Editorial Policy

for the publication of proceedings of conferences
and other multi-author volumes

Lecture Notes aim to report new developments - quickly, informally, and at a high level. The following describes criteria and procedures for multi-author volumes. For convenience we refer throughout to "proceedings" irrespective of whether the papers were presented at a meeting.

The editors of a volume are strongly advised to inform contributors about these points at an early stage.

§1. One (or more) expert participant(s) should act as the scientific editor(s) of the volume. They select the papers which are suitable (cf.§§2-5) for inclusion in the proceedings, and have them individually refered (as for a journal). It should not be assumed that the published proceedings must reflect conference events in their entirety. The series editors will normally not interfere with the editing of a particular proceedings volume - except in fairly obvious cases, or on technical matters, such as described in §§2-5. The names of the scientific editors appear on the cover and title-page of the volume.

§2. The proceedings should be reasonably homogeneous i.e. concerned with a limited and well defined area. Papers that are essentially unrelated to this central topic should be excluded. One or two longer survey articles on recent developments in the field are often very useful additions. A detailed introduction on the subject of the congress is desirable.

§3. The final set of manuscripts should have at least 100 pages and preferably not exceed a total of 400 pages. Keeping the size below this bound should be achieved by stricter selection of articles and NOT by imposing an upper limit on the length of the individual papers.

§4. The contributions should be of a high mathematical standard and of current interest. Research articles should present new material and not duplicate other papers already published or due to be published. They should contain sufficient background and motivation and they should present proofs, or at least outlines of such, in sufficient detail to enable an expert to complete them. Thus summaries and mere announcements of papers appearing elsewhere cannot be included, although more detailed versions of, for instance, a highly technical contribution may well be published elsewhere later. Contributions in numerical mathematics may be acceptable without formal theorems/proofs provided they present new algorithms solving problems (previously unsolved or less well solved) or develop innovative qualitative methods, not yet amenable to a more formal treatment.

Surveys, if included, should cover a sufficiently broad topic, and should normally not just review the author's own recent research. In the case of surveys, exceptionally, proofs of results may not be necessary.

§5. "Mathematical Reviews" and "Zentralblatt für Mathematik" recommend that papers in proceedings volumes carry an explicit statement that they are in final form and that no similar paper has been or is being submitted elsewhere, if these papers are to be considered for a review. Normally, papers that satisfy the criteria of the Lecture Notes in Statistics series also satisfy this requirement, but we strongly recommend that each such paper carries the statement explicitly.

§6. Proceedings should appear soon after the related meeting. The publisher should therefore receive the complete manuscript (preferably in duplicate) including the Introduction and Table of Contents within nine months of the date of the meeting at the latest.

§7. Proposals for proceedings volumes should be sent to one of the editors of the series or to Springer-Verlag New York. They should give sufficient information on the conference, and on the proposed proceedings. In particular, they should include a list of the expected contributions with their prospective length. Abstracts or early versions (drafts) of the contributions are helpful.

Lecture Notes in Statistics 89
Edited by S. Fienberg, J. Gani, K. Krickeberg,
I. Olkin, and N. Wermuth

P. Cheeseman
R. W. Oldford (Eds.)

Selecting Models from Data

Artificial Intelligence and Statistics IV

Springer-Verlag
New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

P. Cheeseman
Mailstop 269-2
NASA
Ames Research Center
Moffett Field, CA 94035
USA

R.W. Oldford
Department of Statistics
and Actuarial Science
University of Waterloo
Waterloo Ontario, N2L 3G1
CANADA

Library of Congress Cataloging-in-Publication Data
Selecting models from data : artificial intelligence and statistics IV/
P. Cheeseman, R.W. Oldford (eds.)
p. cm. - - (Lecture notes in statistics ; 89)
Includes bibliographical references and index.
ISBN-13:978-0-387-94281-0 e-ISBN-13:978-1-4612-2660-4
DOI: 10.1007/978-1-4612-2660-4

1. Artificial intelligence - - Mathematical models. 2. Artificial intelligence - - Statistical methods. 3. Statistics. I. Cheeseman, P. II. Oldford, R. W. III. Series: Lecture notes in statistics (Springer-Verlag) ; v. 89.
Q335.S4119 1994
519.5'4'028563 - - dc20 94-10820

Printed on acid-free paper.

Camera ready copy provided by the author.

9 8 7 6 5 4 3 2 1

Preface

This volume is a selection of papers presented at the Fourth International Workshop on Artificial Intelligence and Statistics held in January 1993. These biennial workshops have succeeded in bringing together researchers from Artificial Intelligence and from Statistics to discuss problems of mutual interest. The exchange has broadened research in both fields and has strongly encouraged interdisciplinary work. The theme of the 1993 AI and Statistics workshop was: "Selecting Models from Data".

The papers in this volume attest to the diversity of approaches to model selection and to the ubiquity of the problem. Both statistics and artificial intelligence have independently developed approaches to model selection and the corresponding algorithms to implement them. But as these papers make clear, there is a high degree of overlap between the different approaches. In particular, there is agreement that the fundamental problem is the avoidence of "overfitting"–i.e., where a model fits the given data very closely, but is a poor predictor for new data; in other words, the model has partly fitted the "noise" in the original data.

The excitement of the AI and statistics workshop, reflected in these papers, comes from the realization that computers will increasingly be required to draw robust inferences from data, sometimes very large quantities of data, despite the presence of incomplete and inaccurate information. And, because the scale of the problems arising from large computer databases quickly overwhelms the human analyst, it is desirable to have a computer assume as much of the role of the analyst as possible. This requires us to have methodological and domain expert information encoded in such a way that it can be used to produce useful results automatically.

Pioneering efforts in this regard have been made in both the Statistics and the AI communities. Statisticians have developed computationally intensive methods for building flexible and interpretable models. AI, particularly machine learning, has pioneered computer based methods of inference from databases with many notible sucesses, as well as a few failures. One solid inference that can be drawn from these workshops is that AI could benefit from making statistical methods the heart of future programs.

It seems to us that there is enormous potential for development at the intersection of statistics, artificial intelligence and computer science. The papers in this book constitute an important step in this direction.

Finally, the production of a volume of this nature requires the dedicated effort of many people. We owe considerable thanks to the programme committee for their effort in the development of the scientific programme, to the Society for AI and Statistics who initiate and finance the workshops, to Nandanee Basdeo who handled the electronic submission of papers and who bore much of the burden for the logistics of this volume and the workshop, to Gwen Sharp who ably handled much of the workshop administration, and to the Department of Statistics and Actuarial Science at the University of Waterloo which was generous with its staff and computational resources in the production of this volume. Finally, we would like to thank the authors in this volume and the workshop participants who have made the workshops a success.

P. Cheeseman
Ames, California

R.W. Oldford
Waterloo, Ontario

January, 1994.

Previous workshop volumes:

Gale, W.A.,ed.(1986) *Artificial Intelligence and Statistics*, Reading, MA: Addison-Wesley.
Hand, D.J., ed. (1990) Artificial Intelligence and Statistics II. *Annals of Mathematics and Artificial Intelligence*, **2**.
Hand, D.J., ed. (1993) *Artificial Intelligence Frontiers in Statistics: AI and Statistics III*, London, UK: Chapman & Hall.

1993 Programme Committee:

General Chair:	R.W. Oldford	U. of Waterloo, Canada
Programme Chair:	P. Cheeseman	NASA (Ames), USA
Members:		
	W. Buntine	NASA (Ames), USA
	M. Deutsch-McLeish	U. of Guelph, Canada
	Wm. Dumouchel	Columbia U, USA
	W.A. Gale	AT&T Bell Labs, USA
	D.J. Hand	Open University, UK
	H. Lenz	Free University, Germany
	D. Lubinsky	AT&T Bell Labs, USA
	E. Neufeld	U. of Saskatchewan, Canada
	J. Pearl	UCLA, USA
	D. Pregibon	AT&T Bell Labs, USA
	P. Shenoy	U. of Kansas, USA
	P. Smythe	JPL, USA
Sponsors:	Society for AI And Statistics	
	International Association for Statistical Computing	

Contents

Part I

Overviews: Model Selection

1
Statistical strategy: step 1

D.J.Hand

Department of Statistics
Faculty of Mathematics
The Open University
Milton Keynes
MK7 6AA

ABSTRACT Before one can select a model one must formulate the research question that the model is being built to address. This formulation – deciding precisely what it is one wants to know – is the first step in using statistical methods. It is the first step in any statistical strategy. This paper contends that often the research question is poorly formulated, so that inappropriate models are built and inappropriate analyses undertaken. To illustrate this three examples are given: explanatory versus pragmatic comparisons in clinical trials, confused definitions of interaction, and two ways to measure relative change. A plea is made that statistics teaching should focus on higher level strategic issues, rather than mathematical manipulation since the latter is now performed by the computer.

1.1 Introduction

The primary theme of this conference is model selection. Many other papers here present and compare strategies and criteria for such selection. This paper, however, takes a step back, to begin with the question 'Why do we want to select a model in the first place?'

The answer, of course, is that we wish to apply the model in addressing some particular research question. For example, some papers at this meeting try to find the best model to answer the question 'What class does this object belong to?' and others address the question 'What are the causal influences on this variable?'

This means that, before one can even consider model selection methods one has to be quite clear what one wants the model for - - or what research question one is attempting to address with the model.

This process of formulating the research question is thus the first step in a 'statistical strategy'.

A statistical strategy is an explicit statement of the steps, decisions, and actions to be made during the course of a statistical analysis. Such explicit statements are, of course, necessary in order to be able to communicate **how** to do an analysis. One would have thought that this was necessary in order to teach statistics. It is certainly necessary if one intends to build a statistical expert system.

Strangely enough, most of the work on statistical strategy seems to have been motivated by this interest in statistical expert systems, with relatively little coming from other areas such as teaching (one exception is Cox and Snell, 1981). Indeed teaching seems to have concentrated on the formal aspects of statistical techniques. (A step in the right direction is the use of project work, though here no principles of strategy are introduced and it is really an attempt to introduce the 'apprentice' role described below into the teaching curriculum.)

---

[1] *Selecting Models from Data: AI and Statistics IV.* Edited by P. Cheeseman and R.W. Oldford.

One reason for this is undoubtedly the mathematical complexity of statistical methods: knowing when a technique should be used is of little value if one is incapable of undertaking the numerical manipulations. Moreover, the mathematical manipulations had to be understood in their entirety in order to apply them. Half a calculation of an estimator is of little value. In contrast, partial understanding of the role and type of questions a technique can address will allow one to apply the methods (often incorrectly, no doubt) so that it might appear that less effort is needed to teach these aspects.

Similarly, the mathematical aspects are easy to formalise (the mathematics itself is a formalisation) while the non-mathematical aspects are difficult to formalise, so that again the former lends itself to teaching. Perhaps we should here acknowledge the fact that statistics has unfortunately often been perceived as being a part of mathematics, so that in the teaching there will be a natural emphasis on the mathematical aspects at the expense of the other, less clearly defined but at least as important, aspects.

This has meant that while conventional statistics teaching has concentrated on the manipulative aspects of deriving estimators, conducting tests, fitting models, and so on, issues of when to apply what tools and what to do when things go wrong have been left to be learnt in an apprentice role, 'on the job'. It is as if one was being given the individual bricks and left to watch others on the building site to learn how the bricks should be put together to make the house.

However, all of this is changing under the impact of the computer. Nowadays statistical software handles the formal arithmetic and mathematical manipulations, so that a researcher undertaking an analysis need not understand the algorithms involved. A trivial example would be the calculation of least squares regression coefficients. The mathematics involves knowledge of differential calculus, and the arithmetic involves knowledge of how to calculate sums of products and squares. But if a computer is to undertake the calculations, what the researcher (or, more generally, the user of least squares regression) needs to know is that a sum of squared deviations criterion is used and that the resulting estimated regression coefficients are those that minimise this criterion. Being able to reproduce what the computer is doing (much faster than a human could do it) is all very well, but is of little practical value, whereas being able to formulate the research question and then being able to use the tool (the computer) in an appropriate way and place is vital.

Of course, we acknowledge here the importance of knowing what is going on inside the computer if one is concerned with developing new methodology. Most users of statistics, however, are not. They want to apply standard techniques (probably using standard software) – and they therefore need to know if their problems fit such techniques.

This, therefore, leads us to the first step of a statistical strategy: the clear formulation of research aims, and their expression in terms of statistical tools.

It is my contention that all too often clients seek statistical advice without a clear formulation of their research objectives and, perhaps worse, statisticians undertake analyses without clarifying these objectives, so that inappropriate or incorrect conclusions are often drawn.

This conference is the fourth in the series. The previous papers I have presented represent steps towards my present position (Hand, 1986, 1990, 1992). The first was on high level statistical strategy, attempting to illustrate with a strategy for multivariate analysis of variance. Then I thought that the difficulties in formulating strategies, while by no means trivial, could be tackled by straightforward attempts to formalise statisticians' actions in tackling problems. The second paper described the construction and experiences with systems for helping researchers apply statistics to problems. I had here stepped back from the notion of an 'expert system' and described

statistical 'knowledge enhancement systems', aimed at supporting the user's knowledge rather than providing a whole additional layer of expertise. The third focussed on the use of metadata in statistical expert systems, but concluded: 'it is not clear that the researcher involved can always make sufficiently precise statements about the research objectives, and a statistician is needed to identify subtle differences between questions.' In the present paper I go a step further and suggest that even statisticians often do not think carefully enough about the initial step in a strategy – precisely what it is the researcher wants to know. This paper presents some examples to illustrate this contention.

1.2 Some examples

1.2.1 Pragmatic versus explanatory comparisons

The distinction between pragmatic and explanatory studies is best known in medical areas, but it is in fact ubiquitous. It is discussed at length by Schwartz, Flamant, and Lellouch (1980). In brief, and in medical terms, a pragmatic trial is concerned with the practical effectiveness of a treatment, as administered in the way it would be used in practice, while an explanatory trial aims at discovering real biological differences. Although perhaps superficially identical, these two questions in fact are very different. The differences have consequence for both design and analysis and yet the distinction is all too rarely made in statistical practice.

Consider, for example, a comparison between two propose treatments. In a pragmatic study the overriding concern would be that one should not identify the poorer treatment as the better. (Such errors have been termed 'Type III' errors.) In such a situation the Type I error would not matter (if the two treatments were equally effective – the null hypothesis – then it would not matter which was chosen). In contrast, if the study was explanatory then it would be important to keep both Type I error and Type II error small. Since the two types of question have different error structures, we might expect that different sample sizes would be needed to address the questions.

In many studies, especially in clinical trials, measurements are taken over a period of time so that withdrawals from the programme can cause a problem. In a pragmatic study one might decide to retain the dropouts in the analysis, coding such cases as treatment failures – they certainly are not successes. In contrast, in an explanatory study, one wishes to make statements about patients who have stuck rigorously to the treatment regimen, and hence would exclude such dropouts from the analysis. (This is, of course, something of a simplification, and in practice one should consider carefully precisely what the objectives are.)

Yet another difference arises in the subjects being studied. In a pragmatic study one is hoping to make inferences which apply to a future population to which the treatments may be given. The test sample should thus be sampled from that population and should adequately represent it – including all its heterogeneities. In contrast, in an explanatory study one will aim to minimise variation – and hence use as homogeneous a sample as possible (relying on implicitly non-statistical inferences to subjects which differ from those in the sample?).

Schwartz et al (1980) give the following striking example of the difference between the two types of research question. Consider a study to explore the effectiveness of a sensitising drug prior to radiotherapy in treating cancer. We wish to compare two treatments:

T1: radiotherapy alone.

T2: radiotherapy preceded for 30 days by a sensitising drug.

In an explanatory trial one will wish to see how the sensitising drug influences the outcome, without there being any other differences between the two treatment groups apart from the fact that one has used the drug and the other has not. Now, group T2 have to wait 30 days before they can commence radiotherapy. Therefore, to make the groups comparable in this regard, group T1 should also have to wait 30 days. Presumably they should be given a placebo medication throughout this period.

In contrast, in a pragmatic trial one wants to compare the treatments that would actually be used in practice. And in practice, the radiotherapy in T1 would commence immediately, and not after 30 days treatment with a drug known to have no effect!

The two questions, explanatory or pragmatic, are quite distinct and it is essential for the researchers to know which they wish to address even before the data can be collected.

1.2.2 Interaction

Table 1, from Brown and Harris (1978), shows a cross- classification of a sample according to whether or not they had recently experienced a life event, whether or not they had a close relationship with a husband or boyfriend, and whether or not they developed depression. The hypothesis was that a close relationship was protective to some extent against the effect of life events in inducing depression. To explore this hypothesis, Brown and Harris computed the proportions developing depression in each of the four conditions resulting from the cross-classification of the relationship variable and the life event variable, to give Table 2. From this table it is easy to see that $(0.32 - 0.03) = 0.29$ is not equal to $(0.11 - 0.01) = 0.10$, and Brown and Harris consequently concluded that there was an interaction in the effect of these variables on the probability of suffering from depression.

Analysing the same data, however, Tennant and Bebbington (1978) questioned their conclusion. They reasoned that since $0.32/0.03 = 10.7$ was approximately equal to $0.11/0.01 = 11.0$ there was no evidence for an interaction.

The resolution, of course, lies in the definition of interaction. That is, it lies in making precise the initial research question. Brown and Harris are using an additive model whereas Tennant and Bebbington are using a multiplicative model. As with the pragmatic/explanatory distinction outlined above, attempting to choose a model is futile unless one considers why one wants it – precisely what question one intends to use the model to address.

Another example of the confusion arising from the distinction between additive and multiplicative models arises in investigations of the interaction between medicines. A combination of two drugs may be more effective than if the drugs act independently (in which case the combination is said to exhibit 'synergy') or less effective than if the drugs act independently (and the combination is said to exhibit 'antagonism'). For example, childhood acute lymphatic leukaemia shows a 40-50% remission with individual drugs but a 94-95% remission with a combination of three drugs. However, these sorts of statements (and, indeed, the definitions of 'synergy' and 'antagonism') are only meaningful if there is a clearly defined 'independence' model with which to compare the results of the combination. Unfortunately there is no unique definition and many proposals have been made.

Obvious and common ones are the additive and multiplicative ones. Let $E(da)$ signify the effect of a dose da of drug A, $E(db)$ that of a dose b of drug B, and $E(da, db)$ that of a combination of dose da of drug A with db of drug B. Then the additive definition says that there

Table 1. Depression and life events

	No close relationship		Close relationship	
	Event	No event	Event	No event
Depression	24	2	10	2
No depression	52	60	78	191

Table 2. Proportions developing depression

	Event	No event
No close relationship	0.32	0.03
Close relationship	0.11	0.01

is no interaction between the two drugs if

$$E(da, db) = E(da) + E(db) \tag{M1}$$

and the multiplicative definition says that there is no interaction if

$$S(da, db) = S(da) * S(db) \tag{M2}$$

where $S = 1 - E$, defined when E is a fractional effect.

Suppose, however, that $E(da) = E(db) = 50\%$ and $E(da, db) = 90\%$. Then measure $M1$ (with 100% expected under the no interaction model) suggests that the combination is antagonistic while measure $M2$ (with 75% expected under the no interaction model) suggests that the combination is synergistic.

Measure $M1$ is satisfactory if both drugs have linear dose/response relationships and $M2$ is satisfactory if both have exponential dose-response relationships. However, neither is generally valid: neither works if the dose-response relationship is (for example) sigmoidal and neither works if the relationship differs between the drugs.

1.2.3 Measuring relative change

Consider the exchange rate between the £ and the $. At the time of writing it is about 1 = $1.5. Not so long ago, however, it was 1 = $2. This means that the £ has dropped to 75% (= 1.5/2) of its previous value relative to the $. In contrast, however, the $ has increased to 133% (= (1/1.5)/(1/2)) of its previous value. Thus while one currency has lost only 25% in value the other has gained 33%. Which of these it is appropriate to use clearly depends on the question one is trying to answer (or, perhaps, the point one is trying to make?). Tornqvist, Vartia, and Vartia (1985) have more to say about this.

1.3 Conclusion

The purpose of these examples has been to illustrate the assertion that often too little care is taken in formulating the reasons for constructing a model, with the consequence that sometimes an inappropriate model may be selected. Resolution of the problem will not be achieved through more sophisticated search algorithms or ways of playing off accuracy against numbers of parameters. Before such considerations arise it is necessary to decide what the model is for – what question it is being constructed to answer.

In the introductory section it was remarked that adequate formulation of the research question is the first step in any statistical strategy. Given that, it is legitimate to ask how far back one can push the 'question formulation' aspects without requiring extensive application domain knowledge. That is, what general aspects are there which might be used in the question formulation stage? Hand (1993) has considered this in more detail. Certain basic issues can be identified. These are things which all statisticians will know about but which would benefit from being explicitly stated in the early discussions with researchers.

One example is the issue of control: when deciding on the appropriate model form, what other variable, terms, or contrasts should be controlled for? For example, should one use simple or multiple regression? Both result in legitimate models; they merely address different issues. The same phenomenon is evident in Simpson's paradox, which arises because of a confusion between two questions, one conditional and the other unconditional. Another way of looking at this issue, and another interpretation which has wide applicability, is to ask whether one wishes the results of the analysis to be a statement about populations or about individuals.

Trying to identify basic issues and explicitly consider them during the question formulation stage is one strategy for alleviating the difficulties. Another is to try to formulate a model such that the different questions are equivalent under the model assumptions. A trivial example would be the case when the researcher cannot decide whether the mean or median is an appropriate summary statistic to use in addressing the question 'Which of two groups has the larger average?' If it is legitimate to assume symmetry of the distributions then the questions become logically equivalent (and one can then choose on other grounds, such as power of the test). The model – that the distributions are symmetric – has meant that the issue of precisely which question is being asked is irrelevant. Of course, in using this approach, one must be confident that the chosen model is legitimate.

The computer has given us the freedom to concentrate on higher level strategic issues, instead of being forced to focus on lower level arithmetic and algebraic manipulations. These higher level issues are arguably even more important than the lower level issues. No matter how accurate an answer is, it is useless if it addresses the wrong question.

1.4 References

[1] Brown G.W. and Harris T. (1978) Social origins of depression: a reply. *Psychological Medicine, 8*, p577-588.

[2] Cox D.R. and Snell E.J. (1981) *Applied statistics: principles and examples*. London: Chapman and Hall.

[3] Hand D.J. (1986) Patterns in statistical strategy. *In Artificial intelligence and statistics*, ed. W.A.Gale. Reading, Massachusetts: Addison-Wesley, p355-387.

[4] Hand D.J. (1990) Practical experience in developing statistical knowledge enhancement systems. *Annals of Mathematics and Artificial Intelligence, 2*, p197-208.

[5] Hand D.J. (1992) Measurement scales as metadata. In *Artificial Intelligence Frontiers in Statistics*, ed. D.J.Hand. London, Chapman and Hall.

[6] Hand D.J. (1993) Deconstructing statistical questions. In preparation.

[7] Schwartz D., Flamanat R., and Lellouch J. (1980) (Trans. M.J.R.Healy) *Clinical Trials*. London: Academic Press.

[8] Tennant C. and Bebbington P. (1978) The social causation of depression: a critique of the work of Brown and his colleagues. *Psychological Medicine*, *8*, 565-575.

[9] Tornqvist L., Vartia P., and Vartia Y.O. (1985) How should relative changes be measured? *The American Statistician*, *39*, p43-46.

2

Rational Learning: Finding a Balance Between Utility and Efficiency

Jonathan Gratch, Gerald DeJong and Yuhong Yang

Beckman Institute for Advanced Studies
University of Illinois
405 N. Mathews
Urbana, IL 61801

Beckman Institute for Advanced Studies
and
Department of Statistics
Yale University
New Haven, CT 06520

ABSTRACT Learning is an important aspect of intelligent behavior. Unfortunately, learning rarely comes for free. Techniques developed by machine learning can improve the abilities of an agent but they often entail considerable computational expense. Furthermore, there is an inherent tradeoff between the power and efficiency of learning techniques. This poses a dilemma to a learning agent that must act in the world under a variety of resource constraints. This article considers the problem of *rational* learning algorithms that dynamically adjust their behavior based on the larger context of overall performance goals and resource constraints.

2.1 Introduction

The field of machine learning has developed a wide array of techniques for improving the effectiveness of performance elements. For example, learning techniques can discover effective classifiers and enhance the speed of planning systems. Learning techniques can take general performance systems and tailor them to the eccentricities of particular domains. Unfortunately, this capability is not without sometimes considerable cost. This expense arises from three sources: first is the cost of attaining a sufficiently large sample of training problems; second is the cost of properly configuring the learning technique (e.g. the determination of parameter settings) which can substantially influence performance; finally there is the computation complexity of the learning algorithm.

Specific learning algorithms can mitigate some of the sources of learning expense. Learning systems incorporate many compromises, or biases, to insure efficient learning. Typically these compromises appear as implicit design commitments, and are fixed into the implementation of learning techniques. These compromises impose tradeoffs between the efficiency and power of a learning technique - more powerful learning techniques generally require greater expense - and the fixed nature of these commitments limit generality. Ideally, a learning system would possess the flexibility to adapt its bias to the demands of a particular learning situation.

---

[0]This research is supported by the National Science Foundation under grant NSF-IRI-92-09394.

[1]*Selecting Models from Data: AI and Statistics IV.* Edited by P. Cheeseman and R.W. Oldford.

In this article we discuss the use of decision theory to control the behavior of a learning system. Many authors have illustrated the usefulness of decision theory as a framework for controlling inference [Doyle90, Horvitz89]. Decision theory provides a well-understood framework for expressing tradeoffs and reasoning about them under uncertain information. The work described in this article can be seen as an application of decision-theoretic meta-reasoning to the control of learning algorithms. We describe the general principles involved in such an approach. We then present an instantiation of these principles, showing how they can be applied within the COMPOSER learning approach, a statistical approach to improving the efficiency of a planning system [Gratch92a]. The extended system dynamically adjusts its learning behavior based on the resources available for learning.

2.2 Rationality in learning

The goal of learning is to tailor a performance element to be effective in some specific environment. For example, the performance element could be a planner where the input is problem specifications; the output is plans. The environment is simply the tasks faced by a performance element and adaptation to this environment must be judged against some criteria for success. In planning, criteria include accuracy, planning efficiency, and plan quality. We adopt a decision-theoretic view of learning. An environment is characterized by a probability distribution over the set of possible tasks. The user provides a *utility function* that specifies a criterion for success on individual tasks. The effectiveness of a performance element is characterized by its *expected utility* over the task distribution. This is the sum of the utility of each task weighted by the probability of a task's occurrence. For example, in classification problems, the standard utility function assigns a value of one to a correctly classified feature vector, and a value of zero to an incorrectly classified vector. The expected utility of a classifier is equivalent its accuracy over the distribution of feature vectors.

This view of learning facilitates a close analogy between learning and work in rational reasoning. In reasoning there is a decision procedure that must choose, from a set of possible actions, an action with high expected utility. In learning, there is a decision maker that must choose, from amongst a set of possible changes to a performance element, a change that produces a performance element with high expected utility. In fact, one might make the argument that reasoning subsumes learning, by allowing "learn something" to be one of the reasoner's actions. Nonetheless, there are issues unique to learning that justify their individual investigation as a specialization of general reasoning. For example, deliberating about possible actions involves past information on these action's effects. Deliberating about learning involves future information, and its impact on action deliberation.

In reasoning, a reasoner that always chooses the action with maximal expected utility exhibits *substantively rationality* [Simon76] (also called Type 1 rationality [Good71]). Similarly, we can define a *substantively rational learning system* as a system that always identifies the transformed performance element with maximum expected utility. Substantive rationality is seldom attainable in that it may require the use of infinite resources. This has led to a focus on rationality under limited resources. Simon refers to this as *procedural rationality* (also called Type 2 rationality [Good71]) because the focus is on identifying efficient procedures for making adequate decisions. A procedurally rational agent relaxes the strict requirements of substantive rationality in the interest of reasoning efficiency. The analogy between reasoning and learning