Foreword

This volume is a collection of papers presented at the Second International Symposium on Intelligent Data Analysis (IDA-97) held at Birkbeck College, University of London, August 4-6, 1997. Our community is growing fast: The first Symposium, held in Baden-Baden in 1995, received 69 extended abstracts for review; this time we received 107 full papers. We accepted 50, divided equally between oral and poster presentations. Each paper was considered by at least two independent reviewers and the results were discussed in a meeting of program committee members.

IDA-97 was a single-track conference consisting of oral and poster presentations, invited speakers, demonstrations, and exhibitions. The conference Call for Papers introduced a theme, “Reasoning About Data”, and many papers complement this theme, but other, exciting topics have emerged, including exploratory data analysis, data quality, knowledge discovery, and data-analysis tools, as well as the perennial technologies of classification and soft computing. A new and exciting theme involves analyzing time series data from physical systems, such as medical instruments, environmental data, and industrial processes.

We are very grateful to Professor David Hand, of the Open University, and Dr. Larry Hunter, of the National Library of Medicine in Washington D.C., for agreeing to give keynote talks. Each has made major contributions and has helped to define the interdisciplinary field of intelligent data analysis.

The symposium owes much to many hard-working individuals. Michael Berthold handled publicity and publications, and is responsible for this volume. The program committee and additional volunteer reviewers read the papers with great care and wrote excellent, informative, helpful reviews. Terrie Korpita and Peggy Weston at the University of Massachusetts coordinated the international reviewing process, a huge job performed flawlessly. The organization at Birkbeck has been tremendous. Phil Gregg was responsible for the IDA-97 homepage, which served to keep the community abreast of developments and was responsible in no small part for the increasing size of the symposium. Betty Walters, Sylvie Jami, Trevor Fenner, Claude Gierl, and Steven Swift were responsible for daily operations, enquiries, registrations, and the details that grew exponentially as the conference date drew nearer. The Systems group in Computer Science at Birkbeck College provided valuable computing support for the reviewing process and conference exhibitions. The symposium would not exist were it not for the confidence and financial backing of Birkbeck College, particularly Roger Johnson, Ken Thomas, and Guy Fitzgerald.

June 1997

Xiaohui Liu
Paul Cohen
Organization Committee

General Chair:  Xiaohui Liu  
Department of Computer Science  
Birkbeck College, Malet Street  
London WC1E 7HX, UK  
E-mail: hui@cs.bbk.ac.uk

Technical Program:  Paul Cohen  
Department of Computer Science  
University of Massachusetts, Amherst  
Amherst, MA 01003-4610, USA  
E-mail: cohen@cs.umass.edu

Publicity&Publication:  Michael Berthold  
Universität Karlsruhe, IRF  
Am Zirkel 2, 76128 Karlsruhe, Germany  
E-mail: berthold@ira.uka.de

Local Arrangements:  Trevor Fenner  
Department of Computer Science  
Birkbeck College, Malet Street  
London WC1E 7HX, UK  
E-mail: trevor@dcs.bbk.ac.uk

Finance:  Sylvie Jami  
Department of Computer Science  
Birkbeck College, Malet Street  
London WC1E 7HX, UK  
E-mail: s.jami@dcs.bbk.ac.uk

Sponsorship:  Mihaela Ulieru  
Simon Fraser University  
2357 Riverside Drive  
North Vancouver, B.C.  
Canada V7H 1V8  
Email: ulieru@cs.sfu.ca

Exhibition:  Richard Weber  
MIT GmbH, Promenade 9  
52076 Aachen, Germany  
E-mail: rw@mitgmbh.de
Program Committee

Eric Backer  Delft University of Technology, The Netherlands
Riccardo Bellazzi  University of Pavia, Italy
Michael Berthold  University of Karlsruhe, Germany
Carla Brodley  Purdue University, USA
Gongxian Cheng  Birkbeck College, UK
David Dowe  Monash University, Australia
Fazel Famili  National Research Council, Canada
Julian Faraway  University of Michigan, USA
Thomas Feuring  WWU Münster, Germany
Alex Gammerman  Royal Holloway London, UK
David J Hand  Open University, UK
Rainer Holve  Forwiss Erlangen, Germany
Wenling Hsu  AT&T Consumer Lab, USA
Larry Hunter  National Library of Medicine, USA
David Jensen  University of Massachusetts, USA
Frank Klawonn  University of Braunschweig, Germany
David Lubinsky  University of Witwatersrand, South Africa
Ramon Lopez de Mantaras  Artificial Intelligence Research Institute, Spain
Sylvia Miksch  Stanford University, USA
Rob Milne  Intelligent Applications Ltd, UK
Gholamreza Nakhaeizadeh  Daimler-Benz Forschung und Technik, Germany
Claire Nedellec  Université Paris-Sud, France
Erkki Oja  Helsinki University of Technology, Finland
Henri Prade  University Paul Sabatier, France
Daryl Pregibon  AT&T Research, USA
Peter Ross  University of Edinburgh, UK
Steven Roth  Carnegie Mellon University, USA
Lorenza Saitta  University of Torino, Italy
Peter Selfridge  AT&T Research, USA
Rosaria Silipo  University of Florence, Italy
Evangelos Simoudis  IBM Almaden Research, USA
Derek Sleeman  University of Aberdeen, UK
Paul Snow  Consultant, USA
Rob St. Amant  North Carolina State University, USA
Lionel Tarassenko  Oxford University, UK
John Taylor  King's College London, UK
Loren Terveen  AT&T Research, USA
Hans-Jürgen Zimmermann  RWTH Aachen, Germany

# Table of Contents

Intelligent Data Analysis: Issues and Opportunities ..................... 1  
David J. Hand (UK)

## Section I: Exploratory Data Analysis, Preprocessing,  
and Tools

- Decomposition of Heterogeneous Classification Problems ............ 17  
  C. Apte, S. J. Hong, J. R. M. Hosking, J. Lepre, E. P. D. Pednault  
  and B. K. Rosen (USA)
- Managing Dialogue in a Statistical Expert Assistant with a Cluster-Based User Model ........................................... 29  
  M. W. Muller (South Africa)
- How to Find Big-Oh in Your Data Set (and How Not to) .............. 41  
  C. C. McGeoch, D. Precup and P. R. Cohen (USA)
- Data Classification Using a W.I.S.E. Toolbox ......................... 53  
  I. Berry and P. Gough (UK)
- Mill’s Methods for Complete Intelligent Data Analysis .............. 65  
  T. A. O. Cornish (UK)
- Integrating Many Techniques for Discovering Structure in Data .... 77  
  D. E. Gregory and P. R. Cohen (USA)
- Meta-Reasoning for Data Analysis Tool Allocation .................... 89  
  R. Levinson and J. Wilkinson (USA)
- Navigation for Data Analysis Systems ................................ 101  
  R. St. Amant (USA)
- An Annotated Data Collection System to Support Intelligent Analysis of Intensive Care Unit Data ................................. 111  
  C. L. Tsien and J. C. Fackler (USA)
- A Combined Approach to Uncertain Data Analysis .................. 123  
  H. Yu and A. Ramer (Australia)

## Section II: Classification and Feature Selection

- A Connectionist Approach to the Distance-Based Analysis of Relational Data ..................................................... 137  
  K. Schädler and F. Wysotzki (Germany)
Efficient GA Based Techniques for Automating the Design of Classification Models ..................................................... 149
R. Glover and P. Sharpe (UK)

Data Representations and Machine Learning Techniques ................... 161
C. P. Lam, G. A. W. West and T. M. Caelli (Australia)

Development of a Knowledge-Driven Constructive Induction Mechanism ... 173
S. Lo and A. Famili (Canada)

Oblique Linear Tree ...................................................... 187
J. Gama (Portugal)

Feature Selection for Neural Networks through Functional Links
Found by Evolutionary Computation ........................................ 199
S. Haring, J. N. Kok and M. C. van Wezel (The Netherlands)

Building Simple Models: A Case Study with Decision Trees ............... 211
D. Jensen, T. Oates and P. R. Cohen (USA)

Exploiting Symbolic Learning in Visual Inspection ........................ 223
M. Piccardi, R. Cucchiara, M. Bariani and P. Mello (Italy)

Forming Categories in Exploratory Data Analysis and Data Mining ........ 235
P. D. Scott, R. J. Williams and K. M. Ho (UK)

A Systematic Description of Greedy Optimisation Algorithms for
Cost Sensitive Generalisation .............................................. 247
M. van Someren, C. Torres and F. Verdenius (The Netherlands)

Dissimilarity Measure for Collections of Objects and Values .............. 259
P. Valtchev and J. Euzenat (France)

Section III: Medical Applications

ECG Segmentation Using Time-Warping .................................. 275
H. J. L. M. Vullings, M. H. G. Verhaegen and H. B. Verbruggen
(The Netherlands)

Interpreting Longitudinal Data through Temporal Abstractions:
An Application to Diabetic Patients Monitoring .......................... 287
R. Bellazzi, C. Larizza and A. Riva (Italy)

Intelligent Support for Multidimensional Data Analysis in
Environmental Epidemiology .............................................. 299
V. Kamp and F. Wietek (Germany)

Section IV: Soft Computing

Network Performance Assessment for Neurofuzzy Data Modelling ........ 313
S. R. Gunn, M. Brown and K. M. Bossley (UK)
A Genetic Approach to Fuzzy Clustering with a Validity Measure
Fitness Function .......................................................... 325
S. Nascimento and F. Moura-Pires (Portugal)

The Analysis of Artificial Neural Network Data Models .......... 337
C. M. Roadknight, D. Palmer-Brown and G. E. Mills (UK)

Simulation Data Analysis Using Fuzzy Graphs .................. 347
K.-P. Huber and M. R. Berthold (Germany)

Mathematical Analysis of Fuzzy Classifiers ...................... 359
F. Klawonn and E.-P. Klement (Germany; Austria)

Neuro-Fuzzy Diagnosis System with a Rated Diagnosis Reliability
and Visual Data Analysis .............................................. 371
A. Lapp and H.-G. Kranz (Germany)

Genetic Fuzzy Clustering by Means of Discovering Membership
Functions ............................................................... 383
M. Turhan (Turkey)

Section V: Knowledge Discovery and Data Mining
A Strategy for Increasing the Efficiency of Rule Discovery in
Data Mining ............................................................ 397
D. McSherry (UK)

Intelligent Text Analysis for Dynamically Maintaining and
Updating Domain Knowledge Bases .................................. 409
K. Schnattinger and U. Hahn (Germany)

Knowledge Discovery in Endgame Databases .................... 423
M. Schlosser (Germany)

Parallel Induction Algorithms for Data Mining .................. 437
J. Darlington, Y.-K. Guo, J. Sutiwaraphun and H. W. To (UK)

Data Analysis for Query Processing ............................... 447
J. Robinson and B. G. T. Lowden (UK)

Datum Discovery ....................................................... 459
L. Siklóssy and M. Ayel (France)

A Connectionist Approach to Extracting Knowledge from Databases ... 465
Y. Zhou, Y. Lu and C. Shi (China)
Section VI: Estimation, Clustering
A Modulated Parzen-Windows Approach for Probability Density Estimation ............................................................... 479
G. C. van den Eijkel, J. C. A. van der Lubbe and E. Backer
(The Netherlands)

Improvement on Estimating Quantiles in Finite Population
Using Indirect Methods of Estimation ..................................... 491
M. R. García, A. A. Cebrián and E. A. Rodríguez (Spain)

Robustness of Clustering under Outliers ................................. 501
Y. Kharin (Belarus)

The BANG-Clustering System: Grid-Based Data Analysis ............... 513
E. Schikuta and M. Erhart (Austria)

Section VII: Data Quality
Techniques for Dealing with Missing Values in Classification .......... 527
W. Z. Liu, A. P. White, S. G. Thompson and M. A. Bramer (UK)

The Use of Exogenous Knowledge to Learn Bayesian Networks
from Incomplete Databases ........................................... 537
M. Ramoni and P. Sebastiani (UK)

Reasoning about Outliers by Modelling Noisy Data ........................ 549
J. X. Wu, G. Cheng and X. Liu (UK)

Section VIII: Qualitative Models
Reasoning about Sensor Data for Automated System Identification ...... 561
E. Bradley and M. Easley (USA)

Modelling Discrete Event Sequences as State Transition Diagrams ...... 573
A. E. Howe and G. Somlo (USA)

Detecting and Describing Patterns in Time-Varying Data
Using Wavelets .................................................................. 585
S. Boyd (Australia)

Diagnosis of Tank Ballast Systems .......................................... 597
B. Schieffer and G. Hotz (Germany)

Qualitative Uncertainty Models from Random Set Theory ............... 609
O. Wolkenhauer (UK)

Author Index ...................................................................... 621