

Intelligent Data Analysis

Reasoning about Data

Michael Berthold, Paul Cohen, and Xiaohui Liu

The growing importance of the automatic or semiautomatic analysis of data sets in many real-world applications has led to the emergence of the field of intelligent data analysis (IDA), a combination of diverse disciplines including AI and statistics in particular. These fields complement each other: Many statistical methods, particularly those for large data sets, rely on computation, but brute computing power is no substitute for statistical knowledge. Thus, we are seeing the development of intelligent systems for data analysis.

To provide an international forum for the discussion of these topics, a series of symposia on IDA was started in 1995 (Liu 1996). In 1997, the Second International Symposium on Intelligent Data Analysis (IDA97) was held at Birkbeck College, University of London, on 4 to 6 August. Almost 130 people from 20 countries in 4 continents attended. The final program consisted of 2 invited talks and 49 reviewed presentations chosen from 107 submitted papers. The symposium was organized as a single track of oral and poster presentations to give the participants the opportunity to discuss all the research, leading to many informal and fruitful interactions between presenters and participants. Each poster was introduced by its author in a brief talk during special plenary sessions.

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.

At the First International Symposium on Intelligent Data Analysis

(IDA95), it was concluded that there is a need for research in the areas of mixed-initiative, IDA tools (Liu 1996). Although data analysts now have access to a variety of statistical algorithms, these tend to be unintelligent black boxes. Because data sets are too large today to be investigated manually, data analysis tools must themselves determine areas of interest, directions in which to guide the search, and try to relieve the user of the boring aspects of analysis. Another theme at IDA95 was the need to integrate different techniques, sometimes from diverse disciplines. Many of the techniques at the first symposium were "component technology," and little thought was given to how these methods could cooperate in an IDA archi-

The Second International Symposium on Intelligent Data Analysis (IDA97) was held at Birkbeck College, University of London, on 4 to 6 August 1997. The main theme of IDA97 was to reason about how to analyze data, perhaps as human analysts do, by exploiting many methods from diverse disciplines. This article outlines several key issues and challenges, discusses how they were addressed at the conference, and presents opportunities for further work in the field.

tecture or framework. Also, most techniques were demonstrated on relatively small applications; there was little discussion of very large data sets. Thus, the theme of IDA97 was set: reasoning about data and how to analyze it, particularly large amounts of data, perhaps as humans analyze it, by exploiting many methods.

Major Themes of Presentation

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 exploratory data analysis, pre-processing, and tools; classification and feature selection; soft computing; knowledge discovery and data mining; estimation; clustering; and qualitative models. Two entire sessions were devoted to medical applications and data quality.

David Hand of The Open University, United Kingdom, started the symposium with an exciting survey of the issues and opportunities for IDA. His paper serves as an insightful assessment of a field that, although too young and exuberant to know exactly what it is about, clearly has great potential. In addition to promoting the interdisciplinary nature of IDA, Hand made a point of defining what he calls *unintelligent data analysis*, or data analysis that goes too far. To analyze data efficiently and intelligently, skills from a variety of disciplines are required, and it is from real problems that solutions emerge; building abstract methods will not be helpful in the future. Appropriately then, Larry Hunter of the National Library of Medicine presented a challenging new application for the IDA community. The title of his talk sums it up: "Given 3,000,000,000 Nucleotides Induce a Person—or Intelligent Data Analysis in Molecular Biology." Hunter talked about data analysis problems so large, so complex, and so important that none of us will ever again feel entirely comfortable testing our techniques on safe little data sets.

A major theme of the conference was data exploration. Several contributions addressed various issues related to this topic. Chidanand Apte and his colleagues from the IBM T. J. Watson Research Center, New York, proposed a methodology to discover heterogeneity in classification problems, thus enabling the system to explore different areas of the feature space using different analysis techniques. On the other side of the spectrum, Michael

Muller's (University of Witwatersrand, South Africa) approach to the management of dialogue between a user and an expert system enables the system to identify clusters of different user profiles. How to explore the complexity of the underlying model function was addressed in C. McGeoch, D. Precup, and Paul Cohen's paper. They presented a set of heuristic approaches to obtain complexity bounds based on observed data. One interesting focus of exploratory work is exploring the model behind the model; for example, work on judging the complexity of data or subdividing the data into smaller problems was presented. In the future, we look forward to developments in techniques such as subgroup discovery (Wrobel 1997) and metamodeling (Kleijnen and Van Groenendaal 1996).

Another major topic was classification and feature selection. K. Schaedler and F. Wysotzki (both of TU Berlin, Germany) introduced a way to compare objects with high structural complexity using an approach to determine the similarity of arbitrary labeled graphs. R. Glover and P. Sharpe (both of the University of the West of England, Bristol, United Kingdom) focused on an efficient way to select features; they examined modified fitness functions for a genetic algorithm based on test-set sampling. C. Lam, G. West, and T. Caelli (Curtin University, Perth, Australia) compared two different ways to generate decision trees, whereas S. Lo and A. Famili (University of British Columbia, Vancouver, Canada) described a tool for knowledge-driven constructive induction. Other work related to decision trees was presented by J. Gama (University of Porto, Portugal). He described a variant of *c4.5* with oblique decision surfaces. As the size of data sets, and especially their dimensionality, increases, it will become increasingly important to filter out useless or unwanted data. Approaches that cleverly choose an appropriate subset of features are gaining interest, although it remains to be seen how they will interact with the rest of the analysis process.

Interesting applications were reported in the medical field. H. Vullings, M. Verhaegen, and H. Verbruggen (Delft University, The Nether-

lands) used time warping for segmentation to find different subpatterns in an electrocardiogram signal. An approach to analyze longitudinal data from diabetic patients was presented by R. Bellazzi, C. Larizza, and A. Riva, all from the University of Pavia (Italy). V. Kamp and F. Wietek (Oldenburg, Germany) explained the architecture of a software system to support modeling and conducting of descriptive epidemiologic studies. It is interesting, and consistent with the development of other areas of AI, that so many applications come from the medical side. The recently launched journal entitled *Information Technology in Biomedicine* (published as part of the Institute of Electrical and Electronics Engineers transaction series) indicates that there is a growing need for automatic data analysis in this field. In the light of rapidly inflating costs for monitoring and early detection of diseases, automatization offers vast potential savings and sometimes even more reliable early detection of medical problems.

Soft computing is beginning to have a significant impact on IDA, just as it is starting to acknowledge work in statistics and related disciplines. S. Gunn, M. Brown, and K. Bossley, all from the University of Southampton (United Kingdom), evaluated different measures of significance to determine an optimal structure for a neural network. They used B-spline fuzzy networks to describe data in an interpretable form. A genetic approach to fuzzy clustering with a validity-measure fitness function was presented by S. Nascimento and F. Moura-Pires (both of University Lisboa, Portugal). In contrast to these approaches, which require specialized models, C. Roadknight, D. Palmer-Brown, and G. Mills (all of The Nottingham Trent University, United Kingdom) described a way to analyze classical multilayer perceptrons. First, a way to prune useless neurons in the trained network is applied, and afterward, a method that the authors call *equation synthesis* extracts rules from the network. Other work was also presented, ranging from the application of fuzzy graphs (K.-P. Huber and M. Berthold, both of University of Karlsruhe, Germany) to the

mathematical analysis of fuzzy classifiers (F. Klawonn, Fachhochschule Ostfriesland, Germany, and E. Klement, Johannes Kepler University, Austria). Of more practical focus was the work presented by A. Lapp and H.-G. Kranz (both of GH Wuppertal, Germany), who introduced an automated diagnosis system with a rated diagnosis reliability. It seems as if the area of soft computing concentrates more and more on *interpretable models*, models that offer insights into the underlying process. There is often the issue of trade-off between the understandability-simplicity of a model and its predictive-classification accuracy; Lotfi Zadeh's (1996) concept of "information granulation" appears to shed some light on this issue. Using granules that represent instances of similar origin to an adjustable degree, this framework makes it possible to hide unwanted information, thus enabling better understandability.

The discovery of knowledge in databases is receiving a lot of attention these days, and several such papers were presented at IDA97. D. McSherry (University of Ulster, United Kingdom) proposed a strategy to increase the efficiency of rule discovery in databases by means of decomposing into subtasks and focusing on rules that appear most interesting. U. Hahn and K. Schnattinger from the University of Freiburg (Germany) proposed a knowledge-intensive approach to text analysis that works incrementally to expand the underlying domain knowledge base. A pool of hypotheses is generated through clues, and the most credible ones are then introduced into the knowledge base. An application for chess end-game databases was presented by M. Schlosser (FH Koblenz, Germany).

Two other growing areas of IDA involve building models for estimation and clustering techniques. G. van den Eijkel, J. van der Lubbe, and E. Backer (all of Delft University, The Netherlands) presented an approach to modeling underlying probability density functions. In contrast to the classical Parzen window, they try to limit the required number of prototypes by building an interpolated esti-

The Sixth International Workshop on
Agent Theories, Architectures, and Languages (ATAL-99)
 Orlando, Florida, July 15-17, 1999

<http://www.elec.qmw.ac.uk/dai/atal/>

The ATAL workshop series aims to bring together researchers interested in the theory and practice of intelligent agent technology. Over the past five years, ATAL has come to be recognised as the pre-eminent international forum for publishing results on the agent-level, micro aspects of this technology. Specifically, ATAL-99 will address issues such as theories of agency, agent architectures, methodologies and programming languages for agent-based systems, software tools for applying and evaluating agent systems, and agent-oriented software engineering.

Organising committee:

Nick Jennings University of London, UK N.R.Jennings@qmw.ac.uk	Yves Lespérance York University, Canada lesperan@cs.yorku.ca
---	--

Submission details: Email a PostScript/PDF version or send hardcopy of your paper to Nick Jennings to arrive by **April 5, 1999**. See the WWW page for details.

mate. Several methods to determine confidence intervals for quantiles in finite populations were discussed by M. Rueda Garcia, A. Arcos Cebrian, and E. Artes Rodriguez (all of the University of Granada, Spain). Y. Kharin (State University, Belarus) presented a method to cluster multivariate data under the presence of outliers. This so-called *cluster algorithm with smoothing* works by cleaning up the data before applying a conventional clustering algorithm. An approach to clustering very large data sets was proposed by E. Schikuta and M. Erhart (University of Vienna, Austria). They use a hierarchical algorithm that clusters the input space using a multidimensional grid.

Data analysis techniques must deal with distorted data. Most algorithms require clean data; only a few can handle missing values or outliers. W. Liu, A. White, S. Thompson, and M. Bramer (all of the University of Portsmouth, United Kingdom) presented a technique based on decision trees, but the decision paths are built dynami-

cally; therefore, the algorithm makes use of data that contain missing values that are not used in the current context. Similar work using Bayesian networks was presented by M. Ramoni and P. Sebastiani (both of The Open University, United Kingdom). J. Wu, G. Cheng, and X. Liu (all of the University of London, United Kingdom) discussed a method to find outliers in data based on a self-organizing map. A medical application was used to demonstrate this approach: Rather than trying to remove outliers, it is more beneficial to model them and use the model for analysis. Often it will not be clear at the outset which process is producing noise, and in many applications, these outliers aren't bad data points: They simply model an underlying dependency that, at this stage, the analyst is not interested in. Here again an intelligent, partly interactive, approach could be a way for the user to investigate the entire data and pick focuses of attention.

The last session of the symposium dealt with qualitative models and was kicked off by a stimulating presentation by E. Bradley (University of Colorado). Together with her coauthor M. Easley, she introduced a system called PRET that automates the system identification process by adding an AI layer on top of traditional techniques. Using different modules that are either user specified or automatically generated, it assembles a system of ordinary differential equations that matches the domain physics and also user-specified qualitative and quantitative observations. A. Howe and G. Somlo (both of Colorado State University) introduced state-transition diagrams as a way to model discrete-event sequences. Through the detection of sequences of limited length (snapshots) and their influence on following events, a transition diagram is generated. S. Boyd (Macquarie University, Australia) discussed how textual descriptions of sequences can be generated. B. Schieffer and G. Hotz (both

of University Saarbruecken, Germany) described a way to diagnose hybrid systems. Finally, O. Wolkenhauer (Control Systems Center, UMIST, Manchester, United Kingdom) presented a way to detect changes in time-series data through the use of random sets.

Concluding Remarks

It seems that the interdisciplinary and eclectic nature of IDA97 sets it apart from other conferences. The final program offered not one but many perspectives on IDA, making it a meeting that broadened everyone's horizons. People are working productively at the boundaries between AI and statistics, and several contributions addressed topics, or implemented mixtures of techniques, from both disciplines. Many papers presented essentially black-box algorithms, and it would be nice to see more integrated systems, especially intelligent assistants. Some early work was reported (for example, R. Levinson and S. Wilkinson of the University of California at Santa Cruz, M. Muller of the University of Witwatersrand, South Africa). It is possible that the development of integrated, intelligent assistants is being held back by our choice of problems and applications, which tended not to be very challenging. Thus, it was fortuitous that Hunter's invited lecture offered a great challenge and made clear how important it is to deal with real, large data sets if we want to gain new insights into the process of IDA. In light of the extraordinary success of IDA97 (Liu, Cohen, and Berthold 1997), we are looking forward to IDA99, which will be held in August 1999 and hosted by Joost Kok's group in Amsterdam, The Netherlands. We are delighted that Hand has agreed to serve as general chair for this symposium.

Acknowledgments

IDA97 was the joint effort of many people, and we cannot possibly acknowledge them here by name. Thanks go to all the authors, participants, exhibitors, and invited speakers and to the program committee and

auxiliary reviewers for their excellent reviews. The University of Massachusetts at Amherst coordinated the IDA97 reviewing process, and Birkbeck College at the University of London provided all the necessary financial and local support. Finally, we would like to thank the American Association for Artificial Intelligence, the Association of Computing Machinery Special Interest Group on Artificial Intelligence, the British Computer Society Special Group on Expert Systems, the Institute of Electrical and Electronics Engineers Systems, Men, and Cybernetics, and the Society for the Study of Artificial Intelligence and Simulation of Behavior for collaborating on IDA97.

References

- Kleijnen J., and Van Groenendaal, W. 1996. Regression Metamodels and Design of Experiments. Paper presented at the 1996 Winter Simulation Conference, 8–11 December, Coronado, California.
- Liu, X. 1996. Intelligent Data Analysis: Issues and Challenges. *The Knowledge Engineering Review* 11(4): 365–371.
- Liu, X.; Cohen, P.; and Berthold, M., eds. 1997. *Advances in Intelligent Data Analysis: Reasoning about Data*. Lecture Notes in Computer Science 1280. New York: Springer Verlag.
- Wrobel, S. 1977. An Algorithm for Multirelational Discovery of Subgroups. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, eds. J. Komorowski and J. Zytkow, 78–87. Berlin: Springer Verlag.
- Zadeh, L. A. 1996. Fuzzy Logic and the Calculi of Fuzzy Rules and Fuzzy Graphs. *Multivalued Logic* 1:138.

Michael Berthold (M.Sc., 1992; Ph.D., 1997) was with the University of Karlsruhe from 1993 to 1997 and is currently a BISC postdoctoral fellow at the University of California at Berkeley. He was a visiting researcher at Carnegie Mellon University from 1991 to 1992 and at Sydney University in 1994. He also worked as a research engineer at Intel Corp., Santa Clara, California, in 1993. His current research interests include neural networks, fuzzy logic, and intelligent data analysis. He is a member of the steering committee of the Intelligent Data Analysis conference series and is coeditor, with David Hand, of the forthcoming book *An Introduction to Intelligent Data Analysis* (Springer-Verlag). His e-mail

address is berthold@cs.berkeley.edu.

Paul R. Cohen is a professor in the Department of Computer Science at the University of Massachusetts and director of the Experimental Knowledge Systems Laboratory. He received his Ph.D. from Stanford University in computer science and psychology in 1983 and his M.S. and B.A. in psychology from the University of California at Los Angeles and the University of California at San Diego, respectively. He served on the program committees for the Fifth and Sixth International Workshops on Artificial Intelligence and Statistics and is currently serving on the program committee for the Fourth International Conference on Artificial Intelligence Planning Systems (AIPS-98). He was the program cochairman of the Second International Conference on Intelligent Data Analysis. He recently coedited, with Bruce Porter, a special issue of *Artificial Intelligence* on empirical methods for AI. He served as a councilor of the American Association for Artificial Intelligence (AAAI) from 1991 to 1994 and was elected in 1993 as a fellow of the AAAI.

Xiaohui Liu is a senior lecturer in the Department of Computer Science at Birkbeck College, University of London. His main research interests are in the study of computationally intelligent methods, particularly their application to challenging real-world data analysis problems. He received his Ph.D. in computer science from Heriot-Watt University at Edinburgh in 1988. He was organizer and program chairman of IDA-95 and general chairman of IDA-97 and is on the editorial board for the Evaluation of Intelligent Systems online resource and *Intelligent Data Analysis: An International Journal*.