ELSEVIER

Intelligent Data Analysis 2 (1998) 63–66

IDA
INTELLIGENT DATA ANALYSIS

www.elsevier.com/locate/ida

Editorial: Reasoning About Data

Xiaohui Liu [a,1], Paul Cohen [b,2], Michael Berthold [c,3]

[a] *Department of Computer Science, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom*
[b] *Principal Investigator, Experimental Knowledge Systems Laboratory, Department of Computer Science, Box 34610, University of Massachusetts, Amherst, MA 01003-4610, USA*
[c] *Computer Science Division, University of California, Soda Hall room 329, Berkeley, CA 94720, USA*

1. Introduction

Two factors have affected the work of modern data analysts more than any others. First, the size of machine-readable data sets has increased, especially during the last decade or so. Second, computational methods and tools are being developed that enhance traditional statistical analysis. 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, a series of symposia on Intelligent Data Analysis was initiated in 1995 [4]. The second Intelligent Data Analysis conference (IDA-97) was held at Birkbeck College, University of London, 4th–6th August 1997. Almost 130 people from twenty countries in four continents took part in the symposium. A total of 107 papers were submitted to the IDA-97 conference, of which 50 were accepted as either oral or poster presentations. After the conference, five papers were chosen from the conference program and their authors were invited to prepare extended versions for publication in the *Intelligent Data Analysis* (IDA) Journal. A second round of review provided additional feedback to the authors and the papers are now presented in this special issue.

2. Background

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

[1] E-mail: hui@dcs.bbk.ac.uk.
[2] E-mail: cohen@titanic.cs.umass.edu.
[3] E-mail: berthold@cs.berkeley.edu.

PII: S108 8-467X(98)00021-3

capable of performing different aspects of data analysis, they need further support. For example, they need competent interactive tools for exploring complex real-world data sets, powerful graphical techniques for visualizing multi-dimensional data, new computational tools for controlling data quality, effective means of summarizing large data sets into convenient and relevant forms for analysis, intelligent search methods which will find the most interesting structures, and 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. The rapid development of statistical and AI tools has made many aspects of data analysis routine; for example, there is no longer the need to concern ourselves with the mechanics of how to find a particular structure in a data set. Instead we can focus on 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 also study how the analyst's knowledge and strategies can be most effectively captured in IDA tools and applications, the strengths and weaknesses of numerous computational techniques and their potential contributions to the various stages of the data analysis process, 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.

These issues were raised in [4] with a suggestion that subsequent IDA symposia might hopefully address them, discussing and fostering research about how to analyze data, perhaps as human analysts do. Analysts often bring exogenous knowledge about data to bear when they decide how to analyze 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, visualization or variable. As a result, "reasoning about data" was chosen to be the guiding theme of the 1997 symposium.IDA-97. Many papers presented at IDA-97 complement the major theme "reasoning about data," but other, exciting topics have emerged, including exploratory data analysis [8], data quality [7] and knowledge discovery [2], as well as the perennial technologies of classification [3,6] and soft computing [9]. A new and interesting theme involves analyzing time series data from physical systems, such as medical instruments, industrial processes or the environment. Here are several highlights of the conference. "General Issues": We were very fortunate to have Professor David J. Hand of the Open University, UK and Dr. Larry Hunter of the National Library of Medicine, USA, to give two exciting, complementary keynote addresses to the conference. Professor Hand started the symposium with an excellent overview of the various issues and opportunities for intelligent data analysis, while Dr. Hunter presented some interesting and challenging IDA applications from molecular biology. Some success stories (e.g., in gene finding) were told, while other applications present considerable difficulties for current analysis tools. "Data Exploration": A wide range of issues were explored, ranging from the collaboration between computer and analyst to pre-partioning training data with different characteristics for optimal classification, and from a "Scientists' Empirical Assistant" to the design of a "Data Analysis Game" that allows the implementation of a learning framework suitable for exploring different analysis problems. "Medical applications": Medical applications have inspired many methods and developments in AI, and they have also provided interesting platforms for developing IDA systems and applications. Many papers presented at the conference used medicine as the underlying application area.These included the use of time-warping for segmentation to find different sub-patterns in an ECG signal, a temporal abstraction approach to analyzing longitudinal data from diabetic patients, an annotated data collection system to support the analysis of intensive care data, and the use of a data exploration system to support modeling in epidemiological studies. "Qualitative models": Several excellent papers that were presented at the conference were concerned with the qualitative abstraction of quantitative information for analysis or reasoning purpose. These include the use of state transition diagrams as a way to model discrete event sequences, the generation of textual descriptions from time-varying data, the detection of changes in time series data, and the use of geometric reasoning to infer qualitative information from quantitative data. "Data Quality": Data is now viewed as a key organizational resource and the use of high-quality data for decision making has received increasing attention. There are many dimensions of data quality: accuracy, completeness, consistency, timeliness etc. and much work has been done in the IDA community on the management of noisy, missing and outlying data. This is reflected by many papers presented at the conference, which included the management of missing data by building dynamic decision paths and using exogenous knowledge in Bayesian Belief Networks, and the analysis of outlying data using relevant domain knowledge. We have selected one paper from each of the above topics for this special issue, which we shall introduce below. However, there were many other interesting papers presented at IDA-97 that, unfortunately, cannot be accommodated here. These include papers related to topics such as data mining, soft computing, estimation and clustering. The IDA-97 proceedings were published in Springer's Lecture Notes in Computer Science Series [5].

3. Contents of this Issue

The first paper by David J. Hand, "Intelligent Data Analysis: Issues and Opportunities", discusses the interdisciplinary nature of data analysis, issues arising from the analysis of large, complex data,and opportunities for intelligent data analysis. Several manifestations of intelligent (and unintelligent) data analysis are illustrated. Finally a cautionary note is given that new analysis techniques should be rooted in real-world problems and that developing techniques in the abstract should be discouraged. The paper by Apte et al. tackles an important problem in classification: examples in a classification domain may correspond to groups that are very different in nature and may require different learning strategies. The authors propose a pre-processing phase where the examples are split into separate subsets before learning actually takes place. The split is based on feature merit measures of different kinds, leading to a measure called an importance profile angle (IPA). A DNA classification problem is addressed using the proposed method. The paper by Bellazzi et al. uses the concept of Temporal Abstraction to transform longitudinal data into a new time series containing more meaningful information, whose features are then interpreted using statistical and probabilistic techniques. The methods are applied to the analysis of diabetic patients' data, and the use of the temporal abstraction framework for handling meta-data is also discussed. The paper by Bradley and Easley discusses the role of an intelligent data analyzer in a qualitative modeling tool called PRET. Geometric reasoning techniques are used to infer qualitative information from quantitative data and the capabilities of the analyzer are demonstrated using a real-world modeling example. Current methods to learn Bayesian Networks from incomplete databases share the assumption that the unreported data are missing at random. Ramoni and Sebastiani propose a method called "Bound and Collapse" which moves away from this assumption. Exogenous knowledge about the pattern of missing data is used instead. The proposed method has a good intuitive feel, but the authors go further, relating it mathematically to other methods, comparing performances etc.

4. Concluding Remarks

Many interesting pieces of work were presented at IDA-97 and only a snapshot is provided in this special issue. Nevertheless we observe much room for improvement. For example, there are still cases reporting essentially black-box algorithms and there are still analysis techniques developed without real-world problems in mind. Reported cases of analyzing very large data sets are not as many as we would like to see. Evaluations of a lot of IDA systems and applications are still not as thorough as they should be; related issues and methods may be found in [1]. We need to obtain a deeper understanding of the processes and principles involved in data analysis and to gain further experience in analyzing large, real-world messy data. In the light of the great success of IDA-97 we are looking forward to IDA-99 which will be held in August 1999 hosted by Professor Joost Kok's group in Amsterdam. We are delighted that Professor David J. Hand has agreed to serve as General Chair for that symposium.

Acknowledgments

We would like to thank all the authors for working with us on a very tight schedule. We are grateful to the following referees who helped tremendously with the review despite our short notice: Francesco Bergadano, Pavel Brazdil, Sylvain Letourneau, David Lubinsky, Bradley Whitehall, Gerhard Widmer and Janusz Wnek. In addition we would like to thank all the authors, speakers, reviewers and organizers of IDA-97 who made this issue possible. Last but not least, we thank Fazel Famili for all his help in the process of editing this issue.

References

[1] Cohen, P., *Empirical Methods for Artificial Intelligence*, MIT, Cambridge, Massachusetts, 1995.

[2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT, 1996.

[3] Hand, D.J., *Construction and Assessment of Classification Rules*, Wiley, 1997.

[4] Liu, X., Intelligent data analysis: Issues and challenges, *The Knowledge Engineering Review*, 11(4), 365–371, 1996.

[5] Liu, X., Cohen, P., Berthold, M. (eds), *Advances in Intelligent Data Analysis: Reasoning about Data*, Lecture Notes in Computer Science 1280, Springer-Verlag, 1997.

[6] Michie, D., Spiegelhalter, D. J., and Taylor, C.C. (eds), *Machine Learning, Neural and Statistical Classification*, London: Ellis Horwood, 1994.

[7] Tayi, G. and Ballou, (eds), *Examining Data Quality*, Communications of the ACM, special issue, 41:2, 1998.

[8] Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

[9] Zadeh, L.A., *Soft Computing and Fuzzy Logic*, IEEE Software, 48–56, 1994.