


STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION

V. Batagelj
H.-H. Bock
A. Ferligoj
A. Žiberna

Data Science and Classification



 Springer

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

Vladimir Batagelj · Hans-Hermann Bock
Anuška Ferligoj · Aleš Žiberna
Editors

Data Science and Classification

With 67 Figures and 42 Tables

 Springer

Prof. Dr. Vladimir Batagelj
Department of Mathematics, FMF
University of Ljubljana
Jadranska 19
1000 Ljubljana, Slovenia
vladimir.batagelj@fmf.uni-lj.si

Prof. Dr. Anuška Ferligoj
Faculty of Social Sciences
University of Ljubljana
Kardeljeva pl. 5
1000 Ljubljana, Slovenia
anuska.ferligoj@fdv.uni-lj.si

Prof. Dr. Hans-Hermann Bock
Institute of Statistics
RWTH Aachen University
52056 Aachen, Germany
bock@stochastik.rwth-aachen.de

Aleš Žiberna
Faculty of Social Sciences
University of Ljubljana
Kardeljeva pl. 5
1000 Ljubljana, Slovenia
ales.ziberna@fdv.uni-lj.si

ISSN 1431-8814

ISBN-10 3-540-34415-2 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-34415-5 Springer 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, 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. Violations are liable to prosecution under the German Copyright Law.

Springer · Part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin · Heidelberg 2006
Printed in Germany

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.

Production: LE-TeX Jelonek, Schmidt & Vöckler GbR, Leipzig
Softcover-Design: Erich Kirchner, Heidelberg

SPIN 11759263 43/3100/YL – 5 4 3 2 1 0 – Printed on acid-free paper

Contents

Preface	V
The 10th IFCS Conference – a Jubilee	VII
<i>Hans-Hermann Bock</i>	
Contents.....	IX

Part I. Similarity and Dissimilarity

A Tree-Based Similarity for Evaluating Concept Proximities in an Ontology.....	3
<i>Emmanuel Blanchard, Pascale Kuntz, Mounira Harzallah, Henri Briand</i>	
Improved Fréchet Distance for Time Series	13
<i>Ahlame Chouakria-Douzal, Panduranga Naidu Nagabhushan</i>	
Comparison of Distance Indices Between Partitions.....	21
<i>Lucile Denœud, Alain Guénoche</i>	
Design of Dissimilarity Measures: A New Dissimilarity Between Species Distribution Areas	29
<i>Christian Hennig, Bernhard Hausdorf</i>	
Dissimilarities for Web Usage Mining	39
<i>Fabrice Rossi, Francisco De Carvalho, Yves Lechevallier, Alzenny Da Silva</i>	
Properties and Performance of Shape Similarity Measures	47
<i>Remco C. Veltkamp, Longin Jan Latecki</i>	

Part II. Classification and Clustering

Hierarchical Clustering for Boxplot Variables	59
<i>Javier Arroyo, Carlos Maté, Antonio Muñoz-San Roque</i>	
Evaluation of Allocation Rules Under Some Cost Constraints. .	67
<i>Farid Beninel, Michel Grun Rehomme</i>	
Crisp Partitions Induced by a Fuzzy Set	75
<i>Slavka Bodjanova</i>	

Empirical Comparison of a Monothetic Divisive Clustering Method with the Ward and the k-means Clustering Methods	83
<i>Marie Chavent, Yves Lechevallier</i>	
Model Selection for the Binary Latent Class Model: A Monte Carlo Simulation	91
<i>José G. Dias</i>	
Finding Meaningful and Stable Clusters Using Local Cluster Analysis	101
<i>Hans-Joachim Mucha</i>	
Comparing Optimal Individual and Collective Assessment Procedures	109
<i>Hans J. Vos, Ruth Ben-Yashar, Shmuel Nitzan</i>	
<hr/>	
Part III. Network and Graph Analysis	
<hr/>	
Some Open Problem Sets for Generalized Blockmodeling	119
<i>Patrick Doreian</i>	
Spectral Clustering and Multidimensional Scaling: A Unified View	131
<i>François Bavaud</i>	
Analyzing the Structure of U.S. Patents Network	141
<i>Vladimir Batagelj, Nataša Kejžar, Simona Korenjak-Černe, Matjaž Zaveršnik</i>	
Identifying and Classifying Social Groups: A Machine Learning Approach	149
<i>Matteo Roffilli, Alessandro Lomi</i>	
<hr/>	
Part IV. Analysis of Symbolic Data	
<hr/>	
Multidimensional Scaling of Histogram Dissimilarities	161
<i>Patrick J.F. Groenen, Suzanne Winsberg</i>	
Dependence and Interdependence Analysis for Interval-Valued Variables	171
<i>Carlo Lauro, Federica Gioia</i>	
A New Wasserstein Based Distance for the Hierarchical Clustering Of Histogram Symbolic Data	185
<i>Antonio Irpino, Rosanna Verde</i>	

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
<i>Klaus Frieler</i>	
Evaluating Different Approaches to Measuring the Similarity of Melodies	299
<i>Daniel Müllensiefen, Klaus Frieler</i>	
Using MCMC as a Stochastic Optimization Procedure for Musical Time Series	307
<i>Katrin Sommer, Claus Weihs</i>	
Local Models in Register Classification by Timbre	315
<i>Claus Weihs, Gero Szepannek, Uwe Ligges, Karsten Luebke, Nils Raabe</i>	

Part VIII. Gene and Microarray Analysis

Improving the Performance of Principal Components for Classification of Gene Expression Data Through Feature Selection	325
<i>Edgar Acuña, Jaime Porras</i>	
A New Efficient Method for Assessing Missing Nucleotides in DNA Sequences in the Framework of a Generic Evolutionary Model	333
<i>Abdoulaye Baniré Diallo, Vladimir Makarenkov, Mathieu Blanchette, François-Joseph Lapointe</i>	
New Efficient Algorithm for Modeling Partial and Complete Gene Transfer Scenarios	341
<i>Vladimir Makarenkov, Alix Boc, Charles F. Delwiche, Alpha Boubacar Diallo, Hervé Philippe</i>	
List of Reviewers	351
Key words	353
Authors	357