EXPERT SYSTEMS IN STATISTICS

By
MARIA DAMI and ALFRED VELLA
Cranfield Institute of Technology Cranfield, Bedford England

ABSTRACT

The problem of misuse of statistical packages and hence statistical methods has long peen recognised but no real solution has yet appeared, although some useful work has been done in related fields. This problem appears to lend itself to expert system technology.

This paper outlines the history of research in statistical expert systems, showing how the demands of data analysis are different from those of other fields. Furthermore, the paper reviews some recent work and discusses the development of an expert system for the planning of experiments.

1. INTRODUCTION

Prior to the development of expert systems computers were mainly used for developping programs to suit the users needs. The need for more user friendly programs that could be used by people not highly familiar with the program itself, or the programming language, led to the development of packages. Most packages nowadays are easy to be used and very readily available to everyone. Experience though has shown that those highly user friendly packages are as easy to misuse as it is to use. At this stage action needed to be taken to make programs more

«intelligent», so as not only to be user friendly but also to prevent the user from making serious mistakes. In general there was need to make the programs behave more or less like a human consultant would behave in similar situations. This is when the expert systems appeared.

Expert systems have been variously defined. Some of the most succesful, definitions are:

i) «An expert system is one that.

— Handles real world complex problems requiring an expert's interaction.

—Solves these problems using a computer model of expert human reasoning searching the same conclusions that the human expert would reach if faced with a comparable problem».
(S.M. Weiss and C.A. Kulikowski (1984).

ii) «An expert system is a computer program containing rules rather than the procedures and the functions of the conventional software. The rules express human knowledge in the relevant application. A common form of rule is IF condition THEN action or conclusion (eg. IF symptoms are spots and temperature THEN may be measles).

Application of these programs include advisory systems where the user-consults the stored expertise (eg. medical diagnosis) and decision making in or for a processing task (eg. assigning the user exploit a complex software package).

A vital feature of many «Expert systems» is their ability to explain their reasoning intelligibly to a human enquirer. This feature should become universal within such systems because it human human makes:

a) computers more fit for usage;

b) programs easier to change and extend;

c) applications more independent of any particular machine».

British Computer Society Comittee speslalist group on expert systems, [Ferbr 1983].

A great deal of attention and effort has been focused on developing expert systems for application in a great many different domains ranging from geology

through medicine to computer configurations and accountancy. The need for the development of expert systems arises because of the time it takes to train, experts on a specific field and their fallibility. Therefore the scarcity of those experts, and the expence of employing them. Furthermore, since computers are becoming more and more commonplace in any domain, why not implant in them the knowledge of an expert, so that many more users could benefit from this knowledge.

In recent years statistics has also become a candidate for implementation of the expert system approach. Statisticians are experts in planning experiments and in analysing and presenting data. Statistical expertise is hard to find especially in specialised areas. According to Hand (1985a) it typically takes around ten years to train an expert consultant statistician, whose charge fees are comparable with those of doctors in private practice.

2. EXPERT SYSTEMS IN STATISTICS. HOW THE NEED EVOLVED

Nowadays there are a fairly large number of reliable and well-documented statistical packages which operate on mainframe computers, and many of these have been ported onto microcomputers (Neffendorf (1983). Current statistical software contains both arithmetic and algebraic expertise. That is the user provides the numbers and the system carries out the numerical manipulation and produces an answer. More complex established packages (such as SPSS, SAS, and GLIM etc.) require the user to have some programming abilities in order to use them. Modern packages on the other hand are designed to be easy to use by people with little statistical expertise. Unfortunatelly this has led to a misuse of the many statistical methods available.

As the number of statistical software have increased, so has the number of users. Many of these users though often experts in their own fields are «naive» in their understanding of the principles underlying statistical analyses. Thus the already existing problem of misuse of the methods becomes increasingly a pertinent one.

Chambers (1981) summarises it thus: «Statistical software in its present forms made widely available by cheap computing will precipitate much uninformed unguided and simply incorrect analysis. We are obliged to help».

Hand (1986) also refers to software misuse by giving examples: «Some pro-

grams (the MANOVA program in SPSSX, for example) are particularly vulnerable to this kind of criticism (criticism of misuse) by very virtue of the fact that they have been made easy to use. Easy use implies easy misuse, and the problems typically tackled by MANOVA require a good grasp of the theoretical background — and of the way this particular program tackles things — before one can be confindent of avoiding mistakes. Other systems, (eg. MULTIVARIANCE and GENSTAT) are vulnerable to the opposite kind of criticism, these are not very easy to use. One has to be something of an expert oneself to use them all. Since only relative experts will use them, there is less chance of them being misused».

The problem has also been discussed at great length in the statistical literature (eg. Nelder 1977, Hooke 1980, Hunter 1981) and people have already started working towards a solution (eg. Gale and Pregibon 1982, Hand 1984, O'Keefe 1982, and others).

According to Hahn (1985): «A natural consequence of the evolution of statistical computing is the desire to have the computers that now perform user-defined statistical analyses also provide guidance on which analyses to conduct and how to interpret the results», while he stresses the fact that «.. .embedding expert guidance in statistical programs is a technically challenging but highly worthwhile undertaking. Automating statistical consulting and data analysis in the form of a statistical expert system is fraught with difficulties», but as the same adds, «Despite serious technical obstacles, the trend toward more intelligent statistical systems is inevitable».

3. WHAT SHOULD STATISTICAL EXPERT SYSTEMS DO

For one to have a clear understanding of what expert systems are designed to do one should realise the following questions.

3.1. What will expert systems be used for ?

Statistical knowledge based system should be built for either or both of the following reasons.

i) The design of studies and the selection of strategies.

ii) The conducting the analyses, where the computing environment provided will be suitable for studying the behaviour of expert data analysts with particular attention to the strategies of data analysis they adopt.

Reason (i) is clear enough and as Hand [1985] states : «As in medicine, for example, there exist many expert systems designed to give a diagnosis and recommend a therapeutic regime, we could design a statistical system which served an analogous role which, in interaction with the user, recommended a series of analysis which would answer the users questions». For reason (ii) even though programs that carry out analyses do exist in the vast amount of statistical software, statistical expert systems that analyse data do not.

In other words the idea to construct an expert consultant for the design of studies and data analysis, whose function is to assist a practiced —if not expert— data analyst to produce a better more complete, or more explicable analysis of a data set than otherwise might be possible (Thisted [1986]).

Finally we quote from Hand (1986) as to what expert systems are used for. He says: «We can best answer this question by taking a step backwards and asking what consultant statisticians do. The answer that they talk to clients and

a) Refine the research objectives and questions.

b) Choose appropriate statistical techniques to address questions.

c) Apply the techniques.

d) Interpret the results».

.

Having asked what statistical expert systems will be used for, the following question now arises.

3.2. Who is the expert system designed for ?

Amongst statisticians there seems to be the idea that expert systems should necessarily be designed for users at the extremes ie.

a) Statistically naive researchers.

b) Professional statisticians.

(eg see Hand (1985, 1986, 1987), Smith Lee and Hand (1983), Hahn(1985), Gale (1986), Gentle (1985)).

In most other application areas of expert system technology there is usually the assumption that either one or the other of these extremes is the primaty target population to use the system. For example, most medical applications, assume that an expert, or at least someone with a relatively sound background in the field will use the system.

This need for designing expert systems in one or the other extreme is clearly discussed by Rector, Newton and Marsden(1985), who analyse the situation where experts are the potential clients and by Kidd (1985), whose primary concern is the consultive role of an expert system.

In statistics however, the requirements are that both extremes will use the system. Therefore the strategy that must be followed in developing an expert system in statistics should have a number of features.

4. STATISTICAL EXPERT SYSTEM'S DESIGN STRATEGY

The features regarded as necessary for an expert system in statistics to have are summarised bellow.

1. The system should be used by both the novice and the expert. The system' should allow easy cooperation between itself and the user without being too sofi-sticated for the novice or too boring for the expert.

2. The second feature that an expert system should have is the capability to explain itself. For example to explain why a certain question was asked, or why a particular method was chosen. Furthermore it is equally important for the system to be able to answer «why not» questions as it is to answer the «why» questions. O'Keefe [1982], developer of ASA (Automated statistical Analysis), claims: «An expert system should discuss its knowledge as well and as simply as it uses it».

3. Another important feature is that the system should have the ability to

. . . . . .
explain statistical terminology. Statistics has its own terminology and this is a real barrier for the naive user. Some form of classification module must be incorporated. Note also must be made that it is equally important for the system to be able to give explanations whilst in the midst of a consultation session without losing its place.

4. Another feature which should be adopted in systems of which novices are the likely users is that the system must be able to cater for the user making a mistake.

5. One very important feature to all expert systems is that the questions asked during the session should be presented in a sensible or der. Only then will the user and especially the novice be able to understand the nature of his/her problem (that is in most cases unknown) and also to realise the whole way of thinking towards its solution. Furthermore, for simple reasons of user acceptability is not a good idea to leap from one topic to another.

6. In choosing statistical tools, some or dering must be established so that the most powerful techniques available are recommended.

7. The system must be able to give more than one answer at a time. It is usual to find that a researcher has multiple objectives.

8. A statistical expert system must be able to examine the data critically and use the results of this examination, in connection to the users objectives in determiring the choice of methods, and suggest appropriate courses of action. An experienced user must then permitted, if he so wishes, to insist on some particular approach, whilst an inexperienced user can take advice.

9. Both an appropriate test and suitable data exploration methods have to be identified by the system.

10. An absolutely essential requirement for a statistical expert system is that it must be able to adapt to changing circumstances as analysis proceeds, and identify suitable methods. Very common adaptations are for outliers, missing data, non-normal distributions etc.

11. A point that must be stressed for every expert system is that ideally it

should be structured so that it can be easily modified. It is vital for it to be flexible so that new techniques to be added and new knowledge to be introduced when needed and available. The ability for correcting any knowledge found to be inadequate or wrong is also essential.

12. The system should be able to store the results of a consultation session so that it can be recalled on a latter stage. This feature applies for statistical exper systems especially but is also important for systems in other areas, since this feature will help the user enormously when for example the system is designing experiments and chooses a method of analysis. When the data have been collected at a later stage, the method of analysis will be applied directly without requiring the user to go through the whole system again.

13. Finally, the system itself should be as self-documented as possible. Users will be higly dissapointed when they need to study endless documentation on the use and application of the expert system. Especially in statistics, where in most of the time expert system will be used by researchers for whom statistics is not of primary interest. The researchers will use the system infrequently and therefore it has to be self-explanatory.

-

5. SOME PREVIOUS EXPERT SYSTEMS

In the seventies, the first programs for solving problems that require expert knowledge were developed for application in medical diagnosis (ie. MYCIN, CASNET, INTERNIST and PUFF c.f. Spiegelhalter and Knill—Jones (1984), and chemical analysis (DENDRAL: Buckhanam and Fiegenbaum (1978).

Also in the seventies attempts were made to define what types of semantic errors could (or should) be avoided by «more intelligent» statistical analysis, system. Applications of Artificial Intelligence in statistics, consequently was soon introduced by Hajek and Invanek (1982), for hypothesis generation in the field of Explanatory Data Analysis.

Bob Blum's thesis (1982) on RX, was slow to receive the recognition it deserved as an important contribution to Artificial Intelligence in statistics. RX

included hierarchical representations of medical concepts and statistical methods, together with a casual network among the medical concepts.

The prototype system ASA (R.O'Keefe (1982)) illustrated what expert systems can contribute to statistical data analysis. It was one of the first attempts at making a system that was able to say which method was good and which not in specific sircumstances. It also explained in a very superficial way why a question was asked and explain how a particular method was chosen or rejected.

In Portier and Lai (1983) they describe a system called STATPATH which is the only system yet to tackle the problem of user not understanding a question or misunderstanding one. STATP ATH uses production rules to encode a binarty choice tree to help the user select an analytic technique.

Pregibon and Gale (1984) began working on REX late in 1981. REX gave advice on regression analysis searching for problems in the data and proposing actions to remedy any problems found. REX was the first system to use expert system techniques succesfully, to choose commands for a statistical package and to carry out complete analysis. The only problem with REX is that is highly sofisticated and the user either had to be an expert or to have consulted an expert prior to using REX. In Pregibon and Gale (1984) they discuss their ideas for STUDENT.

STUDENT provides quidancc to what tests need to be done and when, interpretation of the results of tests and plots, and instruction in statistical concepts. It has appeared that the system although designed for use by novices, is of interest to expert statisticians because it makes explicit much knowledge that has not been formalised. Most experts have also expressed interest in using such a consultation system because it automates many tasks that they know thay want to do **but** don't always do |Gale (1986a). STUDENT is basically used as a **front end** to REX.

Hand [1987] and colleagues developed a knowledge enhancement system (KENS) which provides the user with information about nonparametric statistical methods.

Finally, Bell and Watts [198η developed an expert system (THESEUS) which concentrates on the area of one-way Analysis of Variance (ANOVA), and related techniques.

6 EXPERT SYSTEMS IN THE PLANNING OF EXPERIMENTS

As mentioned above, the computer has played an important role in the analysis of results from comparative experiments. The data from large experiments can now be analysed quickly and complicated experimental designs can now be dealt with as a matter of routine. It is reasonable therefore, for the experimenter to expect the computer to provide assistance at the design stage of an experiment. Statistically informed experimenters can of course use a number of computerised algorithms to construct designs that have certain optimal properties but there does not appear to be a computer program that can be used by the layman.

The need for developing an expert system for Experimental design comes from our awareness of the problems faced by many experimenters. All too often they perform a series of intricate experiments taking a great deal of time and effort. Once the results have been collected they then begin to look for some meaning in the numbers obtained.

After much soul-searching, sweating in their libraries and many lost hours pouring over these «numbers» they decide to seek the aid of their institutions statistical advisors. Often these advisors are unavailable or very busy. The statistician rightly complains that the analysis stage is not the time to seek advise.

It is for these experimenters that we are designing our expert system. It is able to «join in» with the experimenter at any stage of his experiment. Clearly the earlier the better.

The system will help him design an efficient experiment, help him to choose the best analysis methods and eventually run the analysis under his control. He will be able to «consult» the expert system on different aspects of his work as he requires.

One of the novel features of our system is the existence of a number of «discipline interaces». Briefly these perform the communication between the users and our system. Each of these has a «user model» which defines the «prefered language» of the user. For example if the user is a statistician checking the suitability of some design he has made, then the system uses the language of statistics. Words such as «factors», «treatments» etc. appear freely. On the other hand if the user is an engineer the system must use a different language, checking as it does so that the user agrees with its concept of any technical terms used.

We have not yet completed the development of our expert system but fell that we should seek the views of others through this paper.

One of the major problems faced by an expert system faced by an expert system builder building an expert system for use by novices is that of debugging the knowlege base. No matter how carefully designed and implemented this knowlege will have mistakes in it. It is by encouraging «experts» to use our system that we expect to improve our debugging of the knowlege base.

Thus we believe that it is most important that we target the system towards both the naive researcher and the experienced statistician. For the first the system will be a consultive and teaching aid, whilst for the later it provides a second opinion. This dual user-base will also encourage us to keep the system updated with the latest results from statistics.

Our plans are to develop a system adopting as many as possible of the stra. tegic features described in above. Furthermore we wish to make the system to suggest experimental analysis both for the experiments not yet planned and for the experiments that have been carried but prior to the consulting session.

7. CONCLUSION

There appears to be little that one can say against skillfully embedding improved intelligence into statistical programs. Since nonstatisticians will be using statistical programs irrespective of whether statisticians approve, we should do our best to ensure that this is done in the most meaningful manner possible.

REFERENCES

1. Bell E. and Uatts P. (1987) : «Theseus : An Expert Statistical Proc. Consultant» DO - SES, Luxenburg 1987.
2. Blum R. L. (1982) : «Descovery and Representation of Casual Relationships from a Large Time - Oriented Clinical Database» The RX Project, New York, Spiegel Verlag.
3. Buckhanam, B.G. and Fiegenbaum E. A. (1978) : DENDRAL and META - DENDRAL their application dimensions», Artificial Intelligence, 11, 5-24.
4. Chambers J. M. (1981) : «Some thoughts on Statistical Software», Proc. 13th Symposium on the Interface, Pittsburg, PA.
5. Gale, W. A. and Pregibon D. (1982) : «An Except System for Regressio η Analysis», Proc. 14th Symposium on the Interface, New York.
6. Gale, W. A. (1986) : «Overview» In Artificial Intelligence and Statistics, ed. Gale W. A. Addison Wesley, Massachussetts.
7. Gale, W. A. (1986) (a) : «STUDENT Phase 1-A period in progress» In Artificial Intell-gence and Statistics, ed. Gale W. A., Adisson Wesley, Massacnusetts.
8. Gentle, J. E. (1985) : Comments on Hahn G. J. «More Intelligent Statistical Software and Statistical Expert Systems : Future Directions» The American Statistician 39, 1-16.
9. Hajek, P. and Invanek J. (1982) : «Artificial Intelligence and Data Analysis» In Caussinus H., Ettinger P., and Tomassone R. (eds), COMSTAT 1982 - Part I, Proc. in Computational Statistics, Wien : Physika 54 - 60.
10. Hahn, G. J. (1985) : «More Intelligent Statistical Software and Statistical Expert Systems: Future Directions» The American Statistician, 39, 1-16.
11. Hand, D. J. (1984) : «Statistical Expert Systems : Design». The Statistician, 33, 351 - 369.
12. Hand, D.J. (1985): «Statistical Expert Systems: Necessary Atributes», Journal of the RSS, B. 12 (1).
13. Hand, D. J. (1985) : (a) : The Role of Statistics in Psychiatry», Psychological Medicine, 15, 471 - 476.
14. Hand, D. J. (1986) : «Expert Systems in Statistics», The knowledge Engineering Review, 1, 1 - 10.
15. Hand, D. J. (1987) : «A Statistical Knowledge Enhancement System», Journal of the RSS, A, 150, 334 - 345.

16. Hooke, R. (1980) : «Getting People to Use Statistics Properly», The American Statistician 34 (1), Febr.

17. Hunter, W. G. (1981) : «The Practise of Statistics : The Real World is an Idea Whose, Time has Come», The American Statistician, 35 (2), May.

18. Kidd, A.L. (1982): «The Consultive Role of an Expert System» In people and Computers : Designing the Interface, ed. Johnson and Cook, Cambridge University Press, 248 -254.

19. Neffendorf H. (1983) : «Statistical Packages for Microcomputers : A Listing», The American Statistician, 37 (1), Febr.

20. Nelder J. A. (1977) : «Intelligent Programs, The Next Stage in Statistical Computing», In Recent Developments in Statistics, Barra J. R. et al. (ed), North Holland Publishing Company.

21. O'Keefe, R. (1982) : «An Expert System for Statistics», Appeared in the Expert Systems, '82 Technical Conference. Theory and Practice of knowledge Based Systems. Brunei University.

22. Portier, K. M. and Lai P. (1983): «A Statistical Expert System for Analysis Determination» Proc, of the ASA Statistical Computing Section, 309 - 311.

23. Pregibon D. and Gale W. A. (1984) : «REX : An Expert System for Regression Analysis» Proc. CoMSAT, 84, 242 - 248, Prague Chechoslovakia.

24. Smith, A. M., Lee L. S. and Hand D, J. (1983) : «Interactimve User-Friendly Interfaces to Statistical Packages», The Computer Journal, 26, 199 - 204.

25. Spiegelhalter D. J. and Knill - Jones R. P. (1984) : «Statistical Knowledge Based Approaches to Clinical Dimension - Support Systems, with an Application in Gastroenterology», Journal of the RSS, A. 147, 35 - 77.

26. Rector, A. L., Newton P. D. and Marsden P. H. (1985) : « What Kind of System does an Expert Need ?» In people and Computers :Designing the Interface, Iohnson and Cook (ed), Cambrigdge University Press, 239 - 247.

27. yhisted R. A. (1986) : «Representing Statistical Knowledge and Search Strategies for Expert Data Analysis Systems», Gale W. A. (ed), Addison Wesley, Massachsetts.

28. Weiss S. M. and Kulikowski C. A. (1984) : «A Practical Guide to Designing Expert System's» Roman and Allanheld Publishers, ISBN 0-86598-108-6.