

# Studies in Classification, Data Analysis, and Knowledge Organization

---

## *Managing Editors*

H.-H. Bock, Aachen  
W. Gaul, Karlsruhe  
M. Schader, Mannheim

## *Editorial Board*

F. Bodendorf, Nürnberg  
P.G. Bryant, Denver  
F. Critchley, Milton Keynes  
E. Diday, Paris  
P. Ihm, Marburg  
J. Meulmann, Leiden  
S. Nishisato, Toronto  
N. Ohsumi, Tokyo  
O. Opitz, Augsburg  
F.J. Radermacher, Ulm  
R. Wille, Darmstadt

**Springer-Verlag Berlin Heidelberg GmbH**

## Titles in the Series

- H.-H. Bock and P. Ihm (Eds.)  
Classification, Data Analysis,  
and Knowledge Organization. 1991  
(out of print)
- M. Schader (Ed.)  
Analyzing and Modeling Data  
and Knowledge. 1992
- O. Opitz, B. Lausen, and R. Klar  
(Eds.)  
Information and Classification.  
1993 (out of print)
- H.-H. Bock, W. Lenski,  
and M.M. Richter (Eds.)  
Information Systems and Data  
Analysis. 1994 (out of print)
- E. Diday, Y. Lechevallier, M. Schader,  
P. Bertrand, and B. Burtschy  
(Eds.)  
New Approaches in Classification  
and Data Analysis. 1994  
(out of print)
- W. Gaul and D. Pfeifer (Eds.)  
From Data to Knowledge. 1995
- H.-H. Bock and W. Polasek (Eds.)  
Data Analysis and Information  
Systems. 1996
- E. Diday, Y. Lechevallier  
and O. Opitz (Eds.)  
Ordinal and Symbolic Data  
Analysis. 1996
- R. Klar and O. Opitz (Eds.)  
Classification and Knowledge  
Organization. 1997
- C. Hayashi, N. Ohsumi, K. Yajima,  
Y. Tanaka, H.-H. Bock, and Y. Baba  
(Eds.)  
Data Science, Classification,  
and Related Methods. 1998
- I. Balderjahn, R. Mathar,  
and M. Schader (Eds.)  
Classification, Data Analysis,  
and Data Highways. 1998
- A. Rizzi, M. Vichi, and H.-H. Bock  
(Eds.)  
Advances in Data Science  
and Classification. 1998
- M. Vichi and O. Opitz (Eds.)  
Classification and Data Analysis.  
1999
- W. Gaul and H. Locarek-Junge  
(Eds.)  
Classification in the Information  
Age. 1999
- H.-H. Bock and E. Diday (Eds.)  
Analysis of Symbolic Data. 2000
- H. A. L. Kiers, J.-P. Rasson,  
P. J. F. Groenen, and M. Schader  
(Eds.)  
Data Analysis, Classification,  
and Related Methods. 2000
- W. Gaul, O. Opitz and M. Schader  
(Eds.)  
Data Analysis. 2000
- R. Decker and W. Gaul  
Classification and Information  
Processing at the Turn of the  
Millennium. 2000
- S. Borra, R. Rocci, M. Vichi,  
and M. Schader (Eds.)  
Advances in Classification  
and Data Analysis. 2001
- W. Gaul and G. Ritter (Eds.)  
Classification, Automation,  
and New Media. 2002

Krzysztof Jajuga · Andrzej Sokołowski  
Hans-Hermann Bock (Eds.)

---

# Classification, Clustering, and Data Analysis

Recent Advances and Applications

With 84 Figures and 65 Tables



Springer

Prof. Krzysztof Jajuga  
Wroclaw University  
of Economics  
ul. Komandorska 118/120  
53-345 Wroclaw  
Poland  
jajuga@manager.ae.wroc.pl

Prof. Hans-Hermann Bock  
Technical University of Aachen  
Institute of Statistics  
Wuellnerstrasse 3  
52056 Aachen  
Germany  
bock@stochastik.rwth-  
aachen.de

Prof. Andrzej Sokołowski  
Department of Statistics  
Cracow University  
of Economics  
ul. Rakowicka 27  
31-510 Cracow  
Poland  
sokolows@ae.krakow.pl

ISSN 1431-8814  
ISBN 978-3-540-43691-1

Cataloging-in-Publication Data applied for  
Die Deutsche Bibliothek – CIP-Einheitsaufnahme  
Classification, clustering and data analysis: recent advances and applications / Krzysztof  
Jajuga ... (ed.). – Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan;  
Paris; Tokyo: Springer, 2002  
(Studies in classification, data analysis, and knowledge organization)  
ISBN 978-3-540-43691-1 ISBN 978-3-642-56181-8 (eBook)  
DOI 10.1007/978-3-642-56181-8

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, reuse of illus-  
trations, recitation, broadcasting, reproduction on microfilm 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.

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002  
Originally published by Springer-Verlag Berlin · Heidelberg in 2002

The use of general descriptive names, registered names, trademarks, etc. in this publica-  
tion does not imply, even in the absence of a specific statement, that such names are  
exempt from the relevant protective laws and regulations and therefore free for general  
use.

Softcover-Design: Erich Kirchner, Heidelberg

43/3111 - 5 4 3 2 1 – Printed on acid-free paper

# Preface

The present volume contains a selection of papers presented at the Eighth Conference of the International Federation of Classification Societies (IFCS) which was held in Cracow, Poland, July 16-19, 2002. All originally submitted papers were subject to a reviewing process by two independent referees, a procedure which resulted in the selection of the 53 articles presented in this volume.

These articles relate to theoretical investigations as well as to practical applications and cover a wide range of topics in the broad domain of classification, data analysis and related methods. If we try to classify the wealth of problems, methods and approaches into some representative (partially overlapping) groups, we find in particular the following areas:

- Clustering
- Cluster validation
- Discrimination
- Multivariate data analysis
- Statistical methods
- Symbolic data analysis
- Consensus trees and phylogeny
- Regression trees
- Neural networks and genetic algorithms
- Applications in economics, medicine, biology, and psychology.

Given the international orientation of IFCS conferences and the leading role of IFCS in the scientific world of classification, clustering and data analysis, this volume collects a representative selection of current research and modern applications in this field and serves as an up-to-date information source for statisticians, data analysts, data mining specialists and computer scientists.

This is well in the mainstream of the activities of the International Federation of Classification Societies which considers as one of its main purposes the dissemination of technical and scientific information concerning data analysis, classification, related methods and their applications. Note that the IFCS comprises, as its members, the following twelve regional or national classification societies:

- British Classification Society (BCS)
- Associação Portuguesa de Classificação e Análise de Dados (CLAD)
- Classification Society of North America (CSNA)
- Gesellschaft für Klassifikation (GfKl)
- Japanese Classification Society (JCS)
- Korean Classification Society (KCS)

- Société Francophone de Classification (SFC)
- Società Italiana di Statistica (SIS)
- Sekcja Klasyfikacji i Analizy Danych PTS (SKAD)
- Vereniging voor Ordinatie en Classificatie (VOC)
- Irish Pattern Recognition and Classification Society (IPRCS)
- Central American and Caribbean Society of Classification and Data Analysis (CASCCDA)

Previous conferences of IFCS took place in Aachen (1987), Charlottesville (1989), Edinburgh (1991), Paris (1993), Kobe (1996), Rome (1998), and Namur (2000).

The editors of this volume would like express their gratitude towards the authors of the papers contained in this volume for their valuable contributions. In particular, they thank the reviewers for their professional, careful and often time-consuming work when reviewing the submitted papers. Our special thanks and gratitude go to Dr. Katarzyna Kuziak of Wrocław University of Economics for her outstanding work and great help in the process of reviewing and preparing final version of the book. We express our gratitude to Dr. Andrzej Bąk of Wrocław University of Economics for his professional contribution in the process of the formatting and technical editing of this volume. Thanks also to Dr. Martina Bihn from Springer Verlag (Heidelberg) for the smooth and timely production and dispatch of these Proceedings.

Aachen, Wrocław, and Cracow  
July 2002

*Hans-Hermann Bock*  
*Krzysztof Jajuga*  
*Andrzej Sokotowski*

# Contents

Preface .....	V
---------------	---

---

## Part I. Clustering and Discrimination

---

<i>Clustering</i> .....	3
<b>Some Thoughts about Classification</b> .....	5
<i>Frank Hampel</i>	
<b>Partial Defuzzification of Fuzzy Clusters</b> .....	27
<i>Slavka Bodjanova</i>	
<b>A New Clustering Approach, Based on the Estimation of the Probability Density Function, for Gene Expression Data</b> .....	35
<i>Noël Bonnet, Michel Herbin, Jérôme Cutrona, Jean-Marie Zahm</i>	
<b>Two-mode Partitioning: Review of Methods and Application of Tabu Search</b> .....	43
<i>William Castillo, Javier Trejos</i>	
<b>Dynamical Clustering of Interval Data Optimization of an Adequacy Criterion Based on Hausdorff Distance</b> .....	53
<i>Marie Chavent, Yves Lechevallier</i>	
<b>Removing Separation Conditions in a 1 against 3-Components Gaussian Mixture Problem</b> .....	61
<i>Bernard Garel and Franck Goussanou</i>	
<b>Obtaining Partitions of a Set of Hard or Fuzzy Partitions</b> .....	75
<i>Allan D. Gordon, Maurizio Vichi</i>	
<b>Clustering for Prototype Selection using Singular Value Decomposition</b> .....	81
<i>A.K.V.Sai Jayram, M.Narasimha Murty</i>	
<b>Clustering in High-dimensional Data Spaces</b> .....	89
<i>Fionn Murtagh</i>	
<b>Quantization of Models: Local Approach and Asymptotically Optimal Partitions</b> .....	97
<i>Klaus Pötzelberger</i>	

<b>The Performance of an Autonomous Clustering Technique</b> . . . .	107
<i>Yoshiharu Sato</i>	
<b>Cluster Analysis with Restricted Random Walks</b> . . . . .	113
<i>Joachim Schöll, Elisabeth Paschinger</i>	
<b>Missing Data in Hierarchical Classification of Variables – a Simulation Study</b> . . . . .	121
<i>Ana Lorga da Silva, Helena Bacelar-Nicolau, Gilbert Saporta</i>	
<b>Cluster Validation</b> . . . . .	129
<b>Representation and Evaluation of Partitions</b> . . . . .	131
<i>Alain Guénoche, Henri Garreta</i>	
<b>Assessing the Number of Clusters of the Latent Class Model</b> .	139
<i>François-Xavier Jollois, Mohamed Nadif and Gérard Govaert</i>	
<b>Validation of Very Large Data Sets Clustering by Means of a Nonparametric Linear Criterion</b> . . . . .	147
<i>Israel Lerman, Joaquim Pinto da Costa, Helena Silva</i>	
<b>Discrimination</b> . . . . .	159
<b>Effect of Feature Selection on Bagging Classifiers Based on Kernel Density Estimators</b> . . . . .	161
<i>Edgar Acuña, Alex Rojas, Frida Coaquira</i>	
<b>Biplot Methodology for Discriminant Analysis Based upon Robust Methods and Principal Curves</b> . . . . .	169
<i>Sugnet Gardner, Niel le Roux</i>	
<b>Bagging Combined Classifiers</b> . . . . .	177
<i>Torsten Hothorn and Berthold Lausen</i>	
<b>Application of Bayesian Decision Theory to Constrained Clas- sification Networks</b> . . . . .	185
<i>Hans J. Vos</i>	

---

## Part II. Multivariate Data Analysis and Statistics

---

<b>Multivariate Data Analysis</b> . . . . .	193
<b>Quotient Dissimilarities, Euclidean Embeddability, and Huy- gens' Weak Principle</b> . . . . .	195
<i>François Bavaud</i>	



<b>Conjoint Analysis and Stimulus Presentation – a Comparison of Alternative Methods</b> .....	203
<i>Michael Brusch, Daniel Baier, Antje Treppa</i>	
<b>Grade Correspondence-cluster Analysis Applied to Separate Components of Reversely Regular Mixtures</b> .....	211
<i>Alicja Ciok</i>	
<b>Obtaining Reducts with a Genetic Algorithm</b> .....	219
<i>José Luis Espinoza</i>	
<b>A Projection Algorithm for Regression with Collinearity</b> .....	227
<i>Peter Filzmoser, Christophe Croux</i>	
<b>Confronting Data Analysis with Constructivist Philosophy</b> ...	235
<i>Christian Hennig</i>	
<b><i>Statistical Methods</i></b> .....	245
<b>Maximum Likelihood Clustering with Outliers</b> .....	247
<i>María Teresa Gallegos</i>	
<b>An Improved Method for Estimating the Modes of the Prob- ability Density Function and the Number of Classes for PDF- based Clustering</b> .....	257
<i>Michel Herbin, Noel Bonnet</i>	
<b>Maximization of Measure of Allowable Sample Sizes Region in Stratified Sampling</b> .....	263
<i>Marcin Skibicki</i>	
<b>On Estimation of Population Averages on the Basis of Cluster Sample</b> .....	271
<i>Janusz Wywiał</i>	
<b><i>Symbolic Data Analysis</i></b> .....	279
<b>Symbolic Regression Analysis</b> .....	281
<i>Lynne Billard, Edwin Diday</i>	
<b>Modelling Memory Requirement with Normal Symbolic Form</b>	289
<i>Marc Csernel, Francisco de A. T. de Carvalho</i>	
<b>Mixture Decomposition of Distributions by Copulas</b> .....	297
<i>Edwin Diday</i>	
<b>Determination of the Number of Clusters for Symbolic Ob- jects Described by Interval Variables</b> .....	311
<i>André Hardy, Pascale Lallemand</i>	

<b>Symbolic Data Analysis Approach to Clustering Large Datasets</b> .....	319
<i>Simona Korenjak-Černe, Vladimir Batagelj</i>	
<b>Symbolic Class Descriptions</b> .....	329
<i>Mathieu Vrac, Edwin Diday, Suzanne Winsberg, Mohamed Mehdi Limam</i>	
<b>Consensus Trees and Phylogenetics</b> .....	339
<b>A Comparison of Alternative Methods for Detecting Reticulation Events in Phylogenetic Analysis</b> .....	341
<i>Olivier Gauthier, François-Joseph Lapointe</i>	
<b>Hierarchical Clustering of Multiple Decision Trees</b> .....	349
<i>Branko Kavšek, Nada Lavrač, Anuška Ferligoj</i>	
<b>Multiple Consensus Trees</b> .....	359
<i>François-Joseph Lapointe, Guy Cucumel</i>	
<b>A Family of Average Consensus Methods for Weighted Trees</b> .	365
<i>Claudine Levasseur, François-Joseph Lapointe</i>	
<b>Comparison of Four methods for Inferring Additive Trees from Incomplete Dissimilarity Matrices</b> .....	371
<i>Vladimir Makarenkov</i>	
<b>Quartet Trees as a Tool to Reconstruct Large Trees from Sequences</b> .....	379
<i>Heiko A. Schmidt, Arndt von Haeseler</i>	
<b>Regression Trees</b> .....	389
<b>Regression Trees for Longitudinal Data with Time-dependent Covariates</b> .....	391
<i>Giuliano Galimberti, Angela Montanari</i>	
<b>Tree-based Models in Statistics: Three Decades of Research</b> ..	399
<i>Eugeniusz Gatnar</i>	
<b>Computationally Efficient Linear Regression Trees</b> .....	409
<i>Luis Torgo</i>	
<b>Neural Networks and Genetic Algorithms</b> .....	417
<b>A Clustering Based Procedure for Learning the Hidden Unit Parameters in Elliptical Basis Function Networks</b> .....	419
<i>Marilena Pillati, Daniela G. Calò</i>	

**Multi-layer Perceptron on Interval Data** ..... 427  
*Fabrice Rossi, Brieuc Conan-Guez*

**Part III. Applications**

**Textual Analysis of Customer Statements for Quality Control and Help Desk Support** ..... 437  
*Ulrich Bohnacker, Lars Dehning, Jürgen Franke, Ingrid Renz*

**AHP as Support for Strategy Decision Making in Banking** ... 447  
*Czesław Domański, Jarosław Kondrasiuk*

**Bioinformatics and Classification: The Analysis of Genome Expression Data** ..... 455  
*Berthold Lausen*

**Glaucoma Diagnosis by Indirect Classifiers** ..... 463  
*Andrea Peters, Torsten Hothorn, Berthold Lausen*

**A Cluster Analysis of the Importance of Country and Sector on Company Returns** ..... 471  
*Clifford W. Sell*

**Problems of Classification in Investigative Psychology** ..... 479  
*Paul J. Taylor, Craig Bennell, Brent Snook*

**List of Reviewers** ..... 489

**Index** ..... 491