Data Explorer: A Prototype Expert System for Statistical Analysis

Constantin Aliferis, M.D., Evelyn Chao, M.A. and Gregory F. Cooper, M.D., Ph.D.
Section of Medical Informatics & Intelligent Systems Program
University of Pittsburgh, Pittsburgh PA

ABSTRACT

The inadequate analysis of medical research data, due mainly to the unavailability of local statistical expertise, seriously jeopardizes the quality of new medical knowledge. Data Explorer is a prototype Expert System that builds on the versality and power of existing statistical software, to provide automatic analyses and interpretation of medical data. The system draws much of its power by using belief network methods in place of more traditional, but difficult to automate, classical multivariate statistical techniques. Data Explorer identifies statistically significant relationships among variables, and using power-size analysis, belief network inference/learning and various explanatory techniques helps the user understand the importance of the findings. Finally the system can be used as a tool for the automatic development of predictive/ diagnostic models from patient databases.

INTRODUCTION

Although the traditional focus of Medical Informatics has been the optimal use of existing medical knowledge through the application of computer and information processing technologies [1], a number of researchers in the field, during the last decade, have concentrated on better methods for discovering, and disseminating medical knowledge. These efforts suggest the importance of producing high-quality medical knowledge in the first place, as the sound basis for every subsequent effort to apply it towards the improvement of health care [2-7].

The work done so far addresses different components of the research cycle: proper use of the literature [2], automation of the scientific discovery process [3,4], meta-analysis of published data [5], semi-automated normative decision-support systems using both models of research designs and the user's beliefs and assessments [6,7]. Clearly, an area where many problems remain to be solved is that of the proper analysis of commonly collected data (in the context of everyday clinical information storage).

In many research settings, statistical experts are not always available, nor in general, is the interaction with them a trivial part of the research process [8]. Thus, many medical investigators often rely on their limited statistical background to carry out one of the most sensitive phases of research.

As a result, over the past years, much evidence has accumulated showing that inappropriate statistical analysis is a common pitfall in medical research [9]. A recent study [10] shows that statistical review of papers submitted to a major clinical journal and accepted by the medical reviewers, revealed large percentages of errors in analysis and/or inference. Errors typically include: no statistical significance testing, no attention to distribution shapes, inappropriate statistical tests application, numerical errors, interpretation errors and poor sample size, among others. Of course strict peer review can reveal mistakes, but still errors can be missed (especially when editorial resources are limited) and even if they are caught, much effort, time and money will have been spent.

A more specific but closely related phenomenon has also been identified by statisticians: commercially available statistical packages are often used in ways that misuse or even violate the fundamental assumptions of statistical procedures [11]. The problem is that these packages, although powerful in the sense of supporting a huge array of procedures efficiently and reliably, are completely defenseless against inappropriate application, relying solely on the user's statistical familiarity and skill.

To address this problem, substantial work has been occurred at the intersection of Statistics and Artificial Intelligence (AI) [11]. The application of Expert-System (ES) technology to this domain seems to be a natural and mature problem-solving approach for making available expertise that is locally missing. Unfortunately, the systems produced so far aim at defining statistical "strategies" that will *optimize* the analyses performed. As a result the complexity of the resulting systems has permitted only narrowly focused development efforts [11,12]. Additionally, it seems that the more complex multivariate statistical techniques are very difficult to operationalize algorithmically (although promising efforts have been made in that direction [13]).

In our approach to the problem, we followed a different line of reasoning, which is based on the following principles: First, it is much more important to avoid major errors (and thus false scientific inferences) than to attempt an optimum use of statistical analyses. Second, by having a clear picture of the intended user of our system (a statistically naive, but otherwise competent, medical researcher) we can produce effective explanation techniques. Third, we decided to use multivariate strategies based on belief network learning and inference

0195-4210/92/$5.00 © 1994 AMIA, Inc.

techniques [4,14]. Belief networks (BNs) are graph-based representations of probabilistic dependencies and independences among a set of variables, created in order to provide a consise and precise means for handling uncertainty, and thus to improve the efficiency and soundness of both knowledge acquisition and inference in probabilistic reasoning [23]. A large number of algorithms exist for doing inference with BNs [14,15]. Traditionally, creating a BN is a task performed by human experts, but recently methods for automatically creating BNs from data have been developed. In particular, a promising method is the Bayesian learning of belief networks (BLN) [4], which, given a set of variables and a data set, creates a BN by searching heuristically through the space of all possible BNs, and scoring each candidate BN using a Bayesian scoring function. We believe that this design selection not only provides a uniform representation and explanation basis across many statistical modelling problems, but also provides a useful predictive/diagnostic tool at the same time [14,15].

Fourth, we realized that our system can not stand as a "statistical consultant", since the knowledge and the cognitive abilities of a professional statistician are far beyond the current state of the art in AI. Therefore we directed our efforts at a system that is relatively open, highly modular, extensible and necessarily simple, so that it could be used as a testbed for ideas and solutions. Fifth, the system should be flexible and able to function in a resource- limited environment, since various reasoning and computational components (software performing certain tasks, interfaces to various programs, knowledge about the statistical procedures and the available packages, etc.) may or may not be available to the system at a given time.

In summary our goal has been to use the design principles described above, to assist a statistically naive user in identifying statistically significant relationships (as well as spurious ones), explaining why certain statistical procedures were preferred over others, interpreting the results, and handling resource limitations effectively, such as statistical package incompleteness and/or unavailability, and user ignorance of the actual details of use of the data analytic software employed.

METHODS

1.System Design

Figure 1 presents the conceptual design of the system. The user asks a high-level question. The possible high-level questions include: "Are these two variables statistically associated?", and "Is this statistical association confounded by some other variable?". The user interface transfers this request to the *core system* that incorporates all the reasoning capabilities of Data Explorer. The core then utilizes an array of external data-analytic software modules to analyze the data. Table 1 presents the external modules that we have either implemented interfaces for, or written specifically for the system.

After the external modules have carried out the analyses, the core system interprets their output and explains the results to the user. The core system views the external modules as resources that may be available at a given time or setting, and makes every effort to use the available resources appropriately.

Figure 1. : The conceptual design of Data Explorer

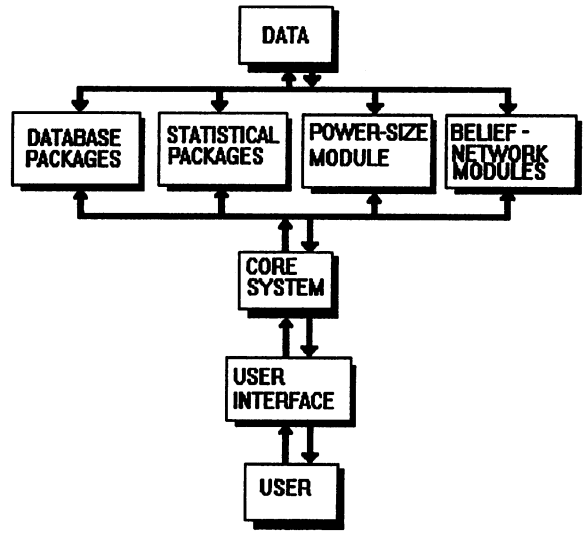

Table 1. : External modules available to the system.

A. Custom-developed modules
1. Power-Size Analysis
2. Belief Network Inference
3. Belief Network Learning
B. Commercially available modules
1. SPSS PC+
(base, statistics and advanced statistics)

The core system consists of the following components: the rule interpreter, a rule base, a statistical packages interface (currently only SPSS [16] is used by the system), a belief-network interface, a power-size module interface, results interpretation, and test selection explanation routines.

The system was developed using the Gold Hill Common Lisp Developer for Windows, Turbo Pascal, and Visual Basic for windows, under MS Windows /MS DOS . It runs on a 486/33 Mhz IBM compatible with 8 MBs of RAM.

2. The Inference Engine

Previous experience with an early prototype written in PROLOG verified published results [11] that the rule-based approach is well suited to the task

at hand, in terms of efficiency and expressiveness.

The goal-directed model in particular, is supported by a) the correspondence between the human reasoning in this domain as often described in textbooks and the classrooms and backward rule activation, b) the nature of the search space (there are few prespecified requests and many possible solutions [17]).

Although PROLOG [18] provided an expressive and efficient development language, we decided to build the system using LISP [19] in order to gain maximum flexibility in procedural interfacing to other software and in explanation. We retained the advantages of PROLOG, by writing a logic-programming based inference engine (in LISP) that provided us with the main PROLOG functionality, as well as procedural attachments used either as predicates or as functions. This arrangement, also allowed us to attach a "daemon" function that updates the explanation module on the rule interpreter, and to write code that reads/interprets/explains the packages output in an efficient and easy way.

3.Knowledge-Base Development

To develop the knowledge base (KB) we: (a) developed the rules for selecting among candidate statistical tests using a three-step procedure. First, high-level user requests were identified. Table 2 presents the user requests supported by Data Explorer. Second, a number of useful tests were selected, based on frequency of use in the literature [19]. Each test was abstracted on paper in such a way that a package and implementation independent representation of the test was produced. This part of the KB development was carried out by one of the authors (E.C.) who served as the domain (i.e. statistics) expert. Third, the test abstract descriptions were used to guide the construction of the rules.

(b) identified and wrote necessary facts (i.e. list of candidate tests to be examined, package support of tests etc.), in first-order logic predicate form, in the corresponding modules (for instance the fact that a power-size analysis module is available to the system, is placed in the power-size interface portion of the core system).

Table 2. : High-level user requests

A. Univariate Statistics :
1. Are two variables statistically associated ?
2. Is variable *a* dependent on variable *b* ?
B. Multivariate Statistics :
1. Is a given variable relationship confounded by another variable ?
2. Build a belief network model from the data, using a set of variables.
3. Predict the values of a set of variables *S1*, given the values of a set of variables *S2*.

4. Explanation

Our explanation goals were to: a) explain *why* the system made certain actions or choices (for example why a given test was applied and another was not), and b) explain *what* the results of the performed actions mean. Our explanation facilities include the following 4 components:

a) *Analyses-tailored explanation*: The system reads the output of the data analysis programs and interprets relationships as significant or not, applicable to the research data or not, confounded or not.

b) *General Reminder*: After the results have been presented, a standard text explaining basic concepts about statistics is printed on screen. The goal is to help the user avoid some very frequent interpretation pitfalls. Currently this component is small and in future work it could be replaced by a more elaborate help system.

c) *Power-size Analysis* : Power-size analysis is an important but often neglected aspect of the design and analysis of research [21,22]. The size analysis addresses the problem of how many observational units should be included in the study, given factors that include the desired levels of statistical accuracy, population characteristics, and degrees of clinical significance. A size smaller than the one required runs the risk of not proving a true relationship, while a size larger than necessary leads to excessive expense and puts subjects at unwarranted risks. On the other hand, power analysis answers the important question of what is the probability that a non-statistically significant relationship is true. The lack of proper power analysis has led some biomedical researchers to view non-significant results as neutral (acceptance of neither the null hypothesis, nor its negation), thus rendering useless studies that in many cases required many resources to be carried out.

Data Explorer uses power-size analysis as follows: if there appears to be a statistically non-significant result, the system calls the Power-Size Module (PSM) asking for the power of the test. If the power level is high, it is concluded that non-significance is probably valid. Otherwise, the PSM module is called again to provide an estimate of the required sample size for a desired (higher) power level. Currently the system uses default values of 0.05 for the desired significance and 0.90 for the desired power levels [9,20].

d) *Actions Explanation*: This part owes much to the explanation design of the ROUNDSMAN expert system [2], and is modelled after the same principles of multiattribute utility theory. More specifically, a combination of 3 techniques is used:

i) *(Modified) lexicographic ordering (LO)*. LO is modified, in the sense that no explicit order is specified, but it is directly derived from the rule ordering in the rule base. The 3 ordering

criteria are (in order of importance) : applicability (whether the test is permitted to be applied), availability (whether the test is available to the system via a package), and optimality (power and robustness of the test).

ii) *Satisficing*. This technique is used within the contexts of applicability and availability. That is, we try to attribute failure of a test to be chosen, either to not being appropriate for the data, or to not being supported by the available statistical software. Such attribution is achieved by maintaining a partial history of failed predicates and a database of "canned" explanation phrases corresponding to each rule predicate. The history is partial since it focuses on the first two levels of the rule hierarchy, thus explaining the top level conditions for the tests selection.

iii) *Dominance*. This refers to the effort preservation principle that makes the system stop trying when the consideration of further tests is guaranteed to be fruitless. Thus if the system does not consider some tests, this is because they are guaranteed to provide no more information than the ones that have been already selected.

The application of this explanation techniques, results in the splitting of the available tests in three groups: the selected test, tests that were more optimal (i.e. more appropriate for the task), and tests that were less optimal. The system then justifies why the more optimal tests failed to be selected (i.e. why they were not applicable or available) and why no more tests are considered.

5. Multivariate Modelling

Data Explorer currently utilizes BLN and BN inference as a means for providing the high-level functionality of table 2, without resorting to standard multivariate techniques (such as ANOVA, Logistic Regression,Multiple Linear Regression, Discriminant Function Analysis, Log-Linear models), which, as explained earlier, are very difficult to automate. Due to space constraints, we will not present here the full theoretical and implementational details of this particular part of the system. For the purposes of this introductory description of the system, we will just outline the three main functions that are provided, and are BN-based:

a) *Identification of confounded relationships*: By building a BN of the variables of interest and examining it, the system can determine whether a relationship established by univariate statistics is spurious or genuine.

b) *Search for the most probable BN given the data*: The output is used for understanding complex probabilistic relationships among variables.

c) *Prediction of the values of a set of variables given evidence on another set of variables*: this function can be used for both prognostic and diagnostic purposes.

EXAMPLE OF USE

Table 3 presents an example of using the system with data from an actual study, in which the role of calcitonin levels in hypogonadic males has been investigated. A number of hypogonadic males have had serum calcitonin levels determined before and after a 6-month therapy with testosterone. In this particular example, the user is interested to know whether calcitonin levels are statistically different before and after the therapy. The system performs a t-test to determine if the observed difference is statistically significant. It explains the findings, and uses power-size analysis to interpret them and to guide the user further. It also justifies why other alternatives were not used.

Table 3: An Example of using Data Explorer

Data Explorer vers. 0.5 Univ. of Pittsburgh 1992

RESULTS EXPLANATION FOR T-TEST :
The value of CALCITONIN was:
22.44 for the first, and 28.85 for the second group.
The variances of the two groups were not found to be significantly different. Therefore the pooled variance estimate was used and it was found that the difference is not statistically significant.
Press any key to get test selection explanations

The following tests were relevant to the request:
Unpaired t-test
Paired t-test
. . .
Kruscall-Walis one way non parametric anova
Although the following tests:
Paired t-test
One-way analysis of variance
were more optimal than the selected test, they were
not used due to the following reasons:
Paired t-test
==> The observations were not paired
One way analysis of variance
==> The system could not construct the appropriate batch files
Unpaired t-test was selected and applied.
The following tests:
Unpaired x-square
. . .
Kendall's W-test
were equally or less optimal than the selected test,

so there was no need to consider applying them.
Size-power module called. Press any key
We can not conclude that there is no true difference, since the power of the test was low (<0.5). Therefore to investigate the possibility of the difference being true, you are advised to increase your sample size to 107 per group.
This will give you a power of 0.9 .
The following are useful reminders for using the above results by Data Explorer most effectively:
[*the general reminder is presented here*]

DISCUSSION

Compared to similar work of applying AI techniques to statistical analysis, Data Explorer represents an effort to avoid major errors caused by the abuse of relatively simple statistical procedures, as opposed to pursuing optimal application of those techniques [11,12]. This enabled us to cover a much broader spectrum of univariate statistics, which according to published results account for 90% of the statistical procedures applied in major clinical journals [20]. Also, the selection of BN inference and learning, as our current multivariate model, offers the advantages of a uniform representation across predictive, diagnostic and discovery problems, ease of automation and explanation, as well as distribution-free applicability. Finally, by applying a series of explanatory techniques based on a well-defined intended user audience, we are able to provide justifications of the system actions and choices.

On the other hand, the current version of Data Explorer is a demonstration prototype that shows the feasibility of our initial goal. We still need to develop the system further and test it more extensively in real world situations. We are currently working towards this end, by enhancing the Knowledge Base and implementing a graphical user interface.

ACKNOWLEDGEMENTS

We thank Rich Thomason, Doug Metzler, and Paul Munro for useful comments and suggestions at the early design stages of the system. Lambros Papandreou provided the patient data of the example. Efi Kokkotou made useful comments on this paper. Support for this research was provided in part by the National Science Foundation under grant IRI 9111590.

Reference

1) E. Shortliffe and L. Perreault (eds.): Medical Informatics: Computer applications in Health Care. Addison-Wesley 1990.
2) G.Rennels: A computational model of reasoning from the clinical literature, Ph.D. dissertation 1986.
3) R.Blum: Discovery and representation of causal relationships from a large time-oriented clinical database: the RX project., Ph.D. dissertation 1981
4) G.Cooper, E. Herskovits: A Bayesian method for the induction of probabilistic networks from data, Machine Learning, 1992; 9: 309-347.
5) D.Eddy, V.Hasselblat, R.Sachter: An introduction to a Bayesian method for meta-analysis: the confidence profile method, Medical Decision Making 1990;10: 15-23.
6) H. Lehmann: A Bayesian computer-based approach to the physician's use of the clinical research literature, Ph.D. dissertation, 1991.
7) H. Jimison: A representation for gaining insight into clinical decision models. In: Proceedings of the 12th Annual SCAMC, 1988, p.110.
8) L. Moses, T. Louis: Statistical consulting in clinical research: the two-way street. Stat Med 1984; 3: 1-5.
9) T. Colton: Statistics in Medicine, Little, Brown 1974.
10) S. Gore, G. Jones, S. Thomson: The Lancet's statistical review process: areas for improvement by authors. Lancet 1992; 340: 100-102.
11) W. Gale (ed.): Artificial Intelligence & Statistics, Addison-Wesley 1986.
12) W. Gale, D. Pregibon : Artificial Intelligence Research in Statistics. AI Magazine, 1985;72-75.
13) A. Afifi, V. Clark : Computer-Aided Multivariate Analysis (2nd ed.) Van Nostrand Reinhold 1990.
14) M. Henrion: An introduction to algorithms for inference in belief networks. In: Uncertainty in Artificial Intelligence 5, M. Henrion and R. Shachter eds. 129-138, 1990, Amsterdam:North Holland.
15) E. Charniak: "Bayesian networks without tears". AI Magazine 1991;12 (4): 50-63.
16) SPSS PC+ vers 4.0 Base manual, Statistics manual, Advanced statistics manual. SPSS Inc 1992.
17) E. Rich, K. Knight: Artificial Intelligence, 2nd ed, Mc Graw-Hill, 1991.
18) W. Clocksin, C. Mellish: Programming in Prolog, Springer-Verlag 1984.
19) G. Steele: Common LISP the language, 2nd ed, Digital Press 1990.
20) J. Bailar, F. Mosteller (eds.): Medical Uses of Statistics (2nd ed.). NEJM books 1992.
21) J. Cohen: Statistical power analysis for the behavioral sciences, Lawrence Erlbaum & Associates, 1988.
22) J. Freiman, T. Chalmers, H. Smith Jr, R. Kuebler : The importance of beta, the type II error and sample sizein the design and interpretation of the randomized controled trial. N Eng J Med 1978;299: 690-4
23) J. Pearl: Probabilistic reasoning in intelligent systems, Morgan- Kaufmann 1988.