolation to the environment of the system.

3. **Symbolic.** The structural view maintains that at the deepest level of abstraction there is a qualitative model, represented by symbols and their interactions. Structural modelling is based on finding the basic units of existence and representing them with symbols. What are the units of existence and how can they be identified with symbols? There are a range of candidates for representation by symbols, e.g. objects, states, and properties of objects. Why assume that modelling requires symbols? Models of planes in wind tunnels explain the behaviour of planes in flight, and similarly, scale models of environments are used in experiments to model ecosystem processes. Symbols are *explicit*, i.e. expressed in a symbolic language that the recipient can comprehend. Recent research in the field of robotics has developed robots with complex behaviours but no central form of representation or purposeful control (Brooks 1991). The robots are composed of many similar units with individual perceptions and behaviour, and no interaction between each other. Complex behaviours such as object avoidance and foraging seemingly emerge from the parallel and independent activities. This work presents a counter-example to the notion that explicit representation of knowledge is necessary for intelligent behaviour.

4. **Stability.** The basis of a structural approach to modelling is that the underlying structure of the system does not change. For example, in a mathematical model the changes are largely restricted to variations in parameters. Changes in the overall structure of the model correspond to fundamental revisions of the model in response to, say, basic conceptual flaws in the coding of the problem. This approach assumes that there is a clear distinction between structures and parameters. In some forms of modelling, such as the induction of decision trees, the structure is the variable component. Changes in the measures used to develop the tree can produce different structures (Mingers 1989a,b). Thus the structure can be an artifact of the method of developing the model, and not related to stable components of the problem.

5. **Uniform architecture.** Uniformity assumes there is a single architecture underlying reality that can be captured in a single formal structure. Adherence to uniform architecture in the model of a system can deteriorate the performance of a modelling system. The search for a uniform method of knowledge representation will fail when timely and correct responses from an explanation system are required (Levesque and Brachman, 1987). Experience in knowledge based systems shows that implementations of more expressive languages require non-consistent fixes, to
constrain computationally expensive aspects. Given the complex nature of reality we either adopt forms of representation that allow the model to perform the tasks required of it, such as explanation, or we accept a degradation in the functionality of the explanation system, such as inefficiency and incompleteness (Doyle and Patil, 1991).

**Conclusions**

The issue of explanation has many facets. A number of the most significant issues have been explored here in the context of explanation-giving computer systems in ecology. The central notion of explanation is that it is paradigmatic, and dependent on a more general multi-component structure. The first section of this paper describes some of the particular forms the components of an explanatory paradigm may take, and the issues arising in those components.

The second section introduces the structural approach to explanation within a three-levelled framework. Each of these levels is identified with one of the aims of modelling. The first level, the event-space, participates in the determination of empirical adequacy. The second level, the concept space, is responsible for supplying an explanation. The third level, the bias, contributes to providing a sense of understanding of the world being modelled. This three-levelled view is amenable to explicit implementation on an automated explanation system.

The third section describes the main ways of biasing concepts that contribute to understanding of models in ecology. Examples of six forms of explanation are given: causal, general laws, mechanism, probabilistic, analogical, and teleological. These forms suggest structural methods of fragmenting graphs that would be appropriate in an implementation of an explanatory system.

The final section presents fundamental characteristics and limitations of the structural approach. It is clear that a number of the assumptions required for using symbol systems to model reality are detrimental to discovering the 'truth of reality'. A sceptic would thus view models as incapable of representing knowledge of the world. Arguments for this view might be that they are primarily rhetorical devices; are disembodied and hence irrelevant to the world they are trying to represent; that the use of symbols introduces unwarranted assumptions; that they are composed of arbitrary stable and variable components; and that a uniform architecture imposes unrealistic and computationally intractable limitations on an expressive symbol based system.

These arguments do not constitute a 'lethal blow' to symbol systems. Many of these objections can be addressed by modifications of symbol systems and lead to
improved performance. Symbol structures remains necessary for communicating knowledge about the world. The graph representation of structure can be automated, and is rich source of insight for understanding explanations. However, it is essentially an abstraction of the world with peculiar assumptions associated with it. These results do not suggest any one form of model is superior and illustrate the many issues associated with explanation. The structures of explanation described could guide the development of more acceptable automated explanation systems.

Acknowledgements

People who contributed to the conceptual development of this paper include Ian Noble, Hugh Possingham, Peter Menzies, Mike Dale, Bob McKay, and Habiba Gitay.
8. A survey of forms of explanation used in ecology

Authors: David R.B. Stockwell and Gerlese Sachse-Äkerlind
Publication status: Submitted to *International Journal of Man-Machine Studies*

Abstract: A number of potential users of expert systems were surveyed with a form containing examples of ecological questions with a multiple choice of explanations. Types of explanation include: causal factors, general laws, mechanisms, probability, analogy and teleology. Three hypotheses for improving explanations are tested: that the type of the explanation is important to explanation preference, the subjective preference hypothesis that individual user preferences are important for improving explanations, and the objective preference hypothesis where factors such as the domain being explained are more important. We discuss the implications for explanation by expert systems.

Introduction

Explanation is central to science as a means of communicating understanding. Yet differences of opinion between philosophers and theoretical ecologists show there is as yet no single strategy for explanation. The field of ecology in particular is pluralistic in its theories and methods of inquiry (Pickett and Kolasa, 1989). A number of forms of explanation have been identified and evaluated as a means of providing justifications for arguments (Osherson et al. 1986). Osherson et al. defined and examined the validity of types of argument on the basis of their form, i.e. syntax. The forms recognised were enthymemes (i.e. an argument with an implicit premise), reverse deduction, conditional plausibility, causal schemas, severe tests, logical form, and similarity. False prediction, gaps, and circularities were found in all these forms. The inference from this work is that no one argument form supports all needs for explanation.

In this study we first define *types* of explanation from a review of the theory and practice of explanation in ecology, related to the categories of Osherson et al. (1986). The types identified are: causal, general law, mechanistic, probabilistic, teleological and analogical. Given a recognition of a range of types of explanation, the next question is to determine the factors that affect the choice of explanation type. This choice is performed by the recipient and expressed as a preference for a particular form of explanation to a question, given all other things are equal. This question is relevant to automated explanation-giving systems such as expert systems, where the acceptability of the system to the user depends on the perceived quality of the explanations provided. The study of forms of explanation and factors influencing their use can lead to better understanding of the process of explanation and in turn, better explanatory systems. These results also have relevance to any givers of explanations, e.g. teachers, in communicating information in the most easily assimilable form.
What are the factors that affect the form of explanations? Osherson et al. (1986) hypothesised that the perceived strength or weakness of an argument might depend on the strength of the person's background convictions. A similar view is implicit in research on models of the user in modifying explanations (Kass and Finin, 1988). Early explanation systems such as expert systems were content with displaying the proof trace - or the logical premises that lead to the conclusion. There has been a recent trend for expert systems to develop models of the user to supplement the quality of the explanation. This subjective preference hypothesis states that explanations can be improved, i.e. made more acceptable to the recipient, if we modify the subjective aspects of the explanation, such as the recipients' knowledge, beliefs, and biases.

A subsequent study of Osherson et al. (1990) tested a quantitative theory of similarity arguments in the limited domain of animal physiology. These results showed that objective components, i.e. similarity and coverage of the objects in the premises of the argument, strongly determine argument strength. It follows that the domain of the argument is likely to have a considerable influence on preferences for explanations. The objective preference hypothesis is that explanations can be best improved by taking into account objective components, i.e. the subject matter (objects) being explained. One of the purposes of this study is to determine the effect of subjective and objective preferences in explanation.

In this study we conduct a survey of preferences for types of explanation. The first question addressed is "What is the preferred type of explanation in ecology?". We then examine the influence of subjective and objective factors in preferences for explanation within the general context of ecology. The subjective factors examined are individual preferences, and educational level. The objective factor is the domain or subject area of the question.

Experimental study of explanatory principles

The questionnaire consisted of six questions, each containing six multiple choice answers (Figure 1). Each answer corresponded to one of the six forms of explanation identified previously. The codes for the questionnaire and a definition of the explanation types are given in Table 1. For each question respondents were asked to indicate their preference for a type of explanation, by circling the number of the one option they most preferred. The respondents were asked to indicate their level of post-graduate education.

The respondents were also asked to indicate which explanations they found acceptable and unacceptable, by placing a tick and a cross beside each multiple choice option, and write another short hypothetical statement to explain the explanation they circled. These data are not reported in this study as the acceptability of explanation appeared to give the same information as the preference selection.
Data from the short hypothetical statement proved ambiguous and a number of respondents experienced misunderstandings about how to answer this question.

The questionnaire was self-administered, by mail or in person by one of the authors, and respondents were directed to complete the questionnaire in about 15 minutes. The survey is essentially multi-factorial. The variable of interest is the explanation type (6). The subjective factors are individual respondents (56) and educational level (2). The objective factor is the individual question (6), representing somewhat different domains within ecology.

A number of interfering factors were anticipated and we attempted as much as possible to hold these factors constant. Each explanation in a question was scrutinised to give the appearance of equal acceptability to the other explanations. In this way the possibility of rejection of an explanation simply because it could be regarded as a poor explanation was minimised. Another factor that we sought to control was the amount of information supplied in each explanation. Each explanation was adjusted to contain a similar amount of new information. An indication of our success in controlling for these factors was shown by the tendency of many respondents to express their great difficulty in deciding on an explanation due to the fact that most of the explanations seemed plausible.

Another possible interfering factor was a respondent's concern with the correctness of the content of the explanations. The participants were directed to regard all given data as true. In addition the possible influence of respondents' specific background knowledge of biology and the novelty of particular explanations was minimised by making the information contained in each question hypothetical. Nevertheless, this effect is hypothesised to be a source of noise in the results.

**Respondent groups**

Our interest was in potential users of an expert system in biology. Two groups were surveyed. One group was composed of 28 final year secondary school science students from Rockhampton Grammar School, Queensland. The second group was composed of 28 academics and postgraduate students from the Research School of Biological Sciences, Australian National University, and attendees of a conference for of the Ecological Society of Australia. The two groups allowed us to sample a defined range of background expertise and examine the effect of respondents' level of scientific education (secondary or post-tertiary) on preferences for types of explanation.
We are interested in your preferences for types of scientific explanations. Below are six hypothetical statements with six possible explanations. You need to do three things with each question:

1. Assume that the statement and the explanations are true. Place a tick beside the explanations you find acceptable and a cross beside those you find unacceptable.
2. Circle the number of the explanation you most prefer. If you think there is not sufficient information to determine which explanation is best, circle the type of explanation that most appeals to you.
3. After you have circled an explanation, write another short hypothetical statement to explain the explanation you have circled (see example below).

You should take about 15 minutes to complete the questionnaire. There are no right or wrong answers.

Example. Pickle's possum is found in Violetberry bushes because ...
1. It has a similar niche to the Tapertail possum which lives in Violetberry bushes too.
2. It's energy yield per unit of foraging time is maximised in Violetberry bushes.
3. It has an enzyme that breaks down Violetberry toxins.
4. Sightings of Pickle's possum are related to the levels of nutrients in plants, and Violetberry bushes have higher than average nutrient levels.
5. It is adapted to disturbed sites and Violetberry bushes grow in disturbed sites.
6. All it's living requirements are met by Violetberries.
Your preferred explanation is true because ...
Violetberry bushes grow in rich soils.

Q1. About 25% of Poona bears die when the temperature drops below -10 degrees Celsius because ...
1. All small animals are badly affected by cold, and Poona bears are small.
2. 5% die if temperature drops below -5 C and 50% die if temperature drops below -15 C.
3. About 25% of individuals have a gene for short hair that lowers cold weather resistance.
4. Small birds often die in winter for lack of food, and Poona bears are small.
5. Loss of weaker members increases the likelihood of other members of a family group surviving.
6. They can't find enough hollows to hide in, so they catch a form of pneumonia and die.
Your preferred explanation is true because ...

Q2. Why did the Stegasaurous, a prehistoric animal, have a long life span?
1. They produced large quantities of a hormone which slowed their metabolism for long periods of time.
2. There were few diseases and predators to kill them in those days.
3. It's like historic buildings, the bigger they are the longer they last.
4. They had to live a long time to compensate for low birth rates.
5. The Stegasaurous was large and generally large animals live a long time.
6. The probability of a Stegasaurous surviving increased enormously when it grew beyond a certain size.
Your preferred explanation is true because ...
Q3. The Cluster beetle produces large numbers of offspring in a good season. Why?
1. Nutrition and numbers of offspring are positively correlated.
2. Producing large numbers of offspring in a good season ensures the survival of the species during poor seasons.
3. The cluster beetle is r selected (i.e. produces many young in good season but few or none in a poor season).
4. Good seasons produce optimal soil conditions for the hatching of eggs.
5. Female cluster beetles produce large numbers of eggs during a good season.
6. They are like grasshoppers that breed up in good seasons too.
Your preferred explanation is true because ...

Q4. *Alto gimbits* and *Alto gumbous* are found in the same general area but rarely close together. Why?
1. You rarely see two trees growing together, do you?
2. There are chemical differences in the soils around the two different species.
3. Whenever two individuals grow together one outcompetes the other.
4. To minimise the competition between them allowing both species to coexist.
5. Both are often found together in a large area but as the size of the area considered decreases they are found together less often.
6. Two similar species cannot co-exist on the same resources.
Your preferred explanation is true because ...

Q5. Why does the Helicopter vine, a rainforest species, have its leaves facing the sun?
1. Most plants that are adapted to grow in low light have their leaves facing the sun.
2. To increase its growth by catching more sunlight.
3. The vine grows up on the sunny side of trees.
4. The probability a leaf will survive is increased by exposure to light.
5. They are drawn to the light like a moths to a candle.
6. The net result of forces generated by changes in water pressure in the stem twists it in the direction of light.
Your preferred explanation is true because ...

Q6. A large number of species of arthropod and bacteria are found on the feet of Heffalumps. Why?
1. As there are different species at different places on the foot, the total number of species on the feet as a whole is high.
2. The area between their toes is like a culture medium.
3. A large number of species maintains the health of the feet.
4. There is a strong correlation between the number of species found on the feet of Heffalumps and the size of their feet, and Heffalumps have big feet.
5. The feet come into contact with an environment rich in bacteria and arthropods.
6. Most mammals have large numbers of bacteria and arthropods on their feet, and Heffalumps are mammals.
Your preferred explanation is true because ...

Please tick your level of education:

_ primary  _ secondary  _ tertiary  _ post-graduate

If you ticked tertiary or postgraduate, what are your fields?
How long did you take to complete the questionnaire?
Types of explanation

We provide explanations as answers to questions beginning with "Why" statements or ending with "because...". The respondent is asked to choose one explanation from six possibilities. Each explanation can be classified into a type of explanation, as defined in Table 1. The types of explanation are described and defined below, and illustrated with examples from ecology.

Table 1. Codes and definitions for types of explanation: A - analogy, C - causal schema, G - specialisation of a general law, P - probabilistic, M - mechanistic, T - teleological.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q0.</td>
<td>A</td>
<td>T</td>
<td>M</td>
<td>P</td>
<td>G</td>
</tr>
<tr>
<td>Q1.</td>
<td>G</td>
<td>P</td>
<td>M</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>Q2.</td>
<td>M</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>Q3.</td>
<td>P</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>M</td>
</tr>
<tr>
<td>Q4.</td>
<td>A</td>
<td>C</td>
<td>M</td>
<td>T</td>
<td>P</td>
</tr>
<tr>
<td>Q5.</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>Q6.</td>
<td>M</td>
<td>A</td>
<td>T</td>
<td>P</td>
<td>C</td>
</tr>
</tbody>
</table>

Explanation with causal factors

Explanation by causation is a central concept in science yet is frequently used loosely. One definition of causation is that it is a decision rule, i.e. x causes y in situation C, if and only if C occurs (Granger 1988). For example, an economist interested in the cause of inflation wants to know how to control inflation. In another context, an individual or a state of affairs could be considered causes. For example, a crime may be regarded as either being caused by a person, or by the economic deprivation that the person is suffering due to the state of the economy.

The modelling of flows of substance through sinks and sources is called causal systems analysis. Systems theory in ecology is concerned with fluxes of nutrients and energy throughout ecosystems. The most well known project based on this approach is the International Biosphere Program (IBP) which attempted to model all of the major global biotic processes (van Dyne 1972). While this approach boasts a sophisticated mathematical arsenal, it is claimed that the predictive power and generality of the approach is low (McIntosh 1980).

The view of causation that seems to generalise these views is that a cause is some event that is sufficient to produce an effect. Thus a causal connection is like the logical implication, 'if precondition then conclusion'. In addition to sufficiency, the precondition and conclusion pair must be on the same level of scale and abstraction, rather than looking to phenomena at a smaller scale to explain the effect. For example, a murder is explained with the causal statement, "person x caused the death
of person y through shooting", rather than "person y died through cerebral trauma". This distinction has been called macro-causation, versus micro-causation (Collier 1990).

Definition: C-causal: The explanation is a precondition/conclusion pair, where the first condition implies the second, or the conclusion is to be expected given the precondition.
Example: Pickle's possum is found in Violetberry bushes because all its living requirements are met by Violetberries.

**Explanation with general laws**

General laws are used to account for patterns in ecological systems. For example, body mass, average local population density and geographic range of North American birds was shown to be constrained within a three dimensional space (Brown and Maurer, 1987). The regional distribution of bracken insects has shown similar constraints with population density, body size and feeding specialisation (Gaston and Lawton, 1988).

These approaches describe the environment broadly. Proponents maintain that the alternative, to reassemble communities from the characteristics of individual species, is unworkable. Qualitative arguments are normally supplied to support and lend theoretical validity to the observed patterns. For example, that body size limits dispersion distance by wind accounts for more widespread distribution of smaller species (Gaston and Lawton, 1988).

Definition: G - specialisation of a general law: The fact to be explained is some specific case of a fact that is more general and believed to be true.
Example: Pickle's possum is found in Violetberry bushes because it is adapted to disturbed sites and Violetberry bushes grow in disturbed sites.

**Explanation with mechanisms**

Mechanism is closely related to the idea of reduction, the replacement of phenomena observed on the same scale by a composition of phenomena at a smaller scale (Collier 1990). Mechanistic approaches in community ecology are those that employ inter-specific ecological concepts. One example is the use of population dynamics of inter-specific competition to determine species composition through time. Further reduction of population dynamics leads to explanation of population dynamics through relationships between individuals. The ecological study of diversity is largely mechanistic. For example, one major concern in ecology is with how species diversity might be maintained. A mechanistic approach searches for processes that favour the recruitment of individuals of rarer species over commoner ones (Connell *et al.* 1984). Thus local interactions at a small scale result in global effects at a larger scale.
A mechanistic approach is critical of general laws, universal concepts such as "niche", "mega-parameters" such as competition ratios, and coefficients (Schoener 1986a). These are regarded as "unreal". So too are arbitrary mathematical models or curve-fitting. A model must be valid, i.e. justified by a correspondence to the objects and mechanisms in the domain. Factors counting against reduction are the fragmentation and complication of the domain that replaces the universal approach. The complexity of the domain of ecology makes it difficult to study more than a handful of domains in depth, emphasising the differences between domains, and leading to a pluralistic approach (Schoener 1986b).

Definition: M - mechanistic: The event is explained by the effect of events within parts of the entity, e.g. events between individuals in a group, or organs of a body, or single genes in a genetic complex.
Example: Pickle's possum is found in Violetberry bushes because it has an enzyme that breaks down Violetberry toxins.

Explanation with probability

A probabilistic theory of explanation states that an event $x$ is caused by $y$ if the probability of the event $x$ is raised by the occurrence of event $y$, i.e. $P(x|y) > P(x)$ (Suppes 1970). The second condition is designed to eliminate statistical correlations: that the event $y$ must "screen off" all other potential causes, i.e. there is no factor $z$ such that $P(x|y \land z) = P(x|z)$.

Probability is used whenever an empirical measurement occurs. In hypothesis testing probabilistic variation is regarded as an annoyance or 'noise', and restricted to a single parameter. Probability as an integral component of explanation is most thoroughly embraced in stochastic modelling. As a simple example, treating the birthrate for a species as a probability distribution with a mean and variance, instead of a single fecundity index, results in a distribution of expected population size with respect to time (Krebs 1978). The stochastic approach is widely used in Population Viability Analysis, models that assess the probability of extinction of a species under environmental conditions (Gilpin 1990).

The probabilistic approach has been criticised as being purely descriptive of statistical relationships between variables and providing no explanation of causes of events. Nevertheless, the theory of probability is the basis of statistics and widely used in science. The advantage of the approach is that it allows analysis and modelling of empirically measurable and repeatable events under uncertainty.

Definition: P - probabilistic: The occurrence of the explanation raises the probability of the event to be explained.
Example: Pickle's possum is found in Violetberry bushes because sightings of Pickle's possum
are related to the levels of nutrients in plants, and Violetberry bushes have higher than average nutrient levels.

**Analogy**

Analogy is widespread in conversational language, and in the communication of unfamiliar concepts. Analogy also provides the motivation for new applications of models. For example the analogy of games or contests for individual interactions has led to the application of game theory to explain many aspects of animal behaviour, such as patterns of ownership of females in the baboon, and strategies of parenteral investment in the raising of offspring (Maynard Smith 1976, 1977).

An analogy is clearly not a valid logical inference. As an inference that runs in parallel to the desired inference but is not a part of it; analogy always fails to hold at some point. It is however a powerful means of exploring, or communicating unfamiliar ideas. By drawing on similar but more familiar explanations of phenomena we can suggest new explanations for unfamiliar phenomena that might be the same.

Definition: A - analogy: The explanation cites some similar example to the case to be explained.

Example: Pickle's possum is found in Violetberry bushes because it has a similar niche to the Tapertail possum which lives in Violetberry bushes too.

**Teleology**

Among all the sciences, biology is one of the few admitting teleological explanation, e.g. explanations of the functions of organs - we speak of having hair in order to stay warm, teeth for cutting or grinding, and so on. Some prefer to be more precise, stating that the purpose of thorns, say, is to defend a plant from herbivores, because thorns were favoured by natural selection (Loehle 1988).

Teleological explanations are criticised for ascribing anthropomorphic intentions to animals and plants - suggesting that they are conscious of the goal of their actions, adaptations and mutations. On the other hand it is a concise form of expression. As long as evolutionary selection, and not intentional states of an organism can be readily assumed, no error should arise.

Definition: T - teleological: the event is explained by referring to the purpose or goal or function of the event.

Example: Pickle's possum is found in Violetberry bushes because it's energy yield per unit of foraging time is maximised in Violetberry bushes.
Data analysis

The variables of interest were: all respondents, individual respondents, individual questions, and respondents' level of scientific education, in terms of the most preferred explanation.

We thus consider the following null and alternative hypotheses in response to each of the questions.

1. Is there a consistent preference across respondents?

$H_0$: Grouping the results for all respondents, the preferred explanation type is evenly distributed.
$H_a$: There is a consistent preference across respondents for some explanation types.

2. Are preferences consistent within respondents?

$H_0$: That the distribution of explanation types chosen by individual respondents is the same as the distribution of the types chosen by the group as a whole. $H_a$: The preference was more consistent within than between respondents.

3. Are preferences consistent across questions?

$H_0$: That each question produces the same distribution of explanation types as the distribution of responses as a whole. $H_a$: Some types of explanation are preferred for particular questions.

4. Are preferences the same at different levels of scientific education?

$H_0$: That the level of scientific education makes no difference to the distribution of preferred explanation types. $H_a$: The level of education affects the preference for explanation type.

Chi square tests were used to compare the distributions. The degree of freedom is equal to the number of categories in the distribution, minus one. The expected distributions of responses in the categories is determined by assuming a random generation process. The observed distribution are the survey results.

Results and Discussion

Is there a consistent preference across respondents?

The missing value category was excluded from analysis of the first hypothesis, giving six categories and five degrees of freedom. We pooled all responses of both groups and questions and analysed the frequency of response in each category. The null hypothesis, that there is no preference for an explanation type was rejected ($P(\chi^2)>0.005$). The histogram in Figure 2 illustrates the preferences for types of
explanation. Causal explanation is most preferred constituting 26% of the total responses. General laws, mechanistic and teleological are the next preferred occupying 18% of the responses. Probabilistic is the next preferred at 14%. Analogy is the least preferred at 5%. Most of the contribution to the analogical response is due to question six and without this question the proportion would have dropped to 0.004% (one response).

Figure 2. The distribution of responses by explanation type. Codes are defined in Table 1. X refers to missing values (e.g. non-response to question).

Are preferences consistent within respondents?

Hypothesis two was tested using a data set generated by simulating random responses to the questions with a distribution of categories identical to the survey results. Once the class distribution of the simulated data had been tested against the survey data to verify that the program operated correctly, the distribution of the number of similar responses for each person was determined for the two data sets and compared with a Chi square distribution having six degrees of freedom.

The null hypothesis, that individual respondents have no stronger preference for a type of explanation than the group as a whole was not rejected. Thus the distribution of the number of identical responses that a respondent gives on a questionnaire does not deviate significantly from the expected distribution given by purely random responses (Figure 3). There is no tendency for individual respondents to prefer certain types of responses more than others.
Are preferences consistent across questions?

In testing hypothesis three each of the responses from each of the questions were tested against the original total responses for all questions. The distributions were compared using a Chi square test with one degree of freedom. The null hypothesis that a question does not differ from the average distribution was strongly rejected on every question (Q1: \( P(\chi^2) < 0.01 \), Q2: \( P(\chi^2) < 0.005 \), Q3: \( P(\chi^2) < 0.005 \), Q4: \( P(\chi^2) < 0.025 \), Q5: \( P(\chi^2) < 0.005 \), Q6: \( P(\chi^2) < 0.005 \), 5 degrees of freedom). The frequency of each form of explanation for each question is given in Figure 4. The most frequent type of explanation response for each question was respectively: Q1 - mechanistic, Q2 - causal, Q3 - probabilistic, Q4 - mechanistic, Q5 - teleological, and Q6 - causal.

There is obviously a strong relationship between the question and the explanation type. As we have attempted to eliminate the specific content of the question, subject areas, or domains are the main factor determining explanation types. The concept of domains is a fundamental concept in AI, based on the idea that sets of phenomena are fundamentally different, and need different classes of methods for problem solving. These results imply that different domains require different forms of explanation. The first question, where a mechanistic explanation was preferred, concerns physiology. Physiology is conventionally explained in a mechanistic fashion: i.e. by reference to internal organs. The mechanistic explanation is also preferred in question four. Competition between organisms is also normally explained mechanistically: i.e. by reference to the physiological basis of competitive advantage. Question two where a probabilistic explanation was preferred, concerns the relationships of fecundity to environmental factors, usually presented in the form of...
simple correlations with nutrients. Question five with the strong teleological response, concerns functional adaptation, the classic domain of teleological explanation. The questions related to causal phenomena are more difficult to relate to existing studies. In question two, prehistoric animals are studied through the fossil record, which permits only reference to other hard-bodied animals. Hence the reference to predators seems likely. The causal preference in question six, that Heffalumps have a large number of species of bacteria between their toes due to contact with them in the environment, could be explained by the evoking of images of large animals in zoo with dirty feet.

Are preferences the same at different levels of scientific education?

In hypothesis four each of the two sub-populations was tested against the original population with a Chi square test and four degrees of freedom (one less for the constraint that the sum of the two populations equals the total population). The null hypothesis that there is no difference was not rejected ($P(\chi^2) < 0.1$). Hence there is no significant change in preferences for types of explanation with level of education in biology (Figure 2). In fact there is some support for rejecting the hypothesis that the samples are drawn from different populations ($P(\chi^2) < 0.9$). This result shows that preference is not strongly affected by educational experience in biology, at least not in comparison between secondary and post-graduate level of study.
Conclusions

The most significant pattern to the responses was that the respondents as a group preferred causal explanations to mechanistic and general law forms, with probabilistic, teleological and analogical explanations seldom preferred. Analysis of the number of similar responses for each question however, showed that the preferences of individual respondents for explanation types are no stronger than the preferences of the group as a whole. Thus there is no support for the effect of individual preferences on choice of explanatory types. In addition, the differences in level of education did not differ significantly from the overall pattern of preferences. These results reject the subjective component hypothesis for the factors examined here.

The strongest effect on respondent preferences for types of explanation was the individual question. Some caution must be observed in interpreting this as support for the objective component hypothesis however, as the response to individual questions could be due to simply the content of the explanations, rather than relationships between the domain of the question and the form of the explanation. Although differences in the apparent content of explanations was minimised to the extent that each of the possible responses were plausible explanations for the question, this possible effect cannot be completely discounted.

In favour of the alternative, that relationships between the domain of the question and the form of the explanation determine explanation preferences, it is reasonable to accept that the form of explanation and the content of the question may be linked to conventional means of explaining phenomena of that type. One possible reason for this is that our perception of the appropriate structure of domains is formed by our common cultural experiences. If this is true, we can conclude that an explanation may be more acceptable if the explanation type is matched to a particular question, i.e. the results would support the objective component hypothesis.

The main results of this exploratory survey are that there are strong preferences for explanation types, and that objective factors related to the phenomena being explained are more important than subjective factors such as the characteristics of the recipients in determining the most preferred explanation. This result is consistent with a view no doubt held by many modellers - that the 'top level' choices of forms of models to adequately represent the domain being modelled are of vital importance to the result of the modelling exercise. The results also suggest a reason for poor explanation by models. If this top level choice is made incorrectly it can lead to models being perceived as having inadequate explanatory power. The consequences of these results are that research into explanatory systems should be directed to modifying the explanation type to match the domain rather than to preferences of individual users. Further study is required into the characteristics of the domains that determine the types of explanation.
Acknowledgments

Thank you John Hamilton of the Rockhampton Grammar School for assisting in the survey of secondary students. Peter Menzies advised on the selection of explanation types for the questionnaire. Thanks also to the survey respondents from the Ecological Society of Australia and the Ecosystem Dynamics Group, Australian National University.
Section 4. Rule set induction for prediction and explanation

A system was developed for learning sets of rules from data (Manuscript 9). The genetic algorithm for rule set induction, GARP, adapts the heuristic search methods of genetic algorithms to develop sets of predictive and explanatory rules. The predictive capacity of a rule was determined from its accuracy. Prediction and explanation were achieved by the selection and application of rules that meet both goals. Stability of structure was determined by monitoring the variation in the number of rules in the rule set (Manuscript 10). The role of biases in stabilising the number of rules in the rule set was demonstrated.

Finally the application of the system to the interpretation of rules in a biological situation was illustrated (Manuscript 11). Rules can determine the factors sufficient for the presence of waterbirds, and hence be used for prediction. They can also determine the factors necessary for waterbirds, and hence produce information important for management of wildlife. Finally, the necessary and sufficient conditions for controlling the waterbirds were determined, useful information for controlling waterbirds. Thus a range of logical relationships within a given data set is determined using induction of rule sets, and used to provide a range of predictions and explanations relevant to ecological questions.
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.

\[ V_1=v_1 \land V_2>v_2 \land \ldots \land V_m=(v_{m1},v_{m2}) \]

where \( v_1, v_2, \ldots, v_m \) are values of the variables \( V_1, V_2, \ldots, V_m \).
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 $\{<1,1,1>, <0,1,1>, <1,0,1>\}$ then the selector $V_1=1$ selects the data set $\{<1,1,1>, <1,0,1>\}$. 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.

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

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...
RULE SET INDUCTION

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) -> $V_1=1,2, V_2=1$, the mutate operator, changing the value of a variable in a rule at random. e.g. $V_1=1,2),V_2=1$ -> $V_1=[0,1),V_2=1$, and the concatenate operator, concatenating two existing rules e.g. $(V_1=[1,2)), (V_2=1) -> V_1=[1,2),V_2=1$.

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].

153
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 $E(\Pr(C_i|M_j))$. Thus, $\Pr(C_i|M_j)$ calculated from given data is a measure of the usefulness of the rule $M_j$.

However, estimation of $E(\Pr(C_i|M_j))$ from $\Pr(C_i|M_j)$ is misleading when there are low numbers of samples, spurious associations, or high prior probabilities of the class $C_i$. Instead of $\Pr(C_i|M_j)$, 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. $\Pr(C_i|M_j) > \Pr(C_i)$). A satisfactory measure of statistical relevance with a standard normal distribution is:

$$z = \frac{n_1 - np_1}{\sqrt{np_1(1-p_1)}} > \alpha \text{ and } n_1 > 4.$$  \hspace{1cm} (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 $\Pr(C_i|M_j) > \Pr(C_i)$. 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 class 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 $V_i=1$ at $t_t$ then $V_i=0$ at $t_i$). There
INDUCTION OF SETS OF RULES

Juveniles

<table>
<thead>
<tr>
<th>age=[0,10)</th>
<th>age=[10,20)</th>
<th>age=[20,30)</th>
<th>age&gt;30</th>
</tr>
</thead>
<tbody>
<tr>
<td>undef.</td>
<td>C&gt;300</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Rules

| 1<age=[10,20)&C>300 | 0<age=[20,30)&aq=[10,20) | 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

10. The effect of bias on prediction and explanation in induced rule sets

Author: David R.B. Stockwell
Publication status: For Machine Learning

Abstract: This paper reports the development of a machine learning system called GARP (Genetic Algorithm for Rule-set Production) for prediction and explanation using sets of rules. A series of experiments are conducted to determine the optimal parameters for predictive accuracy and the recovery of underlying structure in artificial and natural data sets. The best prediction was achieved using a combined significance and probability rule, averaged over a number of rules. Application of bias was necessary to allow the algorithm to converge onto a stable set of rules, however different biases gave better prediction in different data sets. The implications for the development of predictive and explanatory models are discussed.

Introduction

Rule sets have been used as a form of knowledge representation since the first expert systems were developed (e.g. Davis et al. 1977), but developing rule sets manually is time-consuming. There are many advantages in analysing data by machine learning instead of interviewing experts: accuracy, speed, rigour, and automation (Manuscript 5). One of the first machine learning systems for developing rule sets from data was called INDUCE, and diagnosed diseases in soybean plants (Dietterich and Michalski, 1981). Another, BEAGLE, automated acquisition of knowledge from data bases (Forsyth 1981). A general methodology for learning rule sets called Star, used the diagnosis of cancerous cells as an application (Michalski 1983a). Despite the apparent success of these systems, much subsequent work on classification rules has been based on hierarchical forms of rule called the decision tree, such as ID3 (Quinlan 1986). Nevertheless, rule-sets can have advantages over decision trees: robustness and informativeness (Manuscript 9).

Rules are particularly interesting to a modeller as they are applicable to qualitative data, complementing primarily quantitative statistical methods (Jeffers 1991). Rule sets may also be viewed as qualitative models, a factor seen to be of particular value in ecology and forestry science (Stockwell 1990). To consider rule sets as models, in contrast to mere exploratory methods, they need to be able to both predict and explain. Recent work has shown that both prediction and explanation are multi-faceted concepts, with many possible forms (Manuscripts 6, 7, and 8). The predictive nature of rules is seen by restating a rule "if X then Y" as "if X then predict Y". The if-then rule also supports interpretation as a cause-effect relationship, "A causes B". The causal form of model was shown in a survey of students and academics to be the most preferred form of explanation in ecology (Manuscript 8).
GARP finds rules with a genetic algorithm. Genetic algorithms are an effective method for discovering useful patterns in data (De Jong 1988). However, the effectiveness of the algorithm depends critically on system design choices such as the representation language and performance criteria (Forsyth 1981). Little is known about the effect of these factors and their relationships to types of data (De Jong 1988). One of the most difficult problems in practical induction is determining the way the design choices interact with the data sets analysed (Turney 1991). In practice the factors are manipulated experimentally until the algorithm meets set criterion of adequacy in the application domain - a practise is known as 'tweeking'. This paper attempts to contribute to the understanding of the effect of design factors on genetic algorithms developing predictive and explanatory sets of rules through controlled experiments with artificial and natural data sets.

Design choices can be formalised using the concept of bias, i.e. a prior preference for types of representation language or model. Bias is often necessary for inductive or machine learning because the potential number of hypotheses to explain a set of facts is very large, and biases reduce the set of hypotheses to a manageable size (Michalski 1983a). Bias can determine the success or failure of an algorithm in an inductive task (Gordon and Perlis, 1989). Because the biases in an algorithm are rarely made explicit, researchers struggle with the application of a vast number of subtle biases (Michalski 1983a). Formalisation and the quantification of bias has lead to important results on the effect of particular biases on algorithms for learning particular types of structures (Haussler 1988). However, in the application of learning systems to real world induction problems with little background knowledge (e.g. animal response to the environment and diagnosis of diseases), we do not know the form of the underlying structure, and hence do not know prior to analysis what form of bias would be best. The only recourse in this case is to experiment with a range of biases in an attempt to optimise performance of the system.

While bias may be justified on pragmatic, theoretical, or statistical grounds, there are also good reasons for developing methods of data analysis that use less bias, or can manipulate bias flexibly. Less bias allows the development of systems with greater autonomy and generality, as less assumptions are made about the underlying structure. Flexible bias management systems have the generality of a less biased system, and the computational advantages of a strongly biased system (Rendell et al. 1989). It seems likely that flexible bias management will therefore be a feature of learning systems in commercial applications for automated modelling, using large data sets in domains with little background knowledge.

An automated approach to model development is becoming a major priority in natural resource information centres (Green 1991). We need a greater understanding of the best forms of bias to use on data sets of natural origin. In order to determine the usefulness of a range of biases in optimising the performance of the GARP algorithm, this study uses a data set of natural origin containing information on topology, vegetation and wildlife density. The performance of the algorithm on a
range of biases is compared with performance on an artificially generated data set with a known, simple, underlying structure. The primary concern is to specify the requirements for induction of predictive and explanatory rules, how rule sets perform under those requirements, and can be optimised for prediction and explanation using ecological data sets. Discussion of the results follow.

**Methods**

The choice of a form of representation suitable for modelling a particular aspect of reality is a basic design decision in a modelling system, e.g. GARP is based on 'if-then' rules. A given set of rules can vary in the number of rules, their generality, and whether more than one rule in a rule set can apply in a particular situation. Achieving explanation and prediction involves design decisions in the modelling system with implications throughout the whole of a modelling system. The experimental component of this paper consists in manipulating these biases to simultaneously maximise the goals of prediction and explanation.

The forms of prediction used in this study are resampling and resubstitution. Resampling simulates prediction on independent data, as would occur in forecasting, by developing the model on a test set independent of the set used to test the model (Manuscript 6). Resubstitution simulates goodness of fit measures, where the data used to develop the model is also used to test the model. While the fit of model to data is widely used as a test of the utility of a model, it does not detect 'overfitting', or the fit of the model to noise in the data. When models overfit, or overestimate the predictive accuracy of the model, resampling methods can establish the unbiased accuracy of the model (Verbyla and Litvaitis, 1989). Comparison of the resubstitution accuracy to the resampling accuracy gives us a good estimate of the degree of overfitting taking place.

Explanation is defined as structure discovery, the output of the actual underlying model or structure of the system by the induction algorithm (Manuscript 7). Structure recovery is the basis of a formal characterisation of learning which includes scientific discovery (Kelly and Glymour, 1990). An artificial data set is generated with known underlying rules and used to test the performance of the algorithm at recovering those rules. Where the underlying structure is unknown, as in data sets from the real world, stability of the recovered model as indicated by the number of rules in the output model, serves as an indicator of explanatory performance. Stability has been shown to be a useful concept in addressing the question of the best structure to use in curve fitting by linear regression (Turney 1990a). Stability is also a fundamental requirement of a theory of explanation based on underlying structure (Manuscript 7).

First some terminology must be mentioned. *Rules* are precondition-conclusion pairs. That is, they have the form, "if *precondition* is true, then *conclusion*".
model is the precondition that determines the situations where a rule is applied. The model space is the possible preconditions, a set of possible situations that can be identified by rules. Bias refers to an explicit preference for a type of model, i.e. some region of the model space. We now describe the realisation of these concepts in the GARP system.

Rule sets

The precondition of a rule is used to determine if a rule is applicable to a given situation. In most simple rule-based classification systems the precondition is a list of conditions. The conclusion is the category or class predicting the rule. In contrast to a functional representation of a dynamical change (e.g. \( y=f(t)=Ae^{kt} \)), rules are both partial and non-deterministic. Partial relationships, in contrast to total relationships, are those that only apply over part of a range of a variable. For the rest of the range of the variable the value is undefined. Thus if we have a rule, if \"A=a_1 then B\", then the response to the situation \( A=a_2 \) would be "undefined". A non-deterministic relationship, in contrast to a deterministic one, means that more than one rule can apply in a given situation. Therefore unlike a function where a single value is defined for a given input, a set of rules can produce a number of alternative and possibly contradictory conclusions. Given the appropriate combination of rules, rule sets can potentially be total and deterministic, but are not necessarily so. Thus they make less assumptions and are potentially more informative than total and deterministic models (Manuscript 9).

The sets of rules used in GARP constitute a total model:

\[ M = \{M_1, \ldots, M_j\} \cup \{M_0\}, \text{ where } j \text{ is the total number of rules}. \]

\( M_j, i \leq j, \) is a single rule, referred to as a partial model. A given set of rules may contain many or no rules that apply in a particular situation. To ensure that some rule applies in all situations, a null precondition, \( M_0 \), is assigned an output class for all rule sets, \( M \). The null rule has no conditions in the precondition and therefore applies in all situations.

The precondition is composed of a conjunction of conditions, i.e. all conditions must be true for the whole precondition to be true. The precondition \( M_j \) is composed of zero to \( m \) conditions, \( V_1=v_{11}, V_2=v_{21}, \ldots, V_m=v_{m1} \) where

\( V_i \) is a categorical variable, \( i=1 \) to \( m \) the total number of variables,

\( v_{ij} \) is the \( j \)th value of the variable \( V_i \),

A rule is written

\[ M_j \rightarrow c_i \, (u) \] where,

\( M_j \) is the precondition, \( i \leq j \)
Rule set induction

$c_i$ is a class in $C_i$, the set of output classes, where $i = 1$ to $n$ is the total number of classes.

$u$ is the measure of the utility of the rule.

A rule set $M$, is a set of rules with the above form, e.g. $M = \{M_1, M_2, M_3\}$

- $M_1: A_1 = a_{11} \rightarrow c_1$
- $M_2: A_1 = a_{12}, A_3 = a_{31} \rightarrow c_2$
- $M_3: \rightarrow c_3$

Model space and the data

Data items are the symbolic representations of situations in the world. The data, $D$, are single points in the Cartesian space, $V=V_1 \times V_2 \times \ldots \times V_n$, formed by the variables, $V_1, V_2, \ldots, V_n$. A rule relates to the data through the precondition. A precondition selects a subset of $D$ for which the precondition is true, i.e. $M_j(D) = \{x \mid M_j(x) = \text{true}\}$ where $x$ is an element of $D$. The space of possible rules or partial models is the set of possible preconditions. In the representation chosen in GARP this space is composed of the 'boxes' formed by values of variables $V_1 \times V_2 \times \ldots \times V_m$, contained in the total cartesian space, $V=V_1 \times V_2 \times \ldots \times V_n$, where $m \leq n$. For example a model $V_i = v_{11}$ selects a part of the space $V$ containing all points where $V_1 = v_{11}$, or $v_{11} \times V_2 \times \ldots \times V_n$.

One data set, natural, describes the distribution and habitat variables of Greater Gliders, from forest inventory measurements as described in Stockwell et al. (1990) (Manuscript 4). The data set is composed of 400 data items with eight variables. Each variable has between two and six values. The other data set, artificial, is an artificially generated data set containing relationships between variables defined by four rules with a single precondition. The number of data, variables and values of the variables in the artificial data set are designed to match the natural one. The data set was generated by generating data items at random and testing if any of the rules apply. If a rule did not apply then the outcome of a fraction of the rules was changed to the conclusion specified by the rule. A fraction of the data items not covered by the rules were also allowed into the data set. In this way a controlled amount of noise was introduced into a data set, with an underlying structure based on the specified rules.

Bias on the model space

Bias is a term for the computational and structural design choices in an induction algorithm. Bias is imposed in three ways (Gordon and Perlis, 1989). A syntactic bias is a restriction on the language used to express models of the system. A structural bias, defined on the space of possible models, is a distribution or a preference ordering on the possible models. A computation bias applies a preference ordering through the search strategy used to explore the space of models. If the
actual structure is not among the set of structures searched, then the induction system can perform poorly. For example, feature vectors are very poor representations of graphs, giving a large number of very long descriptions for a simple graph structure (Quinlan 1990). This is an example of a mismatch between the syntactic bias of the modelling system and the syntactic bias of the underlying data set. The biases used in GARP are now described.

**Syntactic bias**

A set of rules $M$ is a collection of subsets of the space $V$. From elementary set theory, the set of all possible sets of points in $V$ is the power set of $V$, $P(V)$. As a rule set can select any combination of points in $V$ it can select any set in the power set $P(V)$. Thus unrestricted rules used in GARP are an unbiased representation, under the definition of 'shattering' of the hypothesis space introduced by Haussler (1988). Thus the first form of bias is the unrestricted space of rules.

The second form of syntactic bias is number of conditions in the precondition, or the maximum permissible length of the rule e.g.

Length 1: \{a\rightarrow c_1, b\rightarrow c_1, d\rightarrow c_2, \ldots\}

Length 2: \{a,b\rightarrow c_1, b,c\rightarrow c_1, d,e\rightarrow c_2, \ldots\}

When we start to put restrictions on the types of rules we can use then it may become impossible to select some sets in $P(V)$. For example, if the rules are restricted to single conditions in the precondition it is impossible to select the set $D' = \{x \mid V_1=v_{11} \text{ and } V_2=v_{21}\}$ where $x$ is an element of $D$, as any two rules will contain additional points not in $D'$. This bias is related to parsimony - that hypotheses should be expressed simply. Decision trees are typical of induction methods that impose a bias for shorter representations during the selection of hypotheses (Buntine 1991). The preference for short rules removes long specific descriptions that use more memory, are cumbersome to manipulate, and are more difficult to communicate. Sacrificing longer rules may mean the loss of accurate predictors however.

**Structural bias**

Structural bias uses measures of utility on rules. The two measures of utility examined in GARP are posterior probability, and statistical relevance. Given a finite set of possible outcomes, we should predict that class which in the long run is most probable. The long run probability is the expected probability of a class $E(P(C_i))$. The expected probability cannot be determined accurately from small data sets, so the predictive accuracy is only a guide to the predictive ability of the rule.
The posterior probability of an outcome in a particular situation, is written $P(c_i|M_j)$ where $M_j$ applies in a situation usually supplied by a datum. The posterior probability is a guide to the expected probability of the classes of $C$, given $M_j$, and hence provides a guide to the best rule to apply in a particular situation. An estimate of the expected probability of the class $c_i$ from $M_j$, $E(P(c_i|M_j))$, is given by the conditional probability $P(c_i|M_j)$ calculated from given data

$$P(c_i|M_j) = \frac{P(c_i \text{ and } M_j)}{P(M_j)} = \frac{\text{no. of data with } c_i \text{ and } M_j \text{ true}}{\text{no. of data with } M_j \text{ true}}$$

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 $c_i$. Hence using the posterior probability as a measure of the utility of the rule can be inaccurate when the rules are to be applied to independent data sets. The posterior probability of an outcome for a given data set contains no indication of the confidence we should have in that value, or whether the outcome is relevant to the precondition.

Another estimate of utility is 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. $P(c_i|M_j) > P(c_i)$). Statistical relevance is used in theories of probabilistic causation (Davis 1988). A satisfactory measure of statistical relevance with a standard normal distribution is:

$$z = \frac{n_1 - np}{\sqrt{np(1-p)}} > \alpha \text{ and } n_1 > 4. \quad (1)$$

where $n_1$ is the number of correct classifications, $p$ is the expected fraction of correct classifications, and $n$ is the total number of data (Mendenhall et al. 1981, pp550).

Note that only rules which give a positive value for $z$ 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 other difficulties when there are more than two classes of conclusion. In this case, a rule 'if $A$ then not $c_i$' says nothing about the other classes in $C$. It is therefore not useful for prediction and was not used in the present implementation of GARP.

**Computational bias**

Computational bias, or bias in search, refers to the tendency of the algorithm to focus its efforts on evaluating particular kinds of rules. Three forms of search bias are used: *unbiased* or random, *concatenation*, *mutation*, and *generality*. A search is unbiased if the selection of all hypotheses is equally likely, i.e if the search is randomly distributed throughout the space of possible models. A random unbiased
search occurs in the initialisation stage of genetic algorithms where a number of hypotheses are selected at random.

Given two rules in a rule set, concatenation is the combination of their preconditions, e.g.

\[ M_1: \quad A_2 = a_{21} \rightarrow c_1 \]
\[ M_2: \quad A_1 = a_{12}, A_3 = a_{31} \rightarrow c_2 \]

becomes

\[ M_4: \quad A_1 = a_{12}, A_2 = a_{21}, A_3 = a_{31} \rightarrow c_2 \]

Concatenation leads to rule sets with longer, more specialised rules.

Mutation changes a value of a rule at random.

\[ M_2: \quad A_1 = a_{11}, A_3 = a_{31} \rightarrow c_2 \]

becomes

\[ M_2: \quad A_1 = a_{12}, A_3 = a_{31} \rightarrow c_2 \]

General rules are those that cover the most examples. For example a rule with a single condition \( V_1 \) is more general than a rule with the conditions \( V_1 \) and \( V_2 \). To apply generalisation we only select rules that are maximally general, i.e. there is no more general rule with the same output class. By comparing the preconditions we can exclude the specialised rules from a rule set, e.g.

<table>
<thead>
<tr>
<th>Not generalised</th>
<th>Generalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>a ( \rightarrow ) c1</td>
<td>a ( \rightarrow ) c1</td>
</tr>
<tr>
<td>b ( \rightarrow ) c2</td>
<td>b ( \rightarrow ) c2</td>
</tr>
<tr>
<td>e,d ( \rightarrow ) c3</td>
<td>e,d ( \rightarrow ) c3</td>
</tr>
<tr>
<td>a,c ( \rightarrow ) c1</td>
<td></td>
</tr>
</tbody>
</table>

The main reason for applying generalisation bias is to simplify the model by reducing the number of rules in the rule set, while retaining number of situations where a rule will predict. Generalisation bias may also eliminate some irrelevant rules. A specialised rule may be predictive simply because it is covered by the more general predictive rule, i.e. the specialisation is not relevant to the prediction. For example, given the rules \( a,b \rightarrow c_1 \) and \( b \rightarrow c_1 \), the rule \( a \rightarrow c_1 \) may not be predictive. On the other hand, specialised rules can be more accurate than the general rule, and predictive accuracy could be lost.
GARP - Genetic Algorithm for Rule Production

Empirical results show that genetic algorithms are efficient algorithms to use in search spaces with many local optima (De Jong 1988). They can find rules not only in spaces with a global maxima but also in spaces with many local maximas. This is because genetic algorithms essentially conduct a large number of searches in parallel, each rule acting as an independent search of the model space. Genetic algorithms are *adaptive search techniques*. An adaptive technique starts with an initial set of inferior outputs (e.g. rules), and increases the quality of the rules through iterative improvement by heuristics, or 'rules of thumb' that modify the current rules. The increase in quality, or utility, of the rules is tested after application of the heuristics and used to maintain a size-limited set of best rules so-far. Because the algorithm maintains and improves a current set of best rules found so-far, the model can be evaluated at any time. One can take advantage of the "80/20 rule", that in most cases 80% of the performance is achieved in 20% of the time. Using the set of current best rules we can also monitor the rate of increase in performance, a measure of the efficiency of the algorithm.

```
procedure GENETIC begin
  t = 0;
  Initialize M(t);
  while (not Termination condition) do
    begin
      t=t+1;
      Evaluate M(t);
      Select M(t) from M(t-1),
      Test M(t);
      Recombine M(t);
    end
  Final test M(t)
end
```

Figure 1. Pseudo-code for the genetic algorithm

The algorithm

The pseudo-code for GARP is given in figure 1. The first step is to initialise the current rule set with a set of rules generated at random, *Initialise* $M(t)$. Fifty rules are generated, the number of rules maintained in the current rule set. The utility of these rules is determined from the data with the function *Evaluate* $M(t)$. The algorithm then enters a loop, called a generation. Passing through a *Termination condition*, the number of generations, the program selects a set of rules with a utility better than a given pre-set value, *Select*(M). This selected rule-set is tested to determine the
predictive accuracy, \( Test(M) \). The current rule set is then used to develop a new rule set by applying the heuristic operators, \( \text{Recombine}(M) \). Each of the recombination operators is applied twenty times in each generation. The algorithm returns to the termination condition, repeating the loop and eventually terminating at ten generations. A final test is applied, \( \text{Final Test}(M) \), such as testing the output rule set on an independent data set.

Testing is different to evaluation, as evaluation is a measure based on a single rule, while testing determines the utility of the entire rule set. There is a tension between the utility of a rule as determined by the single rule and the utility of the rule in association with a rule set. When the measure of utility of an individual rule is an imperfect guide to the usefulness of a rule in the rule set, the function \( \text{Select}(M) \) may not select the best rules to include in the rule set for \( Test(M) \).

The procedure used in \( Test(M) \) to determine the best output to assert as a prediction given a set of rules is called a decision strategy. Two decision strategies are examined: the maximum posterior probability and the weighted average of probability. The maximum posterior probability strategy chooses the class \( c_i \) given by the rule \( M_j \) that maximises \( E(P(c_i|M)) \).

\[
E_{\text{max}} = c_i \text{ s.t. } \max_j P(c_i|M_j)
\]

There are advantages in an evaluation rule taking all possible models into account when trying to determine the expected probability (Cheeseman 1990). Choosing the best rule, the one with the highest posterior probability is effectively jumping to conclusions. A weighted average strategy for estimating \( E(P(c_i|M)) \) sums the probability for each class over the rules that apply in that situation.

\[
E_{\text{av}} = c_i \text{ s.t. } \max_j \left( \sum_j P(c_i|M_j) \right)
\]

The posterior probability and the significance constitute two possible measures of utility.

**Experimental Methods**

GARP was written in PDC Prolog and run on an IBM compatible PC (AT). Two protocols for developing and testing models are used: resubstitution and resampling. In resubstitution, the data set used to develop the model is used to test the models analysed. The method of resampling used was to split the data set into a training set and a test set using a random draw of datum without replacement, and then develop and testing of the model on the separate data sets. This was repeated six times to determine average behavior of the algorithm.

The main concern in a modelling algorithm is to maximise predictive accuracy. To do this the rules with the best predictive potential should have the highest utility.
The utility of the rules is determined at two stages, \textit{Evaluate(M)} where the rule is evaluated for inclusion in the set of current best rules, and \textit{Test(M)} where the predictive accuracy of the rule set is tested using the rules in the current best rule set. This means the effects of two different measures of utility need to be determined at two possible places. The first results concern the effect on predictive accuracy of measures of utility in the \textit{Evaluate(M)} and \textit{Test(M)} stages. The predictive accuracy for combinations of these measures is determined for resubstitution and resampling.

The second set of results concerns explanatory capacity, defined as the ability to recover the underlying structure of the data set. The ability of the algorithm to recover the rules used to generate the artificial data set is determined. Stability of the rule set is an issue in both the artificial and the natural data set. The effect of forms of bias: syntactic bias, restriction on rule length, and generalisation bias, on stability is shown. Finally the interaction of these forms of bias on the predictive accuracy of the algorithm is determined.

\textbf{Results}

\textbf{Effect of measures of utility on prediction}

Table 1 shows the effect of posterior probability and statistical relevance at the \textit{Evaluate(M)} stage. Table 2 contains the results of two decision strategies at the \textit{Test(M)} stage: maximum utility versus a weighted average of rule utilities. Each contains the predictive accuracy for resubstitution (predicting on the data used to derive the model) and resampling (predicting on an independent data set) and the difference between resubstitution and resampling. An asterisk represents a significant difference at the 95\% confidence level while two asterisks represent a significant difference at the 99\% confidence level. All errors are standard errors of the mean of six trials.

\textbf{Evaluate(M)}

Table 1 shows the predictive accuracy for combinations of the \textit{Evaluate(M)} criterion and data set. The \textit{Test(M)} criterion was held constant at maximum probability. The first two rows of table 1 compare the accuracy when the \textit{Evaluate(M)} criterion is varied between probability and statistical relevance. The only significant difference was increased predictive accuracy in the resampling test on the artificial data set (6.2±1.5\***).
Table 1. Influence of the Evaluate(M) criterion on predictive accuracy on (a) the natural data set and (b) the artificial data set.

<table>
<thead>
<tr>
<th>Measure</th>
<th>a. Natural data set</th>
<th>b. Artificial data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resubstitution</td>
<td>resampling</td>
</tr>
<tr>
<td>Probability, Maximum Probability</td>
<td>58.7±0.7</td>
<td>47.2±2.6</td>
</tr>
<tr>
<td>Relevance, Maximum probability</td>
<td>58.5±1.1</td>
<td>49.8±1.4</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.2±1.3</td>
<td>2.6±2.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>a. Natural data set</th>
<th>b. Artificial data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resubstitution</td>
<td>resampling</td>
</tr>
<tr>
<td>Probability, Maximum Probability</td>
<td>65.9±0.6</td>
<td>54.0±1.0</td>
</tr>
<tr>
<td>Relevance, Maximum probability</td>
<td>62.8±1.2</td>
<td>60.2±1.1</td>
</tr>
<tr>
<td>Difference</td>
<td>-3.1±1.3</td>
<td>6.2±1.5**</td>
</tr>
</tbody>
</table>

Test(M)

Table 2 contains a comparison of the different Test(M) criteria, maximum posterior probability and weighted average probability, with a constant Evaluate(M) criterion (statistical relevance). The weighted evaluation rule outperforms maximum rule in one case, the resampling prediction on the artificial data set (3.4±1.6*). In all other cases the maximum rule does not significantly increase the prediction in the natural data set, although the consistent increase suggests that a significant increase in accuracy might be gained with a greater number of runs of the program to reduce the variance.

Table 2. Influence of the Test(M) criterion on predictive accuracy on (a) the natural data set and (b) the artificial data set.

<table>
<thead>
<tr>
<th>Measure</th>
<th>a. Natural data set</th>
<th>b. Artificial data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resubstitution</td>
<td>resampling</td>
</tr>
<tr>
<td>Relevance, Maximum probability</td>
<td>58.5±1.1</td>
<td>49.8±1.4</td>
</tr>
<tr>
<td>Difference</td>
<td>2.0±1.7</td>
<td>0.6±2.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>a. Natural data set</th>
<th>b. Artificial data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resubstitution</td>
<td>resampling</td>
</tr>
<tr>
<td>Relevance, Maximum probability</td>
<td>62.8±1.2</td>
<td>60.2±1.1</td>
</tr>
<tr>
<td>Difference</td>
<td>3.4±1.6*</td>
<td>2.1±1.5</td>
</tr>
</tbody>
</table>

168
Forms of prediction

Significant differences occurred between resubstitution and resampling as shown by the differences listed in columns of Tables 1 and 2 marked 'Difference'. The resubstitution accuracy is consistently higher than the resampling accuracy, indicating overfitting of the model to the training data. The degree of overfitting is much higher on the natural data set than on the artificial data set in all but one case, "Probability, Maximum Probability" in Table 1.

These results point to small but significant effects due to different measures and selection rules, allowing determination of the best criteria for general application of the algorithm. The increased predictive accuracy for resampling using combinations of statistical relevance criterion for the selection function, and weighted average criterion for the testing function suggest these will give the best results in general application. The advantage of a conjunction of measures is not consistent however as shown by the few significant increases in predictive accuracy.

Overfitting was the most noticeable effect on predictive accuracy however, with differences of over 10 percent between the accuracy as determined by resubstitution and resampling. As shown by the significance of the differences, none of the combinations of measures eliminated overfitting completely. The introduction of the relevance measure decreased the degree of overfitting significantly (9.3±2.0**). There was no significant increase in overfitting due to the weighted average decision strategy.

Ability to explain

The concept of stability of a model is fundamental to a structural view of explanation (Manuscript 7). Without stability the symbolic representation of the underlying structure of the system would be constantly changing. If the structure does not stabilise then the continued change leaves open the question of when the structure of the natural data set has been recovered, or when to terminate the algorithm. Stability is thus a necessary condition for explanation, but does not guarantee that the actual underlying structure has been found, i.e. it is not sufficient. The method of determining stability adopted in this paper is to monitor the number of rules in the rule set.

Structure recovery on the artificial data set.

The rule sets generated by GARP always contained the four rules used to generate the artificial data set. While the algorithm was successful in recovering this underlying structure, the rule sets also contained additional rules not intentionally introduced into the data set. Some of these rules were specialisations of the
underlying rules, others were unexplainable. This lead to an investigation of the stability of rule set during the operation of the algorithm, and methods of using bias to control the number of rules in the rule set.

The predictive and explanatory performance of the algorithm on the artificial data set is compared in Figure 2. The predictive accuracy quickly reaches a maximum, but the number of rules keeps increasing. While the number of rules must eventually plateau as the number of rules in the model space is finite, up to 200 rules were discovered in long runs of the algorithm. As shown by the flattened curve for
predictive accuracy in Figure 2, larger rule sets do not increase the predictive accuracy.

*Use of syntactic and computational biases*

Bias leads to different explanatory structures and affects predictive accuracy. Theoretical study of bias is only of partial help. Ultimately we need to determine the appropriate bias for the data sets of interest (Turney 1991). In ecological modelling, such as animal habitat prediction, the details of the underlying structure is normally unknown (Davey and Stockwell, 1991). We cannot predict what form of bias will be most efficient at prediction or explanation. A method that works perfectly well at one data set may fail unexpectedly on another. Hence there is a need to experiment with biases to determine their effects on predictive accuracy and structure stability.

![Graph](image-url)

Figure 3 (a) The number of rules remains constant with number of generations when the rules are constrained by the number of conditions and (b) by imposing the generalisation bias. Each symbol represents a single run of the algorithm. The upper row of symbols are the number of rules generated without bias, included for comparison.
Figure 3 illustrates the effect of biases on the increase in the number of rules. Using syntactic bias, a restriction of the precondition of the rules to a single condition, the number of rules in the rule set becomes constant with computation time. With generalisation bias, few rules are found, and the number increases very slowly relative to the unbiased case. The constraint imposed by the syntactic bias is very strong, it restricts the number of possible rules to a small number (5.5 on average). The constraint imposed by generalisation is less severe, and allows specialised rules providing they are not covered by existing rules.

**Interaction of bias with predictive accuracy**

Table 3 lists the predictive accuracy by resampling and resubstitution on the natural and artificial data sets with three types of bias: no bias, syntactic (rule length) bias, and maximally general (generalisation) bias. Statistical relevance was used as the evaluation rule and the weighted average used as a test rule. In the natural data set, generalisation bias gave the highest predictive accuracy for resampling (53.4±0.8, with $\alpha=0.04$ for the probability that the value is equal to the lowest value). In the artificial data set however, syntactic bias gives the highest accuracy in resampling (64.2±0.6, with $\alpha<0.002$ for the probability that the value is equal to the lowest value). On the artificial data set, generalisation bias lead to poorer prediction than no bias at all in the resampling prediction, while on the natural data set, syntactic bias lead to poorer prediction than no bias at all under resampling.

| Table 3. The resubstitution and resampling predictive accuracy with a number of biases. |
|----------------------------------------|--------|--------|--------|--------|
|                                       | Natural |        | Artificial |        |
|                                       |        |        |          |        |
| Resubstitution                        |        |        |          |        |
| No bias                               | 60.5±1.3 | 50.4±1.4 | 62.8±1.2 | 62.3±1.1 |
| Syntactic                             | 52.8±1.9 | 47.9±2.4 | 62.5±1.0 | 64.2±0.6 |
| Generalisation                        | 60.8±1.3 | 53.4±0.8 | 63.7±0.5 | 58.7±1.2 |

Generalisation bias performed better on the natural data set, while the syntactic bias performed better in the artificial data set. The superiority of syntactic bias on the artificial data set is due to consistency of the bias with the underlying structure of the data set. As the data set was generated by a small number of rules of length one, the syntactic bias helps the induction algorithm to find the best set of rules by constraining the recombination to a space of possible models that contains the underlying structure. The generalisation bias may be superior in the natural data set because there are useful rules with a number of conditions that do not have useful generalisations. Clearly different types of bias are better for different data sets.
The optimum rule-length on the natural data set was examined by plotting maximum allowable length of rules against the accuracy (Figure 4). The highest accuracy was achieved when the length of rules was constrained to four. The accuracy increased as the number of conditions increased, and decreased as the number of conditions decreased. Lower accuracy with shorter rules could be due to the constriction of the model space so that potentially good rules are not found. The drop-off at the ends may be due to one of two factors. Very specific rules could accumulate that are detrimental to the predictive accuracy of the system, or the larger space of possible rules means the search encounters less good rules.

![Figure 4. The relationship between accuracy and the number of preconditions in a rule.](image)

**Discussion and conclusions**

Two main conclusions can be generated from these results concerning the GARP algorithm, the effects of bias on induction of rule sets, and the problem of developing predictive and explanatory models in general.

1. The GARP algorithm and appropriate biases can be used to predict and discover underlying structure in data sets. A combination of statistical relevance criterion for the selection function, and weighted average criterion for the testing function gives the best predictive accuracy. The need for two different measures has interesting implications for the prediction explanation problem (Manuscript 2). Posterior probability is primarily a predictive measure, as choosing the rule with the highest value is bound to give the best prediction in a resubstitution regime. Statistical relevance is an explanatory measure, being linked to probabilistic theories of causation (Davis 1988). Either measure is inadequate for prediction on its own. With posterior probability as a measure of utility, some rules with high utility are represented by few data, and hence may be due to chance events. They therefore
EFFECTS OF BIAS

perform poorly as predictors on an independent data set. Rules with high statistical relevance may have a low posterior probability due to a low prior probability, and therefore predict poorly. The combination of the two measures eliminates cases which diminish the predictive accuracy of a rule set.

2. Bias affects predictive accuracy and structure recovered. If the bias is consistent with the underlying structure of the data then it leads to better prediction. A generalisation bias gives better prediction on the natural data set, suggesting that there are higher order effects, i.e. specialised rules, that prove useful for increasing prediction under resampling. Even with generalisation bias the number of rules continued to increase slowly, indicating that the underlying model of the natural data set contains specific rules that are not specialisations of more general rules. The better performance of the rule-length bias on the artificial data set shows that the effect of bias is dependent on the underlying structure of the data set. Simply making the bias less restrictive does not necessarily increase predictive accuracy as the search space of the algorithm is increased, making the discovery of good rules more difficult.

Positive and negative features of the use of bias emerge from these results. The advantages of bias are stability in the search of large hypothesis spaces, and in some cases better prediction. On the negative side the underlying rules may not be found under a certain bias, and inappropriate bias may decrease predictive accuracy. Explicit manipulation of bias is therefore necessary for optimising the performance of predictive modelling systems. These results provide strong support for learning systems to incorporate flexible bias management.

Acknowledgements

Thank you Habbiba Gitay, and Ian Noble for helpful discussions. Stuart Davey and N.S.W. Forestry Commission provided the natural data set.
11. Using rule-sets to explain animal response to the environment

Authors: David R.B. Stockwell and Habiba Gitay
Publication status: For Ecological modelling

Abstract: The application of rule induction to explanation of the biotic response of waterbirds to their environment is described. The nature of explanation in models and sets of rule, and the use of explanations for answering wildlife management questions is defined. While the rule sets fitted the data as well as the regression method, the explanations provided for the data set varied between the two methods. Rule sets can provide a unified method of modelling biotic response based on an alternative theory to continuum modelling.

Introduction

A wide variety of techniques are used for modelling the response of animals and plants to the environment, i.e. biotic response (Gauch 1986). Many are powerful in appropriate conditions, but most existing techniques have limitations: unjustified assumptions, restrictions in applicability, and difficulty of interpretation of the model in terms of ecological processes (Bayes and Mackey, 1991). Expert systems based on sets of rules show promise in modelling in ecology, but the time consuming process of eliciting rules from experts, and lack of knowledge about many species limits their widespread application (Davey and Stockwell, 1991). Rapidly developing predictive models from data using classification rules or decision trees has also shown promise. However, problems have emerged with treating decision trees as explanations: instability under induction and difficulty of interpreting the tree in terms of theory (Stockwell et al. 1990, or Manuscript 4). This paper examines the use of a new tool for modelling biotic response based on the induction of sets of rules, called GARP, Genetic Algorithm for Rule-set Production (Stockwell and Noble, 1991 and Manuscript 9). The method is applied to a data set of bird counts on a number of man-made water bodies subjected to various disturbances. GARP has been developed with the purpose of producing stable small sets of accurate rules on natural data sets (Manuscript 10). It is possible to examine an alternative explanatory model of biotic response in the context of the continuum theory of Austin and Smith (1989).

This paper aims to apply the system to a practical wildlife modelling application and illustrate the use of rules as explanations of biotic response to the environment. Rule set induction and linear regression modelling techniques are compared as explanations for the presence and absence of animals. The explanations serve as answers to questions about wildlife management based on the response of wildlife to the environment. The first section details what it means to explain with a model, and how explanatory models are used, i.e. the questions that can be answered by the explanation. The empirical sections follow, with methods, results and conclusions.
Explaining with a model

The view of explanation here is based on a formal learning theory, called structure recovery. On this view the purpose of a machine learning system is to correctly determine the true or 'underlying' logical structure of a data set (Kelly and Glymour, 1990). The data must be consistent with the underlying structure but noise in the form of errors in the data can be present. In order for a proposed model of the underlying structure to satisfy the requirements of a scientifically adequate model, it must satisfy three goals:

- the model must predict well,
- the model must provide an explanation for data, and
- the model must contribute to our understanding of the system.

Fitting the data

In scientific discovery there is no direct indication that the proposed model is the same as the underlying structure of the system. Because of this, predictive accuracy must be used as an indication of the similarity of the model to the underlying structure. How well the structure predicts the data is indicated by a measure of goodness of fit, such as the residual sum of squares when the output data are continuous. When the dependent or output variable is categorical (e.g. the presence or absence of a bird) then a measure of fit is the accuracy of the model at predicting the output category on a given data set.

Explanation for data

In the case of a rule set the structure is a set of if-then rules. A set of rules in the GARP system has a structure:

\{(if \ A_1 \ then \ B_1, 
if \ A_2 \ then \ B_2, 
... 
if \ A_n \ then \ B_n)\}

where the \( A_i, i=1 \) to \( n \) are called preconditions and specify when the rule is applied, and the \( B_i, i = 1 \) to \( m \) are predicted output classes.

An example of a rule set in GARP is:

if \( V_1=1 \) then Species=present
if \( V_1=2, V_2=[10,20] \) then Species=absent
else Species present
The structure of a linear regression is:

\[ \text{Species} = V_1 + V_2 + \ldots + V_n \]

where \( n \) is the number of variables.

The proposed model is the explanation of a set of data. The proposed model fits the data (is predictive) and is a subset of possible models that could be proposed for the system. For example, the multiple regression equation proposed as the underlying model is usually composed of fewer variables than the total number of variables in the data set. This is called the reduced model. The reduced rule set is similarly a small subset of the possible rules.

Explanations of particular data are conveyed by components of the reduced model. If the output is high in a linear regression model then we might refer to one of the variables in the model that contributed most to this high value as an explanation for the high value. In a rule based system, the specific rules that contributed to the prediction would be identified as an explanation. For example, given two rules in a reduced model, 'if islands are present then Tufted Ducks are present', and 'if islands are absent then Tufted Ducks are absent', then an explanation for the presence of Tufted Ducks might be that islands were present, and that if islands are present then Tufted Ducks are present'.

Reduction of a model is not sufficient for understanding however. The components of the model also need to be consistent with a general theory. One form of general theory found to be acceptable in ecology is causal explanation. While it is not asserted that all models represent causes of the output phenomena, they may often be a 'first approximation' to causal relationships. Further analysis may eliminate correlated input variables, further reducing the number of variables. Belief in an explanation is supported by the sense that the components of the model are approximately causal, i.e. the factor is sufficient to produce the observed effect in the output variable.

Rules are open to causal, and other, interpretations. A rule can represent some relation between the precondition and the conclusion such as a cause and effect process, a transition between states, or a simple statistical correlation. For example the rule \( A \rightarrow B \) could be interpreted as variously, 'if A then B', 'A causes B', 'A changes to B', or that 'A and B are co-occurring'. The data set alone does not contain all the information to determine the interpretation of a rule. Usually additional knowledge is needed about the functioning of the system.

Models can be compared with the above framework. Using prediction, explanation and understanding we compare the performance of rule sets and linear regression. Given a rule set and a linear regression developed on the same data set, we compare the predictive accuracy, the structure of the models and their components, such as the variables that are included in the reduced model. Finally we
compare and contrast the integration of the components with theories of biotic response to the environment.

**Answering questions with a model**

One of the tasks of a modeller is to provide answers to questions about the system being modelled. The answers to questions are explanations. The following sections examine particular types of questions and the particular types of explanations they require, and hence, particular types of multi-variate and rule-based systems.

Typical questions an ecologist may ask are:

- What factors predict the occurrence of a certain species?
- What ranges of variable values does a species need to exist?
- Are there certain parameters that control the occurrence of a certain species?

We look at these question in relation to multi-variate methods and rule sets and how they can be answered.

The first question concerns prediction of the distribution of animals. Given data on certain environmental variables and the presence and absence of the animals within the range of the data set, how do we accurately predict distribution in sites not covered by the data? Multivariate correlation methods are frequently used to determine the answers to these questions (e.g. a linear model, as used for comparison in this study).

Prediction is more exactly the inference of the presence or absence of the birds. Therefore the question can be more exactly stated as 'What evidence would allow us to infer the presence or absence of the birds?' From basic logic, we know that we can infer a fact if the premises are sufficient for the truth of the fact. Hence, the question becomes, 'What factors are sufficient for inferring the presence and absence of the birds?' The answer to such a question is an implication \( A \Rightarrow B \). A rule can be interpreted as an implication and hence used to predict. The use of rule sets for prediction is reported in Manuscript 10.

Translating the second question into more precise language again we have, 'What factors are necessary for a species to exist?'. From elementary logic again we know that necessity is the inverse implication, \( B \Rightarrow A \). That is, the answer to a question 'What environmental factors are necessary for a species to exist?' is equivalent to predicting the environmental factors from the occurrence of the animals. Ordination, where species are positioned in a space defined by environmental factors is often used for answering this question. We could find rules that determine the necessary conditions for an animal, e.g. the presence of \( A \) is necessary for animal \( B \) to occur.
That is to say that if the animal occurs then the condition A is always present. This is a rule $A \Leftrightarrow B$; the direction of the arrow is reversed. These factors may be of interest to a wildlife manager for determining factors that should exist in a reserve, or be created.

The final question we consider is 'Do certain parameters control the number of birds?' In most cases an experiment may be necessary to answer this question. Given that all other factors remain constant, a single factor would be increased, the effect on the animals noted, and then decreased and the effect noted again. If the factor controls the population, the population of the animals would correspondingly increase and decrease. Translating this into more formal language again we have 'If a factor A is present, then animals B are present, if a factor A is absent, then animals B are absent'. If we treat presence and absence as the logical values, truth and falsity, then the formal statement above becomes:

\[ A \Rightarrow B \text{ and } \neg A \Rightarrow \neg B \]
\[ = A \Rightarrow B \text{ and } A \Leftarrow B \]
\[ = A \Leftarrow B \]

Thus the question of factors controlling the animal occurrence is equivalent to a logical bi-conditional, or 'if and only if' relationship. In rules, this is a necessary and sufficient condition, i.e. 'What factors are both necessary and sufficient for the occurrence of birds?' Thus answers that satisfy both of the first two questions satisfy the final question. The outcome variables represent an answer to questions about important environmental factors for the animals, that are necessary (in a logical sense) for their presence or absence. The rules which go both ways, or necessary and sufficient, represent a very strong assertion of a relationship between the environment and the response. That is, there is virtual equivalence between the animal and the stated environmental condition.

**Methods**

The data was collected from 33 abandoned gravel pits in Gloucestershire, England over a period of two summers. The variables consist of seven outcome variables and eleven predictor variables. The outcome was categorised into a presence and an absence class. The predictor variables were a mixture of real values and categorical. The real-value variables were grouped into a small number of classes (2 to 5) determined by inspecting the data for clusters in the values. Two of the seven outcome variables are reported in detail in this paper. The first is the presence and absence of any adult birds. The second is the presence and absence of a single species, the Tufted Duck. The values of the predictor variables are shown in Table 1.
The statistical package SYSTAT was used to generate the linear regression models. The variables regarded as significant had a significance greater than 95% on a standard normal two-tailed test. The complete model was reduced by eliminating those variables that were not significant, except for the constant term, and refitting the data. The linear regression models were used to predict the presence or absence of species by assigning a value of 1 or 0 to the output of the model, 1 if the value was above 0.5 and 0 if below.

The rule-set induction system called GARP was used to generate the rule sets (Stockwell and Noble, 1991 and Manuscript 9). Data represent single observations of the value of a number of variables at a single point in space and time. They are formally described as elements of a cartesian product of the variables:

\[ V_1 \times V_2 \times \ldots \times V_n = \{ <v_{11}, v_{12}, \ldots, v_{1n}>, <v_{21}, v_{22}, \ldots, v_{2n}>, \ldots, <v_{m1}, v_{m2}, \ldots, v_{mn}> \} \]

The predictive accuracy was determined by resubstitution, i.e. using the data set on which the models were developed to determine the predictive accuracy. The number of data were insufficient to support the determination of accuracy on other methods of estimating prediction, such as resampling. As the major concern of this paper is with explanation, resubstitution is a simple and adequate form of determining the goodness of fit.

Table 1 The variables and their values in the Gloucestershire breeding water bird data set. MV is a missing value.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Symbol</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>total adult birds</td>
<td>totad</td>
<td>0 (absent), 1 (present)</td>
</tr>
<tr>
<td>Tufted Duck</td>
<td>TD</td>
<td>0 (absent), 1 (present)</td>
</tr>
<tr>
<td>conductivity of the water (μSiemens)</td>
<td>c</td>
<td>[0,200),[200,300),[300,9999)</td>
</tr>
<tr>
<td>acidity of the water</td>
<td>pH</td>
<td>[1,7.8),[7.8,8.0),[8.0,8.3),[8.3,14)</td>
</tr>
<tr>
<td>calcium content of the water (mg/kg)</td>
<td>Ca</td>
<td>[0,30),[30,40),[40,9999)</td>
</tr>
<tr>
<td>mean water depth of the pit (metres)</td>
<td>depth</td>
<td>[0,2.5),[2.5,9999)</td>
</tr>
<tr>
<td>age since the site was last used for gravel extraction (years)</td>
<td>age</td>
<td>[0,5),[5,10),[10,20),[20,30),[30,9999)</td>
</tr>
<tr>
<td>area of the pit (acres)</td>
<td>area</td>
<td>[0,10),[10,30),[30,9999)</td>
</tr>
<tr>
<td>number of aquatic plant species in the pit</td>
<td>aq</td>
<td>[0,10),[10,20),[20,9999)</td>
</tr>
<tr>
<td>number of riparian plant species surrounding the pit</td>
<td>rip</td>
<td>[0,20),[20,9999),</td>
</tr>
<tr>
<td>water-based disturbance (increasing ranked categories)</td>
<td>water</td>
<td>0,1,2,3,4,MV</td>
</tr>
<tr>
<td>bank-based disturbance (increasing ranked categories)</td>
<td>bank</td>
<td>0,1,2,3,4,5,MV</td>
</tr>
<tr>
<td>number of islands</td>
<td>is</td>
<td>0,1,2,MV</td>
</tr>
</tbody>
</table>
Results

GARP developed rules for the presence or absence of all seven outcome variables. Listed on Table 2 are the rules for total adult birds and Tufted Ducks. Each rule has a significance greater than z=1.96. The accuracy of each rule (p), number of data selected by the rule and total accuracy of the rule set is also shown. The total and reduced linear regression equations are shown in Table 3, where the $r^2$ value and the predictive accuracy are also reported. The predictive accuracy is a measure of goodness of fit. The linear regression model fits the total adult data better than rule sets (93.3% versus 85.5%), while there is no significant difference between the rule set and the linear regression model for the Tufted Duck (86.0±0.6% versus 85.0%). The results of prediction performed on the additional five outcome variables showed no significant difference in predictive accuracy between the two methods.

The rules marked with an asterisk in Table 2 are both significant and useful for prediction. A rule becomes useful for prediction when the accuracy (p value) exceeds the value of the largest proportion of presences or absences. This is because in the absence of any other information, the best strategy for predictive accuracy is to choose the most likely class (Manuscript 6). The expected accuracy of this strategy for prediction will be the frequency of the most likely class. The accuracy of a rule must exceed this value to be of use for prediction. Thus a rule can be significant in raising the probability of an outcome, but may not be of use, (or detrimental) in prediction. When a rule is significant, the probability of the outcome is more likely given the occurrence of the precondition. When the probability of an outcome is low, as is the case for the absence of total adults (p=0.15), a number of rules can raise this low probability (e.g. $0<\text{pH}=[8.3,14]$ p=0.31) but not be sufficiently accurate to exceed the most frequent category, in this case totad=1 with p=0.85. Thus for applying rule sets to prediction, GARP uses the combination of significance and predictive accuracy of rules to determine which rules to apply (Manuscript 10).
Table 2 (a) Rule set generated by GARP for predicting total adult birds and (b) the rule set generated by GARP for Tufted Ducks. All rules have a significance greater than 95% C.L. The rules marked with an asterisk are also of use for prediction, as they have a p value (predictive accuracy) greater than the maximum of the proportion of presences (1) or absences (0).

<table>
<thead>
<tr>
<th>Class</th>
<th>Rule</th>
<th>p</th>
<th>Selected</th>
<th>Total % Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total adults</td>
<td>0&lt;--</td>
<td>0.15</td>
<td>8</td>
<td>85.5±1.7</td>
</tr>
<tr>
<td></td>
<td>0&lt;--rip=[0,20], is=0</td>
<td>0.27</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0&lt;--Ca=[30,40], is=0</td>
<td>0.31</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0&lt;--pH=[8.3,14]</td>
<td>0.31</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0&lt;--depth=[0,2.5], rip=[0,20]</td>
<td>0.33</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0&lt;--Ca=[30,40], rip=[0,20]</td>
<td>0.36</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0&lt;--depth=[0,2.5], bank=1</td>
<td>0.5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1&lt;--</td>
<td>86.7</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>*1&lt;--depth=[2.5,9999]</td>
<td>1.00</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>*1&lt;--c=[300,9999]</td>
<td>1.00</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Average length of rules</td>
<td></td>
<td>1.3±0.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Tufted Ducks   | 0<--                                           | 0.65  | 39       | 86.0±0.6     |
|                | *0<--is=0                                      | 0.82  | 40       |              |
|                | 1<--                                           | 0.35  | 21       |              |
|                | *1<--Ca=[40,9999], aq=[10,20], rip=[0,20]     | 0.64  | 14       |              |
|                | *1<--rip=[20,9999], water=0                   | 0.70  | 10       |              |
|                | *1<--c=[300,9999], ph=[7.8,8.0], Ca=[40,9999], aq=[10,20] | 0.75 | 8 | |
|                | *1<--is=1                                      | 0.83  | 40       |              |
|                | *1<--bank=0                                    | 1.00  | 8        |              |
| Average length of rules |                                | 1.5±0.6 |          |              |

Simple structural measures were recorded: the number of significant variables in the complete and reduced model, the average length (number of preconditions) in the rule sets, and the number of rules found. As the forms of models differ greatly, structural comparisons are indicative rather than exact comparisons. The average number of preconditions in the rules was 1.3 and 1.5, while the number of variables in each of the linear regression models was 3 and 4. The rules sets, however, contain 8 and 6 different variables in the rules.

There was no consistent difference in the goodness of fit of the two methods. The two methods produced models with very different structural characteristics however. The rule sets produce a large number of simple, partial models of the data. The linear regression model uses a number of variables in a single, total model. The methods are empirically indistinguishable but propose very different underlying models of the system.

We now assess the answers provided by the models to the three wildlife management questions.
Table 3. Regression and reduced regression equations applied to the total adult and Tufted Duck data.

<table>
<thead>
<tr>
<th></th>
<th>Regression equation</th>
<th>$r^2$</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total adults</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>$-2.15 + 0.00c + 0.25ph + 0.00Ca + 0.09depth - 0.00age + 0.01area + 0.19aq - 0.00rip - 0.16water + 0.08bank + 0.09is$</td>
<td>0.44</td>
<td>93.3</td>
</tr>
<tr>
<td>reduced</td>
<td>$0.67 + 0.05depth + 0.01area - 0.12water$</td>
<td></td>
<td>90.0</td>
</tr>
<tr>
<td><strong>Tufted ducks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>$3.07 - 0.01c - 3.12ph + 0.03Ca - 0.04depth - 0.01age - 0.02area - 0.13aq + 0.04rip - 0.01water + 0.02bank + 0.35is$</td>
<td>0.33</td>
<td>85.0</td>
</tr>
<tr>
<td>reduced</td>
<td>$0.41 - 0.00c + 0.02Ca - 0.01area + 0.37is$</td>
<td></td>
<td>81.7</td>
</tr>
</tbody>
</table>

What factors predict the occurrence of a certain species?

The linear regression equations show that the presence of total adults increases with increasing depth. Rules regarding depth also were produced by GARP. Pits having depth greater than 2.5 metres have a significantly higher occurrence of adult birds. This rule is also highly accurate, depth greater than 2.5 metres predicts the presence of birds with an accuracy of 100%. Other less accurate rules state that pits of less than 2.5 metres deep associated with low riparian plants or bank disturbance have significantly more absences of birds. The water-based disturbance (water) and area of pits (area) variables were found to be significant in the linear regression, but were not present in any rules. Variables not identified as significant in the linear regression produced significant rules: the presence of islands and numbers of riparian plants. Using rules, deep pits (p=1.00) or pits with high conductivity (p=1.00) are sufficient to predict the presence of adult birds. No rules are accurate enough to predict absence of birds.

The presence and absence of islands was found to produce significant rules for to predicting the presence and absence of Tufted Ducks. Islands (is) was also a significant variable in the linear regression. Area (area) is significant in the linear regression, although it does not appear in any rules. A number of other highly specialised rules predict for the presence of the Tufted Duck while few predict the absence. To predict the presence of Tufted Ducks, the presence of islands (p=0.83) or low bank disturbance (p=1.00) is sufficient. Absence of islands predicts absence of Tufted Duck with high probability (p=0.83)

If we examine a graph of the variable, 'depth', plotted against the numbers of total adults, there is difficulty in discerning any relationships, either linear of rule-based (Figure 1). This is typical of this data set, where linear relationships are very weak and the significance values of the regression model poor. Nevertheless, accurate rules can be found such as the relationship of depth to presence of total adults.
Using rule sets to explain

The presence of Tufted Ducks is significantly related to low bank disturbance and the presence of islands. However only the absence of islands has is predicted with high accuracy (p=0.85). Thus the management directive for this information would be; absence of islands are necessary for the absence of Tufted Ducks.
Table 4. The rules developed by GARP when environmental variables are predicted from the response variable. The rules shown in bold are also present in the previous tables, where the outcome was predicted from environmental variables.

<table>
<thead>
<tr>
<th>Class</th>
<th>Rule</th>
<th>p</th>
<th>Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Adults</td>
<td>age=[30,9999]&lt;-totad=0</td>
<td>0.25</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>aq=[0,10]&lt;-totad=0</td>
<td>0.25</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>water=4&lt;-totad=0</td>
<td>0.38</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>ph=[8.3,14]&lt;-totad=0</td>
<td>0.63</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>depth=[0,2.5]&lt;-totad=0</td>
<td>1.00</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>is=0&lt;-totad=0</td>
<td>1.00</td>
<td>8</td>
</tr>
<tr>
<td>Tufted Ducks</td>
<td>bank=0&lt;-TD=[1,9999]</td>
<td>0.38</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>is=1&lt;-TD=[1,9999]</td>
<td>0.48</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>is=0&lt;-TD=0</td>
<td>0.85</td>
<td>39</td>
</tr>
</tbody>
</table>

What factors control the presence and absence of water-birds?

The rules found in reverse in the rule sets predicting the response of birds (Table 2) are shown in bold in Table 4. A number of variables appear in the reverse set (i.e. aq and age) that did not appear in the rule set for predicting birds from environmental variables (Table 2). One rule appears in the forward directions in the total adult rule sets, while all rules are found in the forward direction in the Tufted Duck rule set. This illustrates an essential asymmetry in the relationship of animals to the environment, not expressed by curve fitting methods.

The rules for predicting environmental factors from the presence of Tufted Ducks involve the same factors as the rules for predicting the Tufted Ducks from the environment. The rules suggest that the presence of islands is both necessary and sufficient for the presence of the Tufted Duck. The absence of islands is also necessary and sufficient for the absence of the species. Lack of bank-based disturbances are important for presence. These variables control the occurrence of the animal in the sense that there is an 'if and only if' relationship between the birds and the factors. The important management statement would be that islands control the occurrence of Tufted Ducks.

Discussion

Regression methods provide information on the general trends in the data. Rules sets provide more specific information, and general trends in cases were a number of rules are formed (i.e. Tufted Ducks). It appears that rules while providing a useful predictive and explanatory tool also find particular factors that predict the occurrence of species. Regression modelling is a modelling system with a bias or a preference for general patterns to explain events, while the rule sets have a bias for particular states of variables as explanations (Manuscript 7). Each represent different aspects of
USING RULE SETS TO EXPLAIN

relationships in the data set. The different methods provide different explanations. Some variables in the linear regression model are not found in the rule-sets, and vice versa.

The rules allow a unified answer to three types of question that are at present handled by completely different multi-variate techniques. Rules that have statistical relevance in the forward condition suggest sufficient conditions for the occurrence of an animal, and may be useful for predicting the occurrence of the animal. The rules with statistical relevance in the backwards direction suggest necessary conditions, and may be useful for prediction of the occurrence of environmental factors that are important to the animal. Bi-direction rules characterise of the factors for controlling the distribution of the animal.

Interpretation of rules as biotic response theory.

A major factor in providing explanation through the linear response equation is the reduction of the model to include only a few variables. However the results of the rule set analysis contradict the view that the number of variables can be reduced significantly. On average, five different variables are implicated in determining factors in each of the variables. One explanation for the large number of variables in the model may be correlation between the variables. Analysis of co-variance in Table 5 reveals that most of the variables are not strongly correlated. The rules sets frequently find specific combinations of variables that are useful as predictors and are not mere specialisations of more general rules. This suggests that the response of animals can be due to a complex number of factors.

Table 5. Matrix of Pearson correlation coefficients

<table>
<thead>
<tr>
<th></th>
<th>c</th>
<th>pH</th>
<th>Ca</th>
<th>depth</th>
<th>age</th>
<th>area</th>
<th>aqu</th>
<th>rip</th>
<th>water</th>
<th>bank</th>
<th>is</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>-0.541</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ca</td>
<td>0.948</td>
<td>-0.555</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>depth</td>
<td>-0.194</td>
<td>-0.251</td>
<td>-0.194</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>0.196</td>
<td>-0.156</td>
<td>0.260</td>
<td>-0.222</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>area</td>
<td>-0.063</td>
<td>-0.056</td>
<td>-0.140</td>
<td>0.094</td>
<td>-0.222</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aqu</td>
<td>0.047</td>
<td>-0.249</td>
<td>0.042</td>
<td>0.152</td>
<td>0.147</td>
<td>0.093</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rip</td>
<td>0.113</td>
<td>-0.279</td>
<td>0.101</td>
<td>0.063</td>
<td>0.215</td>
<td>0.056</td>
<td>0.855</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>water</td>
<td>-0.026</td>
<td>0.115</td>
<td>-0.141</td>
<td>-0.085</td>
<td>0.103</td>
<td>0.325</td>
<td>-0.200</td>
<td>-0.289</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bank</td>
<td>0.111</td>
<td>0.151</td>
<td>0.037</td>
<td>-0.139</td>
<td>-0.017</td>
<td>-0.012</td>
<td>-0.132</td>
<td>-0.336</td>
<td>0.634</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>is</td>
<td>0.077</td>
<td>-0.226</td>
<td>0.033</td>
<td>0.062</td>
<td>-0.003</td>
<td>0.460</td>
<td>0.202</td>
<td>0.144</td>
<td>0.089</td>
<td>-0.174</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Rules can be interpreted in relation to continuum theory as set out by Austin and Smith (1989). In simple continuum theory with no interactions between organisms, the fundamental response of species is a bell-shaped curve to a number of environmental gradients. The bell-shaped curve is caused by a strong response by the organism in some part of the range. An ecological interpretation of rules might be
that they represent a selection by the individuals of the population. The animal has the potential to 'choose' between sites and does so by immigration, emigration, recruitment or mortality. Similarly a plant could 'choose' sites by successful recruitment, and unsuccessful recruitment in other areas. The fundamental response of an animal when represented by a rule is a step function, which can be seen as a qualitative version of a bell-shaped curve (Figure 2).

![Figure 2. Comparison of a bell-shaped response to an environmental resource gradient and a rule. The rule is a step function, coinciding with the apex of the bell-shaped curve.](image)

In many cases the environmental response variable may not be measured over the whole range of response. Three distinct sections could be formed, depending on the part of the response function. The first section is exponentially increasing, the second is a humped response and the third exponentially decreasing (Figure 3). A rule can represent this range of responses (Figure 3a). A linear regression model can be used to represent the increasing and decreasing side of the response, but in the centre section the regression line would be flat, and therefore not significant. Thus the rule set approach can detect significant patterns in the data not detected with linear regression model development as used in this paper. Linear regression would only be valid if the environmental gradient was on one or other side of the curve. If a bell shaped distribution was suspected then different methods of analysis including transformations of the data would normally be used.
Rules can detect partial relationships, a form not detected by curve fitting. A partial response occurs when the value of a variable is significant for part of the range of the function. In the other part of the range, the points are distributed randomly (Figure 4). The rule can model the part that provides a significant response, and ignore the rest of the range. A linear regression function may suggest a significant correlation with the data but would give an erroneous indication of accuracy over part of the range. The problem is due to the assumption of linear modelling that the response extends throughout the whole of the range of the variable. Partial data as illustrated in Figure 4 may often be found in biological studies due to thresholds of survival, environmental conditions beyond which survival cannot occur.
Which is better?

Explanation has three components: empirical adequacy, the proposed underlying model of the system, and the way in which that model contributes to our understanding. We can compare the two methods on this basis. There was no difference in the empirical adequacy. The models as explanations suggested many of the same significant variables, but also many variables differed. Thus the models offer different explanations for the biotic response of the water birds to the environment. Each of the methods is consistent with a simple theory of biotic response, but represent different aspects of it. For example, regression promotes generality, while rule sets promote specificity.

Perhaps the central feature of a good method is few assumptions, or prior knowledge, required to analyse the data. Fewer assumptions allow the development of systems with greater autonomy and generality as making few assumptions about the domain reduces the risk of assumptions not being met in a particular domain. The major assumption of the statistical methods is the global covering of the domain, and the form of the function, e.g. a linear or bell shaped response, over the whole of the range of the environmental variable. Rule sets, however, make local distributional assumptions, i.e. the values within a limited range of the independent variables are constant. They are therefore most applicable to categorical data and categorised continuous data, a natural complement to statistical methods for regression that use continuous independent variables. The present system requires preprocessing of continuous data into categories. While this presents an opportunity to include prior information, it is also time-consuming and intuitive. Classification techniques could be adapted to perform this task automatically, or rules could be defined as fuzzy sets on continuous variables.

Acknowledgments

Tony Fox of the Wildfowl and Wetlands Trust kindly provided the data. Stuart Davey helped in the conceptual development of this paper.
USING RULE SETS TO EXPLAIN

...
Section 5. Conclusions

The conclusions are elaborated in the final manuscript (Manuscript 12). Machine learning provides a methodology for model development that can be analysed through a formal learning theory approach. Theoretical problems are encountered in every aspect of the search for the underlying model, leading to sub-optimal prediction, explanation or both. The difficulties identified are:

- search spaces too large to search satisfactorily by hand,
- the space of potential hypotheses does not contain or only partially contains the underlying model,
- insufficiently strong measures of model adequacy that favour predictive models or explanatory models,
- ill-posed problems leading to instability in the induced model, and
- model development protocols that do not relate to the actual application of the model.

Machine learning presents an opportunity to ameliorate these problems with greater computational resources and more flexible model structures, or representations. A range of theoretical and empirical evidence supports the use of induction of sets of rules by a GARP-like algorithm as predictive and explanatory methods of developing models of animal response to habitat.
Conclusions

How can ecological modellers provide predictive and explanatory models? This is the prediction/explanation problem. My thesis is a novel experimental approach to the problem, using formal learning theory and machine learning methods. One major outcome was a greater understanding of prediction and explanation problems in ecology. Another was a machine learning system which has shown promising results in the automated development of predictive and explanatory models of the distribution of wildlife.

The problem

The widely reported problem of prediction and explanation in ecology is characterised by models which either predict or explain but not both. For example, regression equations developed through fitting data to functions are widely used for modelling of trends, e.g. the response of canopy height to site fertility. These are predictive models, but not explanatory, as they do not characterise important causal processes within the environment. The functions are parametrised from local environmental conditions rather than fundamental physiological growth constants of trees.

Theoretical models are mostly non-predictive explanatory models. The logistic growth equation and models of competition explain through commonsense notions of the way animals and plants should behave under ideal conditions. As practical guides to management of natural resources these models have often been found wanting. One of the most well known examples is the Ricker equation - the main theory of stock-recruitment dynamics. It is used in the prediction of optimal harvesting rates in fisheries, yet as Hall (1988) claimed, it is unsupported by data. The inability of the Ricker equation to predict feasible harvesting rates for maintaining fish populations may have contributed to the failure of many fisheries throughout the world. Thus there are serious implications for failure of models to predict and explain in ecology as predictive and explanatory models are essential for maintenance of natural resources. The application used in this study - the modelling of the biotic response of animals to the environment - is essential to the prediction of wildlife habitat areas and planning for their long-term management.

The approach

What is the prediction/explanation problem, and how can we address it? There are a number of possible reasons for difficulties of prediction and explanation in modelling. Some reasons, such as cultural and technological factors, are outside the domain of a theoretical and experimental study, e.g. lack of data or research effort, and lack of availability of mathematical tools or expertise. The approach to studying
the problem was to develop a computer system which can produce models that both predict and explain. This was used to address three substantive arguments for the failure of models in ecology (Manuscript 2):

- There are fundamental theoretical limitations to developing models that both predict and explain.

The existence of a system that developed predictive and explanatory models in a reliable way would rebut this argument.

- That ecological phenomena are difficult to model because of lack of knowledge of laws and processes.

The development of a system that develops predictive and explanatory models without a great deal of knowledge of laws and processes would constitute a rebuttal of this argument.

- That models fail because they are not complex enough.

The development of a system that produces simple models that are predictive and explanatory would rebut this argument.

Such a system, called GARP, was developed.

In searching for a predictive and explanatory system, I have examined three model development systems: decision trees, Bayesian inferencing and rule set induction. The first method, decision trees, is being increasingly used as a predictive technique in ecology. The second, Bayesian inferencing, though not widely used in ecology, has been used in various forms in diagnostic systems throughout the medical field. The third, called GARP, genetic algorithm for rule set production, extends existing rule induction systems.

**Definition of prediction and explanation**

The meaning of prediction and explanation was explored in section three of this thesis. Machine learning research often uses three explicit levels of structure: data, models and biases. Rendell (1986) characterised machine learning as a simultaneous optimisation problem at each of the three levels (Figure 1). In this thesis these levels are associated with three distinct tasks of models: prediction (Manuscript 6), explanation (Manuscript 7) and understanding (Manuscript 8). Modelling in ecology can be mapped onto this characterisation.
Prediction

When one predicts one asserts information about which little is known. When using a model to predict, these assertions are made on the basis of some other information. Thus in prediction there is the information used to develop a model, called the training set, and the information used to test the model, called the test set. A number of relationships between the test set and the training set correspond to a distinct form of prediction, called a model development protocol. (Manuscript 6):

- **Goodness of fit** is a degenerate case of prediction where the training set is the same as the test set.

- In **resampling**, the training set and test sets are randomly selected subsets of a larger population, simulating replicated experiments.

- In **forecasting**, the training set was formed at an earlier time to the test set.

Each of these distinct relationships between the test set and the training set correspond to a distinct form of prediction, called a model development protocol.

Explanation

Explanation is the explication, or making known in detail, of the underlying model, or 'deep structure' of reality. On this view, the model itself is an explanation for observations as expressed by data. The problem of explaining a set of data is to
find the symbolic structure that generated a given set of data, called \textit{structure recovery} in machine learning. While a structural approach is fundamental to modelling, it also carries with it certain assumptions, most importantly that reality can be expressed as symbols, and that the underlying structure is stable. Symbolic structures are necessary for explication of reality and communication of knowledge. The requirements for symbol formation and stability provide useful criteria for selecting explanatory models.

\textbf{Understanding}

The third aspect of modelling is the sense in which a model conveys an understanding of the world. The main forms identified in this thesis are

- \textit{causation}, partial relationships between states of the world,
- \textit{general laws}, a complete description of events, described by a single function, and
- \textit{mechanisms}, where events are described as combinations of complete descriptions of parts of the world.

These forms of explanation express a preference for particular types of models known in machine learning as \textit{bias}. One form of bias is background knowledge used to determine the form of models in a given problem. Often the form is decided \textit{a priori} using the background knowledge, while data is only used to parametise the model. Assumptions are a form of \textit{a priori} bias, as is parsimony, to prefer the most simple model. Perhaps the most characteristic feature of good machine learning programs is that they contain little background knowledge, i.e. the possible models and biases are left open. Thus machine learning is particularly suited to modelling in a knowledge poor environment, such as ecology.

\textbf{Formal learning paradigm}

The second major concept from machine learning with application to model development is the question of justification: how we can say that a system develops predictive and explanatory models? For this we call on the \textit{formal learning paradigm} as developed by Kelly and Glymour (1990). This paradigm has a number of components (Figure 2),

- the \textit{possible worlds}, or sets of data that could be presented,
- a \textit{data protocol} or way of presenting data, and
- a \textit{theorist} such as a machine learning program, that asserts a model for a given presentation of evidence.
Given these components there are two other concerns,

- **adequacy**, whether the model satisfies given conditions, and
- **convergence**, whether the system outputs a fixed model eventually, given certain changing conditions.

![Diagram of the formal learning paradigm](image)

Figure 2. Diagrammatic illustration of the formal learning paradigm, the language used to express the data and the models,

Adequacy is the determination of whether the model predicts and explains sufficiently well. Prediction in categorical data is determined from the accuracy, the number of data items correctly predicted divided by the total number expressed as a percentage, while prediction in continuous data may be determined by goodness of fit. Prediction can be seen as a measure of the distance of the models from 100% accuracy. Similarly a distance can be defined from the model developed by the theorist to the true underlying model of the system.

Convergence means that given a data protocol we can be sure that after presentation of some finite number of data the theorist asserts a model which does not change given any additional data. In non-convergent model development successive data continue to change the model, an unacceptable situation akin to a scientist continually changing his or her mind about a scientific theory.

Although the formal learning paradigm, as formulated by Kelly and Glymour (1990), is concerned with convergence under incremental presentation of data, three forms of convergence were identified: convergence in data, convergence in computation, and convergence in the number of variables. To have converged in data means that the same model is produced even though more data is provided to the system. Convergence in computation means that given a fixed amount of data there is a time when the system asserts an unchanging model, e.g. even though GARP continually tests new rules at some time it stabilises on a set of best rules.

197
Convergence in the number of variables is the stabilisation of a model as the number of variables in the data is increased. An example of non-convergence under variables is shown in an application of LBS (Manuscript 6).

Finally we need to know if a modelling system is reliable. A modelling system is reliable according to Kelly and Glymour's formal learning paradigm if under a model development protocol, i.e. a succession of model developments under different data presentations, it is guaranteed to converge on the underlying model of the system.

Formalisations of learning are quite useful in the theoretical analysis of types of models where we can examine convergence to a known limit, but are not as applicable to experimental model development. In theoretical analysis of algorithms on generated data sets, the question of whether the system has converged to a known limit is settled by comparison of the output of the theorist with the known underlying model. In the real world we do not know the underlying model of reality, so we do not know what limit the model should converge to. Throughout this thesis, convergence to the limit is examined only in a generated data set where the underlying model is defined. In analysis of natural data sets, only convergence to a limit is required. This form of convergence, similar to Cauchy convergence in mathematical analysis, is demonstrated by 'stability' of successive models in a model development protocol (Figure 3). Thus reliable explanation means that there is a high probability of producing almost similar predictive models in distinct model developments in the model development protocol.

![Diagrammatic illustration of the model development protocol.](image)

**Figure 3.** Diagrammatic illustration of the model development protocol.

**Empirical results**

Three systems were used in the thesis. The first was the decision tree induction algorithms, ID3, CART and CN2. Decision trees have been widely used for
prediction. They classify objects into categories on the basis of their attributes. It was noted that under a resampling model development protocol, a number of different trees were produced by the CART algorithm (Manuscript 4). Under this protocol, successive models did not converge. Thus we can say that explanation has not occurred even though prediction under CART is as good as the other algorithms. While decision trees have been regarded as useful predictive tools, they have also been regarded as poor forms of representation. The explanation provided in this thesis for poor representation is lack of consistency with any of the main forms of scientific model: causal, general law or mechanistic. Because of lack of consistency with a general scientific model, they do not contribute to understanding.

The second system, LBS (Learning Base System), predicts using a matrix of conditional probabilities of attributes given outcomes, such as the probability of a symptom given a disease (Manuscript 5). It is a rather simple learning system. The conditional probabilities are simply recalculated given each new datum presented. Learning Base System was developed with a novel intermediate form of organisation to facilitate user interaction and explanation of the operation of the program. It allows rapid, automated model development and provides a very flexible and computationally efficient platform to perform experiments in model development. This system is predictive in the sense that it predicts as well as the decision trees under the resampling protocol. A problem with prediction occurs when two outcomes occur simultaneously, e.g. two diseases are present. The basic system fails as it would fail to predict either disease, although an adaptation of the system overcomes this problem at the expense of reduced predictive power. The system fails to converge in data because data leading to failure to predict could be presented at any time.

The third system to be developed was GARP, an acronym for Genetic Algorithm for Rule-set Production (Manuscript 9). A rule is an if-then statement such as, "if site quality is high then canopy height is high". This system was used for experimenting with prediction and explanation (Manuscript 10). The system invariably was successful at finding the underlying model of data sets generated using a simple set of rules with noise, although additional rules were also found. On both natural and generated data sets, the predictive accuracy of the system stabilised rapidly with computation time; however as computation time increased, the number of rules in the current rule set increased almost linearly. Thus the system didn't satisfy the Cauchy condition for convergence to an explanation. Two different forms of bias were then applied: precondition length and generalisation. When the possible rules were limited to single preconditions (e.g. if A=a1 then C), convergence occurred in both the artificial and natural data set. On the natural data set however, restriction to a single precondition decreased predictive accuracy. Under generalisation bias the number of rules increased at a very slow rate, but the predictive accuracy remained high. These results show that different types of bias produce different results on different data sets.
Another manipulation of bias was the use of different measures for estimating the utility of rules. Combinations of a significance measure and a probability measure were compared for predictive accuracy. It was found that a combination of the two produced the highest prediction on a resampling model development protocol. The significance measure is biased towards rules that are more likely to be found in succeeding model developments, encouraging stability of the model. The probability measure is biased towards rules that will result in the highest overall predictive accuracy of the rule set, encouraging accuracy of the model. This result shows that a conjunction of explanatory and predictive measures achieves maximum predictive accuracy under resampling.

Another effect of bias was comparison of a linear regression form of modelling with a rule set induction (Manuscript 11). The linear regression model was chosen as a simple modelling system based on general functional relationships. Linear regression and the rules derived from GARP did not differ consistently in overall predictive accuracy. While there was a deal of overlap in the variables found significant and used by GARP, some variables were unique to each method. Thus support can be gained for completely different forms of model.

The major experimental results of the thesis are illustrated in Table 1. The predictive accuracy of the methods used in the thesis is extracted from the various manuscripts (4, 6, and 10). Comparing the predictive accuracy by resubstitution and resampling on the natural data set, the machine learning methods outperformed the more conventional methods on resubstitution accuracy. Comparisons of resampling performance were not performed as the non-machine learning methods do not have a model development protocol that allows development and testing on different data sets. ID3 and LBS produced the highest resampling accuracy, significantly greater than CART and GARP. The decision tree algorithms however had a higher variance than LBS and GARP, illustrating the greater instability of decision trees.

Table 1. Comparison of a range of predictive methods on the Waratah Creek greater glider data set. The three non-machine learning based models are: MR - multiple regression, PCA - a model based on principle components analysis, and KA - a rule based expert system developed by an expert. The machine learning methods are ID3, CN2 and CART - decision tree induction algorithms, LBS - Learning Base System, and GARP, a rule set induction system.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy resubstitution</th>
<th>Accuracy resampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>45%</td>
<td>-</td>
</tr>
<tr>
<td>PCA</td>
<td>41%</td>
<td>-</td>
</tr>
<tr>
<td>KA</td>
<td>48%</td>
<td>-</td>
</tr>
<tr>
<td>ID3</td>
<td>60.7±1.3%</td>
<td>57.3±2.2%</td>
</tr>
<tr>
<td>CN2</td>
<td>50.8±2.6%</td>
<td>45.2±2.2%</td>
</tr>
<tr>
<td>CART</td>
<td>61.2±3.5%</td>
<td>54.8±3.9%</td>
</tr>
<tr>
<td>LBS</td>
<td>58.2±0.5%</td>
<td>56.4±0.5%</td>
</tr>
<tr>
<td>GARP</td>
<td>60.8±1.3%</td>
<td>53.4±0.8%</td>
</tr>
</tbody>
</table>
A preliminary implementation of GARP by David Peters of Parks and Wildlife in Tasmania is proving to be very successful in developing predictive and explanatory models for mapping the distribution of a range of species of wildlife in Tasmania from topographical and climatic data (Peters, pers. comm.). The wildlife experts examining the maps expressed satisfaction with the predicted distributions. The ultimate test of the system will be if surveys of areas where occurrences are predicted but no sightings have been recorded result in discoveries of new populations of animals.

The prediction/explanation problem

Automated model development can be defined as: reliable development of models using a model development protocol such as resampling, where a convergent search of a set of possible models, using a protocol for presentation of data, produces stable and predictive models.

There are a number obstacles to achieving this task. These are the problems of prediction and explanation, and appear in ecological modelling and within the systems used in this thesis. Although the problems have a basis in fundamental problems of acquiring knowledge about the world, the results of this research have shown that machine learning, and the formal paradigmatic approach, can ameliorate many of the difficulties.

Problems of search

We need to search because we do not know the underlying model, and there is no explicit calculation for finding it. Traditional modelling has shied away from search problems despite their prevalence in modelling, eg. we cannot reliably find a reduced linear regression equation without searching the alternative linear models. In practise there are more models than can be searched manually, leading to sub-optimal solutions, and poor prediction and explanation. Hypothesis testing can be seen as a degenerate form of search, where the model space consists of a single model.

Machine learning methods such as GARP test a large number of hypotheses, one for each rule. Thus machine learning methods expand the number of hypotheses that can be tested in a scientist's lifetime, and increase the chance of finding the actual underlying model of the system.
Conclusion: Finding the underlying model requires search of possible models and biases, not just parametisation of equations defined a priori.

Set of possible models

One of the first choices in modelling is the set of possible models containing the underlying model of the system. However, the underlying model may not be included in the set of models chosen. Typically, and particularly in a system with little background knowledge such as ecology, we do not know the general form of the underlying model. Thus the general form chosen for the underlying model may be wrong or only partly correct, leading to poor prediction and explanation.

In a survey of preferred forms of explanation in ecology, causal forms expressing local interactions are preferred over general laws expressing global interactions. Although there may be other reasons for using general laws or mechanistic models, the tendency of people to prefer causal explanations suggests that causal relationships as expressed by rules may be a real property of ecological systems. The large numbers of specific rules found to be significant in natural data sets support the view that specific rules, as well as general relationships or trends as expressed by linear and nonlinear functions, comprise the underlying models of ecology.

Conclusion: Models based on global fits to data using complete functions only express general trends in ecological data; local or partial relationships expressed by rules are also necessary.

Problems of data

Measures indicating the utility of possible models are imperfect because data may be incomplete, wrong or noisy. For example, a high probability of A given B does not entail that A is relevant to B. Significance measures reflect statistical relevance but may not be useful for prediction when the prior probability of the event is low. Using imperfect measures as indicators of the underlying model can be responsible for the development of models that are either predictive or explanatory, depending on the measure used.

The GARP system used a conjunction of measures. The best prediction occurred with a combination of statistical relevance and posterior probability measures. This result shows that predictive and explanatory models can be the intersection of predictive and explanatory models (Figure 4). Other measures may be necessary to obtain models which restrict the rules only to the underlying model. At present it is clear that using statistical relevance and posterior probability, GARP finds the underlying rules, and some additional rules.
**Conclusion:** A conjunction of two measures of utility of models are necessary for achieving the combined goals of prediction and explanation, and additional criteria may be necessary for discovering the underlying model.

**Possible models**

![Venn diagram of the relationships between models.](image-url)

Figure 4. Venn diagram of the relationships between models.

**Problems of stability**

Some forms of model are sensitive to the data protocol. This is called instability and was demonstrated in the CART algorithm by the induction of a number of different decision trees under resampling. This problem can be seen as a variety of ill-posed problems or inverse problems, in contrast to forward or well-posed problems where the solution is a specific well-defined outcome. Instability is invariably due to an ill-defined model structure.

Decision trees can be unstable because the models are developed incrementally, i.e. the later branches in the tree are determined by former choice of branch conditions. Where there are no clear initial choices of branch condition, random variations in the training data set can produce choices of variables for branching leading to large differences in the final tree. This can happen when many of the variables are correlated as is often the case with ecological data sets. In contrast, in rule sets, the choice of a rule as part of a model does not directly influence the choice of any other rule. Thus rule sets are less sensitive to random variations in the training data and are more stable.
Conclusion: Instability due to interactions within components of the model can be reduced by using forms of representation composed of non-interacting components, such as sets of rules.

Problems of prediction

Prediction involves more than fitting models to data. A good fit of a model to data explains the given data, but may not necessarily predict on an independent set of data. The main reason for this is overfitting, i.e. a component of the model is expressing noise and not the underlying model. Failure to appreciate this methodological error can lead to explanation without prediction.

There is a sense in which the predicted data set is a different set to the one on which the model was developed. Although it would be better to use the actual data for developing the model, this cannot be done where the data from the situation where predictions are needed is unobtainable, e.g. in the future. Although there is no solution to this problem, we can use a modelling protocol to give us an unbiased estimate of the predicability of the system with given variability in the environmental situations. An example of this is the use of resampling to simulate the application of a model to a future year (Manuscript 6). While resampling methods are computationally intensive methods of developing and testing models, they are a useful application of increased computational power.

Conclusion: The model development protocol should simulate multiple developments and applications of the model under conditions approximating the variation environmental situations to give a measure of the reliability of the model development system.

These conclusions suggest that there are a number of fundamental difficulties in prediction and explanation, some of which contribute to reduction in both prediction and explanation, and some that reduce one or the other. However no fundamental limitations to prediction and explanation have been found given the definitions of prediction and explanation used in this study. Machine learning and the formal learning paradigm assists in finding this underlying model reliably.

Implications for modelling in ecology

The second argument was that ecological phenomena are difficult to model because of lack of knowledge of laws and processes.

The theory of biotic response to the environment is based on a functional response to environmental gradients - in its basic form a bell shaped curve. These are general laws - they apply over the whole range of a variable. I have shown from the survey that causal models, where specific events are related to other specific events,
are regarded as satisfactory forms of explanations, and can be expressed as rules. The use of rules as a method of modelling embodies a theory that biotic response is based on a large number of local effects with potentially high order interactions with the environment, rather than a few general trends with no interactions. Local effects with high order interactions could come about through biological thresholds such as decision-making by organisms about factors such as acceptable nesting sites or food sources, or sharp changes in response to variables. Thus rule sets provide an alternative form of modelling without the use of general laws or mechanistic models for representing processes. It therefore rebuts the second argument.

The third argument was that models fail because they are not complex enough. The results suggest that prediction and explanation can be achieved with simple qualitative models and developed using lots of data. GARP presents a method of analysis appropriate to current and projected database technology integrated with a psychological and theoretical basis for understanding the biotic response to the environment. The promising results for this system suggest that modelling systems do not need to be more complex, in terms of incorporating additional terms into existing models, but different from traditional analytical approaches. The characteristics of this approach are:

- development of models directly from survey data,
- partial models (i.e. local, or covering part of the range of the variable) rather than complete or global,
- search of large numbers of possible models rather than relying on background knowledge of how the system should behave,
- predictive and explanatory criteria to determine adequacy of models,
- models composed of sets of independent models, rather than single models with interacting components,
- and protocols for model development generating multiple models to establish stability.
References


Hall, C.A.S. (1988) An assessment of several of the historically most influential theoretical models used in ecology and of the data provided in their support. Ecological Modelling, 43, 5-31.


Appendix 1.

Publications during doctoral research


Presentation only


In press


*In review*


*In preparation*

Stockwell, D.R.B. A structural view of scientific explanation by explanatory modelling systems in ecology.