Studies in Classification, Data Analysis, and Knowledge Organization

Managing Editors
H.-H. Bock, Aachen
W. Gaul, Karlsruhe
M. Vichi, Rome

Editorial Board
Ph. Arabie, Newark
D. Baier, Cottbus
F. Critchley, Milton Keynes
R. Decker, Bielefeld
E. Diday, Paris
M. Greenacre, Barcelona
C. Lauro, Naples
J. Meulman, Leiden
P. Monari, Bologna
S. Nishisato, Toronto
N. Ohsumi, Tokyo
O. Opitz, Augsburg
G. Ritter, Passau
M. Schader, Mannheim
C. Weihs, Dortmund
Titles in the Series:

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 (Eds.)
Classification and Information Processing at the Turn of the Millenium. 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

K. Jajuga, A. Sokolowski, and H.-H. Bock (Eds.)
Classification, Clustering and Data Analysis. 2002

M. Schwaiger and O. Opitz (Eds.)
Exploratory Data Analysis in Empirical Research. 2003

M. Schader, W. Gaul, and M. Vichi (Eds.)
Between Data Science and Applied Data Analysis. 2003

H.-H. Bock, M. Chiodi, and A. Mineo (Eds.)
Advances in Multivariate Data Analysis. 2004

D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul (Eds.)
Classification, Clustering, and Data Mining Applications. 2004

D. Baier and K.-D. Wernecke (Eds.)
Innovations in Classification, Data Science, and Information Systems. 2005

M. Vichi, P. Monari, S. Mignani and A. Montanari (Eds.)
New Developments in Classification and Data Analysis. 2005

D. Baier, R. Decker, and L. Schmidt-Thieme (Eds.)
Data Analysis and Decision Support. 2005

C. Weihs and W. Gaul (Eds.)
Classification the Ubiquitous Challenge. 2005

M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nünberger and W. Gaul (Eds.)
From Data and Information Analysis to Knowledge Engineering. 2006
Preface

This volume contains a refereed selection of papers presented during the 10th Jubilee Conference of the International Federation of Classification Societies (IFCS) on Data Science and Classification held at the Faculty of Social Sciences of the University of Ljubljana in Slovenia, July 25-29, 2006. Papers submitted for the conference were subjected to a careful reviewing process involving at least two reviewers per paper. As a result of this reviewing process, 37 papers were selected for publication in this volume.

The book presents recent advances in data analysis, classification and clustering from methodological, theoretical, or algorithmic points of view. It shows how data analysis methods can be applied in various subject-specific domains. Areas that receive particular attention in this book are similarity and dissimilarity analysis, discrimination and clustering, network and graph analysis, and the processing of symbolic data. Special sections are devoted to data and web mining and to the application of data analysis methods in quantitative musicology and microbiology. Readers will find a fine selection of recent technical and application-oriented papers that characterize the current developments in data science and classification. The combination of new methodological advances with the wide range of real applications collected in this volume will be of special value for researchers when choosing appropriate newly developed analytical tools for their research problems in classification and data analysis.

The editors are grateful to the authors of the papers in this volume for their contributions and for their willingness to respond so positively to the time constraints in preparing the final versions of their papers. Without their work there would be no book. We are especially grateful to the referees – listed at the end of this book – who reviewed the submitted papers. Their careful reviews helped us greatly in selecting the papers included in this volume.

We also thank Dr. Martina Bihn and the staff of Springer-Verlag, Heidelberg for their support and dedication for publishing this volume in the series Studies in Classification, Data Analysis, and Knowledge Organization.

Vladimir Batagelj, Ljubljana
Hans–Hermann Bock, Aachen
Anuška Ferligoj, Ljubljana
Aleš Žiberna, Ljubljana

May 2006
The 10th IFCS Conference – a Jubilee

The International Federation of Classification Societies (IFCS) was founded July 4, 1985 in Cambridge (UK) at a time when classification problems were frequently being encountered in such diverse fields as biology, marketing, social sciences, pattern recognition, picture processing, information retrieval, and library science. These often massive problems had to be solved by quantitative or computerized methods based on data and measurements. In fact, the IFCS founding agreement paved the way for an intensive bilateral and multilateral cooperation and scientific as well as personal contacts among the members of the six ‘national’ classification societies from United Kingdom (BCS), North America (CSNA), Germany (GfKl), Japan (JCS), France (SFC), and Italy (SIS) that formed the IFCS in those times.

A main activity of IFCS is the organization of a biennial conference series. The first one with the title ‘Classification, Data Analysis, and Related Methods’ was held in Aachen (Germany) from June 29 to July 1, 1987 with about 300 participants from 25 countries and more than 180 scientific papers. Since that time, eight more IFCS conferences have taken place at different locations around the world (see the table below), always with a broad international participation from inside and outside the IFCS. Typically, a selection of scientific papers was published in a Proceedings volume in the Springer series ‘Studies in Classification, Data Analysis, and Knowledge Organization’ so that the results became available worldwide.

<table>
<thead>
<tr>
<th>Year</th>
<th>Place</th>
<th>Hosting Society</th>
<th>Organizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>Aachen (Germany)</td>
<td>GfKl</td>
<td>H.-H. Bock</td>
</tr>
<tr>
<td>1989</td>
<td>Charlottesville (USA)</td>
<td>CSNA</td>
<td>H. Bozdogan</td>
</tr>
<tr>
<td>1991</td>
<td>Edinburgh (UK)</td>
<td>BCS</td>
<td>D. Wishart, A. Gordon</td>
</tr>
<tr>
<td>1993</td>
<td>Paris (France)</td>
<td>SFC</td>
<td>E. Diday</td>
</tr>
<tr>
<td>1996</td>
<td>Kobe (Japan)</td>
<td>JCS</td>
<td>Ch. Hayashi</td>
</tr>
<tr>
<td>1998</td>
<td>Rome (Italy)</td>
<td>SIS</td>
<td>A. Rizzi, M. Vichi</td>
</tr>
<tr>
<td>2000</td>
<td>Namur (Belgium)</td>
<td>SFC, VOC</td>
<td>J.-P. Rasson, H. Kiers</td>
</tr>
<tr>
<td>2002</td>
<td>Cracow (Poland)</td>
<td>SKAD</td>
<td>A. Sokolowski</td>
</tr>
<tr>
<td>2004</td>
<td>Chicago (USA)</td>
<td>CSNA</td>
<td>F.R. McMorris</td>
</tr>
<tr>
<td>2006</td>
<td>Ljubljana (Slovenia)</td>
<td>SDS</td>
<td>A. Ferligoj, V. Batagelj</td>
</tr>
</tbody>
</table>

As a consequence, more and more national groups or societies working in the classification and data analysis field joined the IFCS: the VOC from Belgium/Netherlands, the CLAD from Portugal, the Polish society (SKAD),
the KCS from Korea, the Irish Pattern Recognition and Classification Society (IPRCS), and finally the Central American and Caribbean society (SoCAD).

The 10th IFCS conference is being hosted by the Statistical Society of Slovenia (SDS) at the University of Ljubljana (Slovenia), in July 2006, with Anuska Ferligoj and Vladimir Batagelj chairing. It will, without any doubt, be a new highlight in the history of IFCS, provide a challenging marketplace for scientific and applied research results, and foster further cooperation and contacts within the worldwide classification community.

This Jubilee Conference is certainly an occasion for recalling the history and achievements of the 20 years of IFCS’s life. But it also marks the beginning of another decade of tasks and activities for IFCS: with new challenges for research and application, with interesting scientific conferences, with an intensive cooperation among IFCS members, and hopefully also a large impact on the worldwide development of our favorite domains: data analysis and classification.

May 2006

Hans-Hermann Bock
Contents

Preface ....................................................... V
The 10th IFCS Conference – a Jubilee ............... VII
Hans-Hermann Bock

Contents ....................................................... IX

Part I. Similarity and Dissimilarity

A Tree-Based Similarity for Evaluating Concept Proximities in an Ontology ........................................ 3
Emmanuel Blanchard, Pascale Kuntz, Mounira Harzallah, Henri Briand

Improved Fréchet Distance for Time Series ........... 13
Ahlame Chouakria-Douzal, Panduranga Naidu Nagabhushan

Comparison of Distance Indices Between Partitions .......... 21
Lucile Denœud, Alain Guénoche

Design of Dissimilarity Measures: A New Dissimilarity Between Species Distribution Areas .................. 29
Christian Hennig, Bernhard Hausdorf

Dissimilarities for Web Usage Mining .................... 39
Fabrice Rossi, Francisco De Carvalho, Yves Lechevallier, Alzennyr Da Silva

Properties and Performance of Shape Similarity Measures .... 47
Remco C. Veltkamp, Longin Jan Latecki

Part II. Classification and Clustering

Hierarchical Clustering for Boxplot Variables .............. 59
Javier Arroyo, Carlos Maté, Antonio Muñoz-San Roque

Evaluation of Allocation Rules Under Some Cost Constraints . 67
Farid Beninel, Michel Grun Rehomme

Crisp Partitions Induced by a Fuzzy Set ..................... 75
Slavka Bodjanova
Empirical Comparison of a Monothetic Divisive Clustering Method with the Ward and the k-means Clustering Methods 83
Marie Chavent, Yves Lechevallier

Model Selection for the Binary Latent Class Model: A Monte Carlo Simulation 91
José G. Dias

Finding Meaningful and Stable Clusters Using Local Cluster Analysis 101
Hans-Joachim Mucha

Comparing Optimal Individual and Collective Assessment Procedures 109
Hans J. Vos, Ruth Ben-Yashar, Shmuel Nitzan

Part III. Network and Graph Analysis

Some Open Problem Sets for Generalized Blockmodeling 119
Patrick Doreian

Spectral Clustering and Multidimensional Scaling: A Unified View 131
François Bavaud

Analyzing the Structure of U.S. Patents Network 141
Vladimir Batagelj, Nataša Kejžar, Simona Korenjak-Černe, Matjaž Zaveršnik

Identifying and Classifying Social Groups: A Machine Learning Approach 149
Matteo Roffilli, Alessandro Lomi

Part IV. Analysis of Symbolic Data

Multidimensional Scaling of Histogram Dissimilarities 161
Patrick J.F. Groenen, Suzanne Winsberg

Dependence and Interdependence Analysis for Interval-Valued Variables 171
Carlo Lauro, Federica Gioia

A New Wasserstein Based Distance for the Hierarchical Clustering Of Histogram Symbolic Data 185
Antonio Irpino, Rosanna Verde
Symbolic Clustering of Large Datasets .......................... 193
Yves Lechevallier, Rosanna Verde, Francisco de A.T. de Carvalho

A Dynamic Clustering Method for Mixed Feature-Type Symbolic Data .................................................. 203
Renata M.C.R. de Souza, Francisco de A.T. de Carvalho, Daniel Ferrari Pizzato

Part V. General Data Analysis Methods

Iterated Boosting for Outlier Detection .......................... 213
Nathalie Cheze, Jean-Michel Poggi

Sub-species of Homopus Areolatus? Biplots and Small Class Inference with Analysis of Distance ........................... 221
Sugnet Gardner, Niël J. le Roux

Revised Boxplot Based Discretization as the Kernel of Automatic Interpretation of Classes Using Numerical Variables ..................................................... 229
Karina Gibert, Alejandra Pérez-Bonilla

Part VI. Data and Web Mining

Comparison of Two Methods for Detecting and Correcting Systematic Errors in High-throughput Screening Data .... 241
Andrei Gagarin, Dmytro Kevorkov, Vladimir Makarenkov, Pablo Zentilli

kNN Versus SVM in the Collaborative Filtering Framework .. 251
Miha Grčar, Blaž Fortuna, Dunja Mladenič, Marko Grobelnik

Mining Association Rules in Folksonomies ....................... 261
Christoph Schmitz, Andreas Hotho, Robert Jäschke, Gerd Stumme

Empirical Analysis of Attribute-Aware Recommendation Algorithms with Variable Synthetic Data ..................... 271
Karen H. L. Tso, Lars Schmidt-Thieme

Patterns of Associations in Finite Sets of Items ............... 279
Ralf Wagner
Part VII. Analysis of Music Data

Generalized N-gram Measures for Melodic Similarity ........... 289
Klaus Frieler

Evaluating Different Approaches to Measuring the Similarity of Melodies ................................................................. 299
Daniel Müllensiefen, Klaus Frieler

Using MCMC as a Stochastic Optimization Procedure for Musical Time Series ........................................... 307
Katrin Sommer, Claus Weihs

Local Models in Register Classification by Timbre ............... 315
Claus Weihs, Gero Szepannek, Uwe Ligges, Karsten Luebke, Nils Raabe

Part VIII. Gene and Microarray Analysis

Improving the Performance of Principal Components for Classification of Gene Expression Data Through Feature Selection ................................................................. 325
Edgar Acuña, Jaime Porras

A New Efficient Method for Assessing Missing Nucleotides in DNA Sequences in the Framework of a Generic Evolutionary Model ...................................................... 333
Abdoulaye Baniré Diallo, Vladimir Makarenkov, Mathieu Blanchette, François-Joseph Lapointe

New Efficient Algorithm for Modeling Partial and Complete Gene Transfer Scenarios ........................................... 341
Vladimir Makarenkov, Alix Boc, Charles F. Delwiche, Alpha Boubacar Diallo, Hervé Philippe

List of Reviewers ............................................................. 351

Key words ................................................................. 353

Authors ....................................................................... 357