The Knowledge Engineering Review, Vol. 11:4, 1996, 365–371

Intelligent data analysis: issues and challenges

XIAOHUI LIU

Department of Computer Science, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK. Email: hui@dcs.bbk.ac.uk

1 Introduction

Two phenomena have probably affected modern data analysts' lives more than anything else. First, the size of real-world data sets is getting increasingly large, especially during the last decade or so. Second, modern computational methods and tools are being developed which add further capability to traditional statistical analysis tools. These two developments have created a new range of problems and challenges for analysts, as well as new opportunities for intelligent systems in data analysis.

To provide an international forum for the discussion of these topics, I chaired a symposium on Intelligent Data Analysis (IDA–95) as part of the annual conference of the International Institute for Advanced Studies in Systems Research and Cybernetics, held 16–20 August 1995 in Baden-Baden, Germany. About 60 people from 20 countries on four continents took part in the symposium. There were plenty of informal and fruitful interactions between presenters and participants. We were also able to have single-track oral and poster presentations. Each poster was introduced by its author in a brief talk during a special plenary session.

2 Background

The job of a data analyst typically involves problem formulation, advice on data collection (it is not uncommon for the analyst to be asked to analyse data which have already been collected), effective data analysis, and interpretation and reporting of the findings (Chatfield, 1988). The data analysis part has often been described as an iterative process in which "exploratory analysis" and "confirmatory analysis" are the two principal components.

Exploratory data analysis (Tukey, 1977), or data exploration, resembles a detective's job closely: understanding evidence collected, looking for clues, applying relevant background knowledge and pursuing and checking the possibilities that clues suggest. This initial phase of data examination typically involves data quality checking (errors, outliers, missing values), data modification (variable transformation, error correction), data summary (summary statistics, tables, visualisations), data dimensionality reduction (principal component analysis, correspondence analysis), use of other multivariate data-analytic techniques (multidimensional scaling, clustering), and informal use of inferential methods (multiple regression, significance tests). Data exploration is not only useful for data understanding, but also helpful in generating possibly interesting hypotheses for a later study–normally a more formal or "confirmatory" procedure for analysing the data. Such a procedure often assumes a potential model structure, and may involve estimating the model parameters and testing hypotheses about the model. The fitted model needs to be evaluated by checking the residuals from the model to see if it needs to be modified or refined.

For the last decade or so, we have witnessed two phenomena which may have affected what modern data analysts do more than anything else. First, the problem of "data explosion" has become increasingly apparent. Many analysts no longer have the luxury of focusing on problems with a manageable number of variables (say a dozen) and cases (typically several hundred), and problems involving hundreds of variables and millions of cases are not uncommon. Data are

collected at an astonishing speed in almost every sector of the modern society, from supermarket transactions to telephone calls, from electronic medical records to data generated by scientific instruments. The World Wide Web is probably the most notable medium for storing huge quantities of data (text, sound, images, etc.) generated by millions of people around the world. There is no doubt that the sheer quantity of data collected is currently far more than our capabilities of making good sense of them.

On the more positive side, data analysts, armed with traditional statistical packages for data exploration, model building and hypothesis testing, are now provided with further analysis capability. For example, On-Line Analytic Processing (OLAP) tools have been developed to provide the interactive capability of analysing multi-dimensional data stored in data warehouses. Data mining tools have been developed to provide further data analysis capabilities. Techniques used in these tools include Bayesian belief networks, case-based reasoning, fuzzy and rough sets, genetic algorithms and genetic programming, heuristic search, knowledge-based systems, machine learning, neural networks and visualisation.

The above two developments have created a new range of problems and challenges for the analysts. First, although we begin to see work on the use of approximation methods, efficient algorithms, parallel processing and high performance computing in the context of analysing very large data sets, much more needs to be done to enable the effective implementation of such applications. Research issues associated with such data sets include developing effective ways of managing and visualising these data, checking data quality, summarising them into convenient and relevant forms for analysis, sampling them with minimum amount of bias, intelligent search for potentially useful structures, detecting anomalous and peculiar patterns and avoiding missing interesting ones (Hand, 1996; Elder IV & Pregibon, 1996).

Second, data analysis is a complex process in which exploratory analysis and confirmatory analysis may be carried out iteratively. At any stage of the analysis process, there is often a large set of possible operations that could be performed and what to do next often depends on the results obtained so far, the problem-solving context, data characteristics and the analyst's strategy. This decision is complicated further by the availability of not only statistical methods and tools, but their counterparts in AI, databases and visualisation, many of which perform similar functions. In this context, it is of particular interest to develop intelligent assistants which will be able to help human analysts make these difficult decisions. They should also be able to perform complex and laborious operations using their computational power so that the analysts can focus on the more creative part of the data analysis using knowledge and experience. Relevant issues include how to divide up work between human and computer; how to ensure that the computer and human stay "in synch" as they work on parts of a data analysis problem; how to seamlessly integrate human domain and common sense knowledge to inform otherwise stupid search procedures such as stepwise regression; how to present data so human eyes can see patterns; how to develop an integrated data analysis environment where the analyst can use relevant methods and tools from different paradigms without too much trouble; and how to help the analyst choose the "best" one from a set of competing methods.

In short, IDA–95 was motivated by the need to address the problems and challenges faced by data analysts, brought about by the analysis of large data sets and by the availability of a variety of data analysis methods.

3 Major themes of presentations

Work reported at the symposium included a variety of research topics on the theory and application of various techniques to data analysis problems. The principal topics covered include classification, clustering, data preprocessing, data visualisation, exploratory data analysis, integrated systems, knowledge discovery, methodological issues and structure learning. Computational techniques used include statistics, decision trees, fuzzy logic, genetic algorithms, knowledge-based systems, neural

networks, pattern recognition, planning, probabilistic reasoning, rough sets, search and user interface.

3.1 Data exploration

One of the major themes of the conference was data exploration. Many authors addressed specific issues related to this topic. Fazel Famili (National Research Council, Canada) emphasised the importance of pre-processing data before confirmatory data analysis can be most effectively carried out. Real-world examples in the manufacturing industry were used to demonstrate his points. He also chaired a panel discussion on this topic (see below). Geoff Holmes and C Nevill-Manning, and Tony Smith and Geoff Holmes (University of Waikato, New Zealand) proposed two different methods for addressing the problem of feature selection and reported preliminary experimental results. One is based on Holte's 1R algorithm and the other is on the notion of "feature relevance". WinViz developed by Hing-Yan Lee et al. (ITI, Singapore) is a system providing the graphical display of multidimensional data using the concept of "parallel coordinates". WinViz has been used to support data exploration in different applications. David McSherry and Sally McClean (University of Ulster, UK) considered the problem of identifying the likeliest value of an unknown, or inconsistently represented, attribute of an entity in a distributed database environment. Results obtained using a probabilistic model of diagnostic reasoning were presented. The problem of clustering was addressed in different contexts (E Bauman and A Dorofeyuk, Institute of Control Science, Russia; Frank Klawonn and Rudolf Kruse, University of Braunschweig, Germany; Frank Reine, Carl Schenck AG, Germany; van den Eijkel *et al.*, Delft University of Technology, Holland).

Worth particular mention in the context of data exploration was a system developed by Rob St. Amant and Paul Cohen from the University of Massachusetts at Amherst, USA. AIDE (Assistant for Intelligent Data Exploration) is an interactive system which cooperates with the analyst to accomplish the data analysis task. AIDE encodes and applies experienced analysts' knowledge of plans, tactics, and statistical strategy, while the analyst provides relevant knowledge in the problem-solving context and steers the analysis process. This mixed initiative takes advantage of the strengths of the computer and the analyst and sheds interesting light on the question of how the analyst may be helped during the iterative process of data analysis.

Although it is sometime undervalued for its "common sense", "ad hoc" or "empirical" approach, data exploration lies in the very heart of detecting interesting patterns or structures from data, and is more demanding than many classical confirmatory procedures which have become very easy to perform with a computer (Chatfield, 1988). Therefore, it is particularly pleasing to see that many pieces of work have been presented on various issues related to this topic. However, it would be nice to see more of this kind of work in future IDA conferences, especially for dealing with large data sets.

3.2 IDA applications

Another major theme was the diversity of practical IDA applications, ranging from engineering to medicine, and from service industry to image processing. Evangelos Simoudis *et al.* (IBM, USA) described the development of customer vulnerability models for frozen orange juice and reported the results of applying the models to a database containing the supermarket purchases of 15,000 households over a three-year period. Several generic conclusions were made, which may be helpful for developing similar applications, regarding the effect of conceptual clustering and symbolic induction algorithms. Hector Garcia of Danish Hydraulic Institute described the development of a system using intelligent agents and numerical algorithms for the real-time control of urban drainage networks. The system has been applied to one of such networks in the city of Gothenburg, Sweden, with a surface area of approximately 152 square kilometres and serves a population of 973,000. Interesting results were also reported on the restoration of distorted sound patterns (Andrzej

Czyzewski, Technical University of Gdansk, Poland), the classification of instrument sounds (Bozena Kostek, Technical University of Gdansk, Poland), and the search for the optimal path for autonomous vehicles (Osamu Ono and Buhei Kobayashi, Meiji University, Japan).

In medicine, Silvia Miksch and Johannes Gartner (Austrian Research Institute for Artificial Intelligence) presented a "fact-finding" process which integrates data validation, data interpretation and task visualisation into a framework. This process was applied to two different medical domains: artificial ventilation of newborn infants and shift-scheduling. Alberto Riva and Riccardo Bellazzi (University of Pavia, Italy) applied a set of techniques, including probabilistic learning and statistical analysis, to data from home-monitoring of patients affected by Insulin Dependent Diabetes Mellitus. They found that data pre-processing is essential and the use of domain knowledge crucial in selecting features and appropriate analysis methods.

To support IDA applications in different areas, Integral Solutions Limited of UK developed a system called Clementine. One of the main strengths of Clementine is that the system has been developed with the needs of non-technologist end-users in mind. The details of the techniques (neural networks, rule induction, etc.) are hidden from these users, and interfaces between the system and users are intuitive and easy to learn (e.g., visual programming). Colin Shearer described many of these interesting features in his presentation.

The practical applications reported at the symposium provided a good snapshot of the range of application domains where intelligent data analysis could lead to important benefits. Admittedly, there are many obstacles for applying IDA techniques to real world problems; these include the lack of efficient tools for analysing very large data sets, and lack of truly integrated, effective, and friendly data analysis environments. However, important initial steps have been taken. A survey of issues in developing knowledge discovery applications in industrial settings can be found in (Piatetsky-Shapiro *et al.*, 1996).

3.3 General principles

There were a number of interesting pieces of work which were of a more theoretical or technical nature. Paul Cohen and Tim Oates (University of Massachusetts at Amherst, USA) presented four algorithms for finding different types of structure in streams. These algorithms are rather general and they have been utilised in a variety of application areas. Alois Heinz (Albert-Lugwigs-Universitat, Germany) proposed Adaptive Fuzzy Neural Trees as an appropriate tool for intelligent data analysis. His experiments demonstrate that this hybrid approach has advantages over single approaches such as decision trees, fuzzy logic and neural networks. Michael Berthold and Klaus-Peter Huber (Universitat Karlsruhe, Germany) propose a method for building a special type of neural network using Rectangular Basis Functions that allow the direct extraction of rules from the network. It is found that the networks can be constructed very rapidly without any need for user-interaction.

Klaus-Peter Huber and Michael Berthold also conducted a comparative study between two learning algorithms for generating rules from data. Many such comparative studies (e.g. Michie *et al.*, 1994), can be found in the literature, with some reporting very little difference between methods, while others report sharp contrasts. This raised an issue which concerns many of the analysts working in practical problem solving: if there are many methods which appear to be applicable to the problem or sub-problem under consideration, which one should be used? One approach would be that one tries every method and chooses the best. The assumption behind this approach is that there are sufficient resources available for the experiments, but this is not always realistic in practice. Could we anticipate the results of a particular comparative study so that an "optimal" or near-optimal method can be chosen without trying all of the possibilities? A deep understanding of data characteristics, properties of individual methods and problem solving contexts, and certainly results from those comparative studies which have already been carried out, should be able to help here.

Huw Roberts *et al.* (BT Laboratories, UK) addressed the problem of integrating misclassification costs into a tree induction algorithm. This is in response to the practical need that predictive or

classification accuracy is not always the only, or even the most important, criterion for evaluating learning methods. Credit scoring is one of the most quoted applications where misclassification cost is more important than the predictive accuracy (Nakhaeizadeh, 1995). Other important factors in deciding that one method is preferable to another include the computational efficiency and the understandability/ interpretability of the methods (Brodley & Smyth, 1996; Weiss & Kulikowski, 1991). A comprehensive study of issues and methods in assessing the performance of AI programs can be found in Cohen (1995).

There are many general issues concerning the effective development and application of intelligent data analysis techniques as discussed in sections 2 and 4. At this point it is perhaps worth mentioning a dilemma faced by data analysts. We are witnessing the continuing development of new and increasingly sophisticated techniques both in the statistics and computing communities. On the one hand, they may be regarded as additional methods for effectively analysing data. However, they may also lead to confusion for practitioners and inappropriate use of these techniques may be common. Despite the development of ever more complex and innovative techniques, it seems to be desirable to clarify the general principles needed to effectively apply the techniques we already have. For example, are these techniques competitive or complementary? Which techniques are particularly effective on which aspects of data analysis?

3.4 Panel on data preprocessing

Fazel Famili (NRC, Canada) chaired a panel on "The Importance of Data Pre-processing in Intelligent Data Analysis", and the panelists included Wen-Ling Hsu (AT&T, USA), Wei-Min Shen (University of Southern California, USA), Evangelos Simoudis (IBM, USA) and Richard Weber (MIT, Germany). The main objectives of this panel were to emphasise the importance of data preprocessing and to discuss some of the important issues in dealing with real world data.

Data pre-processing was described as all the actions taken for the purpose of: (i) solving problems that exist with the raw data (e.g., corrupt or out of range data), (ii) understanding the nature of the raw data for efficient use of a data analysis tool (e.g., use of principal component analysis and data visualisation, elimination of irrelevant data), and (iii) data transformation/conversion and the introduction of new features for extracting more accurate knowledge from a given set of data (e.g., linear/nonlinear transformation, scale levels and simulation of new attribute vectors).

The following questions were raised. What are useful data pre-processing techniques? How do we know which technique to apply in a given situation? In which applications will data pre-processing be useful and in which might it lead to loss or change of useful knowledge? Is it possible to develop a tool box (advisor) for data pre-processing? Should the user (e.g., domain expert) be involved/ consulted in all stages of data pre-processing? How can we incorporate our experiences into new tools that are developed? How could the Intelligent Data Analysis community benefit from the research done on data pre-processing?

A number of real world applications including aerospace, retail and telecommunications were used during the discussion of the above issues. The general consensus was that it is possible to develop data pre-processing tools to be customised and used in different applications. It was also felt that a domain expert should be involved in the data pre-processing process, as inappropriate data preprocessing may result in loss of information. Answers to many other questions are frequently dependent on domain-specific characteristics.

Although there is a growing acceptance that data pre-processing is often crucial to the success in the development of solutions, this is not always reflected in the literature. Take neural networks as an example. An overwhelming majority of publications in the area report on the development of learning algorithms, network performance assessment and related work. Most of the papers spend little space on data pre-processing issues. A panel on this topic would serve the purpose of creating interests in a systematic study of related issues and seeing how it can be effectively integrated into the "mainstream" data analysis.

4 Concluding remarks

Problems arising from effective analysis of large data sets have made the data analyst's job more challenging than ever. Although data analysts now have access to a variety of statistical and AI tools capable of performing different aspects of data analysis, they certainly need further support. For example, they need competent interactive tools for exploring complex real-world data sets; they need powerful graphical techniques for visualising multi-dimensional data; they need new computational tools for controlling data quality; they need effective means of summarising large data sets into convenient and relevant forms for analysis; they need data sampling methods with a minimum amount of bias; they need intelligent search methods which will find the most interesting structures; and they need more integrated and "friendly" data analysis environments where "boring aspects" of the analyst's job may be kept to the minimum so that interesting aspects of the job can be focused on. Of course many of these needs were there before large data sets came into the picture. However, searching for interesting structures in large data sets has called for better ways of doing these things and suggested new areas for research.

Take data quality control as an example. We begin to see "data cleaning" companies being set up to respond to the need to automate (at least partially) the laborious process of managing imperfect data as this often consumes a significant amount of resources. Outliers, or outlying observations, are particularly difficult to handle as some of them are measurement errors, while others may represent something "significant" from the viewpoint of the application domain. In this situation an outright rejection of outliers based on some statistical tests is normally not a very good idea. So a careful analysis of outliers by data analysts is normally desirable. However, if there are many of them (this is often the case when dealing with large data sets), manual analysis may become insufficient. Some preliminary work has been performed to explore the possibility of automating certain aspects of outlier analysis (Liu *et al.*, 1994).

Apart from the need to research into the problems and opportunities brought about by the analysis of large data sets, it is of strategic importance to obtain a deep understanding of how data analysts carry out their task and to see how their analysis procedures may be encoded, extended, or improved. The rapid development of statistical and AI tools has made many aspects of data analysis routine, e.g., there is no longer the need to concern ourselves with the mechanics of how to find a particular structure in a data set (Hand, 1996). Instead we can concentrate on considering the "high-level" issues such as what kind of structures should be sought, what questions should we be asking, what would be the most appropriate method of analysis, and how the results should be interpreted. We can find out how the analyst's knowledge and strategies can be most effectively captured in IDA tools and applications. We can study the strengths and weaknesses of numerous computational techniques and their potential contributions to the various stages of the data analysis process. We can see how these techniques may be most appropriately used and how we may assist data analysts in their quest to uncover interesting and useful structures from large data sets.

In IDA-95, the Intelligent Data Analysis community demonstrated great enthusiasm and resourcefulness in responding to the above challenges by presenting many interesting pieces of research. For example, much interesting work on data exploration was reported; many useful real world IDA applications were described; some fundamental research problems were addressed; and we also begin to see some papers (e.g., St Amant and Cohen), but not many, on developing intelligent assistants for data analysts. These papers were concerned with human-machine data analysis, and asked: what kind of intelligence is needed to make this type of analysis work effectively?

Although IDA-95 was successful in bringing together people with different backgrounds to discuss important issues in intelligent data analysis, there is much to be desired. First, some people presented essentially black-box techniques. The boxes might be intelligent, but it was unclear how these boxes may contribute to the overall data analysis process. Second, there were not many papers on the effective integration of AI and statistical techniques to approach data analysis tasks. For example, papers on statistical analysis tend to have little AI content, whereas AI approaches tend to

talk about machine learning and knowledge discovery without referring much to specific data analysis issues. Third, we need more work addressing issues arising from analysing very large data sets. Last, but not least, IDA–95 was less focused than it should have been. This problem might be addressed in subsequent IDA symposia because the potential areas for IDA are very large and trying to address every possible issue in any given symposium could lead to the possibility of only scratching surfaces and not making serious progress.

A survey after IDA–95 supported the idea of making IDA a regular, biennial conference. IDA–97, the second in the series, will be focusing on "Reasoning about Data". We are interested in intelligent systems that reason about how to analyse data, perhaps as human analysts do. Analysts often bring exogenous knowledge about data to bear when they decide how to analyse it; they use intermediate results to decide how to proceed; they reason about how much analysis the data will actually support; they consider which methods will be most informative; they decide which aspects of a model are most uncertain and focus attention there; they sometimes have the luxury of collecting more data, and plan to do so efficiently. In short, there is a strategic aspect to data analysis, beyond the tactical choice of this or that test, visualisation or variable.

IDA–97 will be held at Birkbeck College, University of London, August 4–6 1997. Details can be found on the World Wide Web via URL: http://web.dcs.bbk.ac.uk/ida97.html. The IDA–95 proceedings was published by the International Institute for Advanced Studies and Cybernetics (IIAS Press 1995; ISBN 0–921836–29–5).

Acknowledgement

IDA–95 was a joint effort of many people and I cannot possibly acknowledge all of them here by names. My deep gratitudes go to all the authors, panelists, and participants in IDA–95; to the Program Committee and auxiliary reviewers for their great help; to Fazel Famili for organising the data pre-processing panel. I would also like to thank Gongxian Cheng, John Wu and other members of the Intelligent Data Analysis Group at Birkbeck for their assistance in organising the event, and my management at Birkbeck for their encouragement and support for IDA–95, and IDA–97 to be held at Birkbeck. Finally my appreciations go to Paul Cohen, Trevor Fenner and Evangelos Simoudis for their detailed, valuable comments on early drafts of this paper, and to Fazel Famili for providing much of the materials regarding the panel activities in section 3.

References

Brodley, CE and Smyth, P, 1996. "Applying classification algorithms in practice" *Statistics and Computing* (in press).

Chatfield, C, 1988. *Problem Solving: a Statistician's Guide* Chapman & Hall.

Cohen, P, 1995. *Empirical Methods for Artificial Intelligence* MIT Press.

Elder IV, J and Pregibon, D, 1996. "A statistical perspective on knowledge discovery in databases" In UM Fayyad, G Piatetsky-Shapiro, P Smyth and R Uthurusamy (eds), *Advances in Knowledge Discovery and Data Mining* AAAI/MIT Press.

Hand, DJ, 1996. "Intelligent data analysis and deep understanding" *Proc. Intelligent Data Management 96* Unicom, London, pp. 26–39.

Liu, X, Cheng, G and Wu, J, 1994. "Noise and uncertainty management in intelligent data modeling" *Proc. AAAI-94* Seattle, WA, pp. 263–268.

Michie, D, Spiegelhalter, DJ and Taylor, CC, (eds) 1994. *Machine Learning, Neural and Statistical Classification* Ellis Horwood.

Nakhaeizadeh, G, 1995. "What Daimler-Benz has learned as an industrial partner from the machine learning project StatLog?" *Proc. Workshop on Applying Machine Learning in Practice* 22–26.

Piatetsky-Shapiro, G, Brachman, R, Khabaza, T, Kloesgen, W and Simoudis, E, 1996. "An overview of issues in developing industrial data mining and knowledge discovery applications" In E Simoudis, J Han and U Fayyad, (eds) *Proc. Second International Conference on Knowledge Discovery and Data Mining* AAAI Press.

Tukey, JW, 1977. *Exploratory Data Analysis* Addison-Wesley.

Weiss, SM and Kulikowski, CA, 1991. *Computer Systems that Learn* Morgan Kaufmann.