

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, Birmingham
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

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Singapore

Tokyo

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
Analysis of Symbolic Data. 2000

Henk A.L. Kiers · Jean-Paul Rasson
Patrick J.F. Groenen · Martin Schader (Eds.)

Data Analysis, Classification, and Related Methods

With 96 Figures



Springer

Professor Dr. Henk A.L. Kiers
University of Groningen
Heymans Institute (PA)
Grote Kruisstraat 2/1
NL-9712 TS Groningen

Dr. Patrick J.F. Groenen
Leiden University
Data Theory Group
Department of Education
P.O. Box 9555
NL-2300 RB Leiden

Professor Dr. Jean-Paul Rasson
University of Namur
Directeur du Department
de Mathématique
Facultés Universitaires
Notre-Dame de la Paix
Rempart de la Vierge, 8
B-5000 Namur

Professor Dr. Martin Schader
University of Mannheim
Lehrstuhl
für Wirtschaftsinformatik III
Schloß
D-68131 Mannheim

*Proceedings of the 7th Conference of the
International Federation of Classification Societies (IFCS-2000)
University of Namur, Belgium
11-14 July, 2000*

Cataloging-in-Publication Data applied for
Data analysis, classification and related methods / Henk A.L. Kiers ... (ed.). – Berlin; Heidelberg;
New York; Barcelona; Hong Kong; London; Milan; Paris; Singapore; Tokyo: Springer, 2000
(Studies in classification, data analysis, and knowledge organization)

ISBN-13: 978-3-540-67521-1 e-ISBN-13: 978-3-642-59789-3

DOI: 10.1007/978-3-642-59789-3

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 illustrations, 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.

Springer-Verlag is a company in the BertelsmannSpringer publishing group.
© Springer-Verlag Berlin · Heidelberg 2000

The use of general descriptive names, registered names, trademarks, etc. in this publication 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

SPIN 10725385

43/2202-5 4 3 2 1 0 – Printed on acid-free paper

Preface

This volume contains a selection of papers presented at the Seventh Conference of the International Federation of Classification Societies (IFCS-2000), which was held in Namur, Belgium, July 11–14, 2000. From the originally submitted papers, a careful review process involving two reviewers per paper, led to the selection of 65 papers that were considered suitable for publication in this book.

The present book contains original research contributions, innovative applications and overview papers in various fields within data analysis, classification, and related methods. Given the fast publication process, the research results are still up-to-date and coincide with their actual presentation at the IFCS-2000 conference. The topics captured are:

- Cluster analysis
- Comparison of clusterings
- Fuzzy clustering
- Discriminant analysis
- Mixture models
- Analysis of relationships data
- Symbolic data analysis
- Regression trees
- Data mining and neural networks
- Pattern recognition
- Multivariate data analysis
- Robust data analysis
- Data science and sampling

The IFCS (International Federation of Classification Societies)

The IFCS promotes the dissemination of technical and scientific information concerning data analysis, classification, related methods, and their applications. The IFCS is a federation of the following member 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)

Previous IFCS-conferences were held in Aachen (Germany, 1987), Charlottesville (USA, 1989), Edinburgh (UK, 1991), Paris (France, 1993), Kobe (Japan, 1996), and Rome (Italy, 1998).

Acknowledgements

First of all, we wish to express our gratitude towards the authors of the papers in the present volume, not only for their contributions, but also for their diligence and timely production of the final versions of their papers. Secondly, we thank the reviewers (listed at the end of this book) for their careful reviews of the originally submitted papers, and in this way, for their support in selecting the best papers in this publication.

We also thank M. Bihn, F. Holzwarth, and R. Milewski of Springer-Verlag, Heidelberg, for their support and dedication to the production of this volume.

Finally, the technical and administrative support we received from J.M. Baan, E. de Boer, K. Friesen, D. Jacquemin, B. Kip, H.J. Kreusch, and A. Verstappen-Remmers is gratefully acknowledged.

Groningen, Namur, Leiden, Mannheim
July 2000

Henk A.L. Kiers
Jean-Paul Rasson
Patrick J.F. Groenen
Martin Schader

Contents

Part I. Cluster Analysis

Cluster Analysis and Mixture Models

Classifier Probabilities 3

J. A. Hartigan

Cluster Analysis Based on Data Depth 17

Richard Hoberg

An Autonomous Clustering Technique 23

Yoshiharu Sato

Unsupervised Non-hierarchical Entropy-based Clustering 29

M. Jardino

**Improving the Additive Tree Representation of a
Dissimilarity Matrix Using Reticulations** 35

Vladimir Makarenkov, Pierre Legendre

Double Versus Optimal Grade Clusterings 41

Alicja Ciok

**The Effects of Initial Values and the Covariance Structure
on the Recovery of some Clustering Methods** 47

Istvan Hajnal, Geert Loosveldt

What Clusters Are Generated by Normal Mixtures? 53

Christian Hennig

A Bootstrap Procedure for Mixture Models 59

Suzanne Winsberg, Geert deSoete

Fuzzy Clustering

A New Criterion of Classes Validity 63

Arnaud Devillez, Patrice Billaudel, Gérard Villermain Lecolier

**Application of Fuzzy Mathematical Morphology for
Unsupervised Color Pixels Classification** 69

A. Gillet, C. Botte-Lecocq, L. Macaire and J.-G. Postaire

A Hyperbolic Fuzzy k-Means Clustering and Algorithm for Neural Networks	77
<i>Norio Watanabe, Tadashi Imaizumi, Toshiko Kikuchi</i>	
<i>Special Purpose Classification Procedures and Applications</i>	
A Toolkit for Development of the Domain-Oriented Dictionaries for Structuring Document Flows	83
<i>Pavel P. Makagonov, Mikhail A. Alexandrov, Konstantin Sboychakov</i>	
Classification of Single Malt Whiskies	89
<i>David Wishart</i>	
Robust Approach in Hierarchical Clustering: Application to the Sectorisation of an Oil Field	95
<i>Jean-Paul Valois</i>	
A Minimax Solution for Sequential Classification Problems ...	101
<i>Hans J. Vos</i>	
<i>Verification and Comparison of Clusterings</i>	
Comparison of Ultrametrics Obtained With Real Data, Using the P_L and VAL_{Aw} Coefficients	107
<i>Isabel Pinto Doria, Georges Le Calvé, Helena Bacelar-Nicolau</i>	
Numerical Comparisons of two Spectral Decompositions for Vertex Clustering	113
<i>P. Kuntz, F. Henaux</i>	
Measures to Evaluate Rankings of Classification Algorithms	119
<i>Carlos Soares, Pavel Brazdil, Joaquim Costa</i>	
A General Approach to Test the Pertinence of a Consensus Classification	125
<i>Guy Cucumel, François-Joseph Lapointe</i>	
<i>Dissimilarity Measures</i>	
On a Class of Aggregation-invariant Dissimilarities Obeying the Weak Huygens' Principle	131
<i>F. Bavaud</i>	
A Short Optimal Way for Constructing Quasi-ultrametrics From Some Particular Dissimilarities	137
<i>B. Fichet</i>	

Missing Data in Cluster Analysis

Estimating Missing Values in a Tree Distance	143
<i>A. Guénoche, S. Grandcolas</i>	
Estimating Trees From Incomplete Distance Matrices: A Comparison of Two Methods	149
<i>Claudine Levasseur, Pierre-Alexandre Landry, François-Joseph Lapointe</i>	
Zero Replacement in Compositional Data Sets	155
<i>J. A. Martín-Fernández, C. Barceló-Vidal, V. Pawlowsky-Glahn</i>	
EM Algorithm for Partially Known Labels	161
<i>C. Ambroise, G. Govaert</i>	

Part II. Discrimination, Regression Trees, and Data Mining

Discriminant Analysis

Detection of Company Failure and Global Risk Forecasting . . .	169
<i>Mireille Bardos</i>	
Discriminant Analysis by Hierarchical Coupling in EDDA Context	175
<i>Isabel Brito, Gilles Celeux</i>	
Discrete Discriminant Analysis: The Performance of Combining Models by a Hierarchical Coupling Approach . . .	181
<i>Ana Sousa Ferreira, Gilles Celeux, Helena Bacelar-Nicolau</i>	
Discrimination Based on the Atypicality Index versus Density Function Ratio	187
<i>H. Chamlal and S. Slaoui Chah</i>	

Decision and Regression Trees

A Third Stage in Regression Tree Growing: Searching for Statistical Reliability	193
<i>Carmela Cappelli, Francesco Mola, Roberta Siciliano</i>	
A New Sampling Strategy for Building Decision Trees from Large Databases	199
<i>J.H. Chauchat, R. Rakotomalala</i>	

Generalized Additive Multi-Model for Classification and Prediction 205
Claudio Conversano, Francesco Mola, Roberta Siciliano

Radial Basis Function Networks and Decision Trees in the Determination of a Classifier 211
Rossella Miglio, Marilena Pillati

Clustered Multiple Regression 217
Luis Torgo, J. Pinto da Costa

Neural Networks and Data Mining

Artificial Neural Networks, Censored Survival Data, Statistical Models 223
Antonio Ciampi, Yves Lechevallier

Visualisation and Classification with Artificial Life 229
Alfred Ultsch

Pattern Recognition and Geometrical Statistics

Exploring the Periphery of Data Scatters: Are There Outliers? 235
Giovanni C. Porzio, Giancarlo Ragozini

Discriminant Analysis Tools for Non Convex Pattern Recognition 241
Marcel Rémon

A Markovian Approach to Unsupervised Multidimensional Pattern Classification 247
A. Sbihi, A. Moussa, B. Benmiloud, J.-G. Postaire

Part III. Multivariate and Multidimensional Data Analysis

Multivariate Data Analysis

An Algorithm with Projection Pursuit for Sliced Inverse Regression Model 255
Masahiro Mizuta, Hiroyuki Minami

Testing Constraints and Misspecification in VAR-ARCH Models 261
Wolfgang Polasek, Shuangzhe Liu

**Goodness of Fit Measure based on Sample Isotone Regression
of Mokken Double Monotonicity Model** 267
Teresa Rivas Moya

Multiway Data Analysis

**Fuzzy Time Arrays and Dissimilarity Measures
for Fuzzy Time Trajectories** 273
Renato Coppi, Pierpaolo D'Urso

Three-Way Partial Correlation Measures 279
Donatella Vicari

***Analysis of Network and Relationship Data
and Multidimensional Scaling***

Statistical Models for Social Networks 285
Stanley Wasserman, Philippa Pattison

**Application of Simulated Annealing in some
Multidimensional Scaling Problems** 297
Javier Trejos, William Castillo, Jorge González, Mario Villalobos

Data Analysis Based on Minimal Closed Subsets 303
S. Bonnevey, C. LARGERON-LETEÑO

Robust Multivariate Methods

A Robust Method for Multivariate Regression 309
Stefan Van Aelst, Katrien Van Driessen, Peter J. Rousseeuw

Robust Methods for Complex Data Structures 315
Ursula Gather, Claudia Becker, Sonja Kuhnt

Robust Methods for Canonical Correlation Analysis 321
Catherine Dehon, Peter Filzmoser, Christophe Croux

Part IV. Data Science

Data Science and Data Collection

From Data Analysis to Data Science 329
Noboru Ohsumi

Evaluation of Data Quality and Data Analysis 335
Chikio Hayashi

Collapsibility and Collapsing Multidimensional Contingency Tables—Perspectives and Implications 341
Stefano De Cantis, Antonino M. Oliveri

Sampling and Internet Surveys

Data Collected on the Web 347
Vasja Vehovar, Katja Lozar Manfreda, Zenel Batagelj

Some Experimental Surveys on the WWW Environments in Japan 353
Osamu Yoshimura, Noboru Ohsumi

Bootstrap Goodness-of-fit Tests for Complex Survey Samples 359
Andrea Scagni

Part V. Symbolic Data Analysis

Classification and Analysis of Symbolic Data

Regression Analysis for Interval-Valued Data 369
L. Billard, E. Diday

Symbolic Approach to Classify Large Data Sets..... 375
Francisco de A.T. de Carvalho, Cezar A. de F. Anselmo, Renata M.C.R. de Souza

Factorial Methods with Cohesion Constraints on Symbolic Objects 381
N.C. Lauro, R. Verde, F. Palumbo

A Dynamical Clustering Algorithm for Multi-nominal Data... 387
Rosanna Verde, Francisco de A. T. de Carvalho, Yves Lechevallier

Software

DB2SO : A Software for Building Symbolic Objects from Databases 395
Georges Hébrail, Yves Lechevallier

Symbolic Data Analysis and the SODAS Software in Official Statistics 401
Raymond Bisdorff, Edwin Diday

Strata Decision Tree SDA Software 409
M. Carmen Bravo

**Marking and Generalization by Symbolic Objects
in the Symbolic Official Data Analysis Software.....** 417
Mireille Gettler Summa

List of Reviewers 423

Index 425