

# Lecture Notes in Computer Science

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

2189

**Springer**

*Berlin*

*Heidelberg*

*New York*

*Barcelona*

*Hong Kong*

*London*

*Milan*

*Paris*

*Tokyo*

Frank Hoffmann David J. Hand Niall Adams  
Douglas Fisher Gabriela Guimaraes (Eds.)

# Advances in Intelligent Data Analysis

4th International Conference, IDA 2001  
Cascais, Portugal, September 13-15, 2001  
Proceedings



Springer

## Volume Editors

Frank Hoffmann

Royal Institute of Technology, Centre for Autonomous Systems  
10044 Stockholm, Sweden  
E-mail: hoffmann@nada.kth.se

David J. Hand

Niall Adams

Imperial College, Huxley Building  
180 Queen's Gate, London SW7 2BZ, UK  
E-mail: {d.j.hand,n.adams}@ic.ac.uk

Douglas Fisher

Vanderbilt University, Department of Computer Science  
Box 1679, Station B, Nashville, TN 37235, USA  
E-mail: dfisher@vuse.vanderbilt.edu

Gabriela Guimaraes

New University of Lisbon, Department of Computer Science  
2825-114 Caparica, Portugal  
E-mail: gg@di.fct.unl.pt

## Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Advances in intelligent data analysis : 4th international conference ;  
proceedings / IDA 2001, Cascais, Portugal, September 13 - 15, 2001. Frank  
Hoffmann ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ;  
London ; Milan ; Paris ; Tokyo : Springer, 2001  
(Lecture notes in computer science ; Vol. 2189)  
ISBN 3-540-42581-0

CR Subject Classification (1998): H.3, I.2, G.3, I.5.1, I.4.5, J.2, J.1, J.3

ISSN 0302-9743

ISBN 3-540-42581-0 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York  
a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2001  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign  
Printed on acid-free paper SPIN: 10840583 06/3142 5 4 3 2 1 0

# Preface

These are the proceedings of the fourth biennial conference in the *Intelligent Data Analysis* series. The conference took place in Cascais, Portugal, 13–15 September 2001. The theme of this conference series is the use of computers in intelligent ways in data analysis, including the exploration of intelligent programs for data analysis. Data analytic tools continue to develop, driven by the computer revolution. Methods which would have required unimaginable amounts of computing power, and which would have taken years to reach a conclusion, can now be applied with ease and virtually instantly. Such methods are being developed by a variety of intellectual communities, including statistics, artificial intelligence, neural networks, machine learning, data mining, and interactive dynamic data visualization. This conference series seeks to bring together researchers studying the use of intelligent data analysis in these various disciplines, to stimulate interaction so that each discipline may learn from the others. So as to encourage such interaction, we deliberately kept the conference to a single track meeting. This meant that, of the almost 150 submissions we received, we were able to select only 23 for oral presentation and 16 for poster presentation. In addition to these contributed papers, there was a keynote address from Daryl Pregibon, invited presentations from Katharina Morik, Rolf Backhofen, and Sunil Rao, and a special ‘data challenge’ session, where researchers described their attempts to analyse a challenging data set provided by Paul Cohen. This acceptance rate enabled us to ensure a high quality conference, while also permitting us to provide good coverage of the various topics subsumed within the general heading of intelligent data analysis.

We would like to express our thanks and appreciation to everyone involved in the organization of the meeting and the selection of the papers. It is the behind-the-scenes efforts which ensure the smooth running and success of any conference. We would also like to express our gratitude to the sponsors: Fundação para a Ciência e a Tecnologia, Ministério da Ciência e da Tecnologia, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Fundação Calouste Gulbenkian and IPE Investimentos e Participações Empresariais, S.A.

September 2001

Frank Hoffmann  
David J. Hand  
Niall Adams  
Gabriela Guimaraes  
Doug Fisher

# Organization

IDA 2001 was organized by the department of Computer Science, New University of Lisbon.

## Conference Committee

General Chair:	Douglas Fisher (Vanderbilt University, USA)
Program Chairs:	David J. Hand (Imperial College, UK) Niall Adams (Imperial College, UK)
Conference Chair:	Gabriela Guimaraes (New University of Lisbon, Portugal)
Publicity Chair:	Frank Höppner (Univ. of Appl. Sciences Emden, Germany)
Publication Chair:	Frank Hoffmann (Royal Institute of Technology, Sweden)
Local Chair:	Fernando Moura-Pires (University of Evora, Portugal)
Area Chairs:	Roberta Siciliano (University of Naples, Italy) Arno Siebes (CWI, The Netherlands) Pavel Brazdil (University of Porto, Portugal)

## Program Committee

Niall Adams (Imperial College, UK)  
Pieter Adriaans (Syllogic, The Netherlands)  
Russell Almond (Educational Testing Service, USA)  
Thomas Bäck (Informatik Centrum Dortmund, Germany)  
Riccardo Bellazzi (University of Pavia, Italy)  
Michael Berthold (Tripos, USA)  
Liu Bing (National University of Singapore)  
Paul Cohen (University of Massachusetts, USA)  
Paul Darius (Leuven University, Belgium)  
Fazel Famili (National Research Council, Canada)  
Douglas Fisher (Vanderbilt University, USA)  
Karl Froeschl (University of Vienna, Austria)  
Alex Gammernan (Royal Holloway, UK)  
Adolf Grauel (University of Paderborn, Germany)  
Gabriela Guimaraes (New University of Lisbon, Portugal)  
Lawrence O. Hall (University of South Florida, USA)  
Frank Hoffmann (Royal Institute of Technology, Sweden)  
Adele Howe (Colorado State University, USA)  
Klaus-Peter Huber (SAS Institute, Germany)  
David Jensen (University of Massachusetts, USA)  
Joost Kok (Leiden University, The Netherlands)  
Rudolf Kruse (University of Magdeburg, Germany)  
Frank Klawonn (University of Applied Sciences Emden, Germany)

## VIII Organization

Hans Lenz (Free University of Berlin, Germany)

David Madigan (Soliloquy, USA)

Rainer Malaka (European Media Laboratory, Germany)

Heikki Mannila (Nokia, Finland)

Fernando Moura Pires (University of Evora, Portugal)

Susana Nascimento (University of Lisbon, Portugal)

Wayne Oldford (University of Waterloo, Canada)

Albert Prat (Technical University of Catalunya, Spain)

Peter Protzel (Technical University Chemnitz, Germany)

Giacomo della Riccia (University of Udine, Italy)

Rosanna Schiavo (University of Venice, Italy)

Kaisa Sere (Abo Akademi University, Finland)

Roberta Siciliano (University of Naples, Italy)

Rosaria Silipo (Nuance, USA)

Floor Verdenius (ATO-DLO, The Netherlands)

Stefan Wrobel (University of Magdeburg, Germany)

Hui XiaoLiu (Brunel University, UK)

Nevin Zhang (Hong Kong University of Science and Technology, Hong Kong)

## Sponsoring Institutions

Fundação para a Ciência e a Tecnologia, Ministério da Ciência e da Tecnologia

Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Fundação Calouste Gulbenkian

IPE Investimentos e Participações Empresariais, S.A.

# Table of Contents

## The Fourth International Symposium on Intelligent Data Analysis

Feature Characterization in Scientific Datasets . . . . .	1
<i>Elizabeth Bradley (University of Colorado), Nancy Collins (University of Colorado), W. Philip Kegelmeyer (Sandia National Laboratories)</i>	
Relevance Feedback in the Bayesian Network Retrieval Model: An Approach Based on Term Instantiation . . . . .	13
<i>Luis M. de Campos (University of Granada), Juan M. Fernández-Luna (University of Jaén), Juan F. Huete (University of Granada)</i>	
Generating Fuzzy Summaries from Fuzzy Multidimensional Databases . . . .	24
<i>Anne Laurent (Université Pierre et Marie Curie)</i>	
A Mixture-of-Experts Framework for Learning from Imbalanced Data Sets . . . . .	34
<i>Andrew Estabrooks (IBM), Nathalie Japkowicz (University of Ottawa)</i>	
Predicting Time-Varying Functions with Local Models . . . . .	44
<i>Achim Lewandowski (Chemnitz University), Peter Protzel (Chemnitz University)</i>	
Building Models of Ecological Dynamics Using HMM Based Temporal Data Clustering – A Preliminary Study . . . . .	53
<i>Cen Li (Tennessee State University), Gautam Biswas (Vanderbilt University), Mike Dale (Griffith University), Pat Dale (Griffith University)</i>	
Tagging with Small Training Corpora . . . . .	63
<i>Nuno C. Marques (Universidade Aberta), Gabriel Pereira Lopes (Centria)</i>	
A Search Engine for Morphologically Complex Languages . . . . .	73
<i>Udo Hahn (Universität Freiburg), Martin Honeck (Universitätsklinikum Freiburg), Stefan Schulz (Universitätsklinikum Freiburg)</i>	
Errors Detection and Correction in Large Scale Data Collecting . . . . .	84
<i>Renato Bruni (Università di Roma), Antonio Sassano (Università di Roma)</i>	



A New Framework to Assess Association Rules .....	95
<i>Fernando Berzal (University of Granada), Ignacio Blanco (University of Granada), Daniel Sánchez (University of Granada), María-Amparo Vila (University of Granada)</i>	
Communities of Interest .....	105
<i>Corinna Cortes (AT&amp;T Shannon Research Labs), Daryl Pregibon (AT&amp;T Shannon Research Labs), Chris Volinsky (AT&amp;T Shannon Research Labs)</i>	
An Evaluation of Grading Classifiers .....	115
<i>Alexander K. Seewald (Austrian Research Institute for Artificial Intelligence), Johannes Fürnkranz (Austrian Research Institute for Artificial Intelligence)</i>	
Finding Informative Rules in Interval Sequences .....	125
<i>Frank Höppner (University of Applied Sciences Emden), Frank Klawonn (University of Applied Sciences Emden)</i>	
Correlation-Based and Contextual Merit-Based Ensemble Feature Selection .....	135
<i>Seppo Puuronen (University of Jyväskylä), Alexey Tsymbal (University of Jyväskylä), Iryna Skrypnyk (University of Jyväskylä)</i>	
Nonmetric Multidimensional Scaling with Neural Networks .....	145
<i>Michiel C. van Wezel (Universiteit Leiden), Walter A. Kosters (Universiteit Leiden), Peter van der Putten (Universiteit Leiden), Joost N. Kok (Universiteit Leiden)</i>	
Functional Trees for Regression .....	156
<i>João Gama (University of Porto)</i>	
Data Mining with Products of Trees .....	167
<i>José Tomé A.S. Ferreira (Imperial College), David G.T. Denison (Imperial College), David J. Hand (Imperial College)</i>	
$S^3$ Bagging: Fast Classifier Induction Method with Subsampling and Bagging .....	177
<i>Masahiro Terabe (Mitsubishi Research Institute, Inc.), Takashi Washio (I.S.I.R., Osaka University), Hiroshi Motoda (I.S.I.R., Osaka University)</i>	
RNA-Sequence-Structure Properties and Selenocysteine Insertion .....	187
<i>Rolf Backofen (University of Munich)</i>	
An Algorithm for Segmenting Categorical Time Series into Meaningful Episodes .....	198
<i>Paul Cohen (University of Massachusetts), Niall Adams (Imperial College)</i>	

An Empirical Comparison of Pruning Methods for Ensemble Classifiers . . .	208
<i>Terry Windeatt (School of Electronics Engineering Guildford), Gholamreza Ardeshir (School of Electronics Engineering Guildford)</i>	
A Framework for Modelling Short, High-Dimensional Multivariate Time Series: Preliminary Results in Virus Gene Expression Data Analysis . . . . .	218
<i>Paul Kellam (University College London), Xiaohui Liu (Brunel University), Nigel Martin (Birkbeck College), Christine Orengo (University College London), Stephen Swift (Brunel University), Allan Tucker (Brunel University)</i>	
Using Multiattribute Prediction Suffix Graphs for Spanish Part-of-Speech Tagging . . . . .	228
<i>José L. Triviño-Rodríguez (University of Málaga), Rafael Morales-Bueno (University of Málaga)</i>	
Self-Supervised Chinese Word Segmentation . . . . .	238
<i>Fuchun Peng (University of Waterloo), Dale Schuurmans (University of Waterloo)</i>	
Analyzing Data Clusters: A Rough Sets Approach to Extract Cluster-Defining Symbolic Rules . . . . .	248
<i>Syed Sibte Raza Abidi (Universiti Sains Malaysia), Kok Meng Hoe (Universiti Sains Malaysia), Alwyn Goh (Universiti Sains Malaysia)</i>	
Finding Polynomials to Fit Multivariate Data Having Numeric and Nominal Variables . . . . .	258
<i>Ryohei Nakano (Nagoya Institute of Technology), Kazumi Saito (NTT Communication Science Laboratories)</i>	
Fluent Learning: Elucidating the Structure of Episodes . . . . .	268
<i>Paul R. Cohen (University of Massachusetts)</i>	
An Intelligent Decision Support Model for Aviation Weather Forecasting . . .	278
<i>Sérgio Viademonte (Monash University), Frada Burstein (Monash University)</i>	
MAMBO: Discovering Association Rules Based on Conditional Independencies . . . . .	289
<i>Robert Castelo (Utrecht University), Ad Feelders (Utrecht University), Arno Siebes (Utrecht University)</i>	
Model Building for Random Fields . . . . .	299
<i>R.H. Glendinning (Defence Evaluation and Research Agency)</i>	
Active Hidden Markov Models for Information Extraction . . . . .	309
<i>Tobias Scheffer (University of Magdeburg), Christian Decomain (SemanticEdge), Stefan Wrobel (University of Magdeburg)</i>	

Adaptive Lightweight Text Filtering ..... 319  
*Gabriel L. Somlo (Colorado State University), Adele E. Howe (Colorado State University)*

A General Algorithm for Approximate Inference in Multiply Sectioned Bayesian Networks ..... 330  
*Zhang Hongwei (Tsinghua University), Tian Fengzhan (Tsinghua University), Lu Yuchang (Tsinghua University)*

Investigating Temporal Patterns of Fault Behaviour within Large Telephony Networks ..... 340  
*Dave Yearling (BTexact Technologies), David J. Hand (Imperial College)*

Closed Set Based Discovery of Representative Association Rules ..... 350  
*Marzena Kryszkiewicz (Warsaw University of Technology)*

Intelligent Sensor Analysis and Actuator Control ..... 360  
*Matthew Easley (Rockwell Scientific), Elizabeth Bradley (University of Colorado)*

Sampling of Highly Correlated Data for Polynomial Regression and Model Discovery ..... 370  
*Grace W. Rumantir (Monash University), Chris S. Wallace (Monash University)*

**The IDA'01 Robot Data Challenge**

The IDA'01 Robot Data Challenge ..... 378  
*Paul Cohen (University of Massachusetts), Niall Adams (Imperial College), David J. Hand (Imperial College)*

**Author Index** ..... 383