# 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, and O. Opitz (Eds.) Ordinal and Symbolic Data Analysis. 1996
I. Balderjahn, R. Mather, 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
W. Gaul, O. Opitz, M. Schader (Eds.) Data Analysis. 2000
R. Decker and W. Gaul (Eds.) Classification and Information Processing at the Turn of the Millennium. 2000
S. Borra, R. Roccia, M. Vichi, and M. Schader (Eds.) Advances in Classification and Data Analysis. 2000
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. Nürnberg, and W. Gaul (Eds.) From Data and Information Analysis to Knowledge Engineering. 2006
V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna (Eds.) Data Science and Classification. 2006
S. Zani, A. Cerioli, M. Riani, M. Vichi (Eds.) Data Analysis, Classification and the Forward Search. 2006
P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.) Selected Contributions in Data Analysis and Classification. 2007
R. Decker, H.-J. Lenz (Eds.) Advances in Data Analysis. 2007
C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.) Data Analysis, Machine Learning and Applications. 2008
Preface

This volume contains the revised versions of selected papers presented during the 31st Annual Conference of the German Classification Society (Gesellschaft für Klassifikation – GfKl). The conference was held at the Albert-Ludwigs-University in Freiburg, Germany, in March 2007. The focus of the conference was on Data Analysis, Machine Learning, and Applications, it comprised 200 talks in 36 sessions. Additionally 11 plenary and semi-plenary talks were held by outstanding researchers. With 292 participants from 19 countries in Europe and overseas this GfKl Conference, once again, provided an international forum for discussions and mutual exchange of knowledge with colleagues from different fields of interest. From altogether 120 full papers that had been submitted for this volume 82 were finally accepted.

With the occasion of the 30th anniversary of the German Classification Society the associated societies Sekcja Klasyfikacji i Analizy Danych PTS (SKAD), Vereniging voor Ordinatie en Classificatie (VOC), Japanese Classification Society (JCS) and Classification and Data Analysis Group (CLADAG) have sponsored the following invited talks: Paul Eilers - Statistical Classification for Reliable High-volume Genetic Measurements (VOC); Eugeniusz Gatnar - Fusion of Multiple Statistical Classifiers (SKAD); Akinori Okada - Two-Dimensional Centrality of a Social Network (JCS); Donatella Vicari - Unsupervised Multivariate Prediction Including Dimensionality Reduction (CLADAG).

The scientific program included a broad range of topics, besides the main theme of the conference, especially methods and applications of data analysis and machine learning were considered. The following sessions were established:

I. Theory and Methods
Supervised Classification, Discrimination, and Pattern Recognition (G. Ritter); Cluster Analysis and Similarity Structures (H.-H. Bock and J. Buhmann); Classification and Regression (C. Bailer-Jones and C. Hennig); Frequent Pattern Mining (C. Borgelt); Data Visualization and Scaling Methods (P. Groenen, T. Imaizumi, and A. Okada); Exploratory Data Analysis and Data Mining (M. Meyer and M. Schwaiiger); Mixture Analysis in Clustering (S. Ingrassia, D. Karlis, P. Schlattmann and W. Sei-
II. Applications
Marketing and Management Science (D. Baier, Y. Boztug, and W. Steiner); Banking and Finance (K. Jajuga and H. Locarek-Junge); Business Intelligence and Personalization (A. Geyer-Schulz and L. Schmidt-Thieme); Data Analysis in Retailing (T. Reutterer); Econometrics and Operations Research (W. Polasek); Image and Signal Analysis (H. Burkhardt); Biostatistics and Bioinformatics (R. Backofen, H.-P. Klenk and B. Lausen); Medical and Health Sciences (K.-D. Wernecke); Text Mining, Web Mining, and the Semantic Web (A. Nürnberger and M. Spiliopoulou); Statistical Natural Language Processing (P. Cimiano); Linguistics (H. Goebl and P. Grzybek); Subject Indexing and Library Science (H.-J. Hermes and B. Lorenz); Statistical Musicology (C. Weihs); Archaeology and Archaeometry (M. Helfert and I. Herzog); Psychology (S. Krolak-Schwerdt); Data Analysis in Higher Education (A. Hilbert).

Contributed Sessions (by CLADAG and SKAD)
Latent class models for classification (A. Montanari and A. Cerioli); Classification and models for interval-valued data (F. Palumbo); Selected Problems in Classification (E. Gatnar); Recent Developments in Multidimensional Data Analysis between research and practice I (L. D’Ambra); Recent Developments in Multidimensional Data Analysis between research and practice II (B. Simonetti).

The editors would like to emphatically thank all the section chairs for doing such a great job regarding the organization of their sections and the associated paper reviews.

Cordial thanks also go to the members of the scientific program committee for their conceptual and practical support as well as for the paper reviews: D. Baier (Cottbus), H.-H. Bock (Aachen), H. Bozdogan (Tennessee), J. Buhmann (Zürich), H. Burkhardt (Freiburg), A. Cerioli (Parma); R. Decker (Bielefeld), W. Gaul (Karlsruhe), A. Geyer-Schulz (Karlruhe), P. Groenen (Rotterdam), T. Imaizumi (Tokyo), K. Jajuga (Wroclaw), R. Kruse (Magdeburg), S. Lang (Innsbruck), B. Lausen (Erlangen-Nürnberg), H.-J. Lenz (Berlin), F. Murtagh (London), H. Ney (Aachen), A. Okada (Tokyo), L. Schmidt-Thieme (Hildesheim), C. Schnoerr (Mannheim), M. Spiliopoulou (Magdeburg), C. Weihs (Dortmund), D. A. Zighed (Lyon).

Furthermore we would like to thank the additional reviewers: A. Hotho, L. Marinho, C. Preisach, S. Rendle, S. Scholz, K. Tso.

The great success of this conference would not have been possible without the support of many people mainly working in the backstage. We would like to particularly thank M. Temerinac (Freiburg), J. Fehr (Freiburg), C. Findlay (Freiburg), E. Patschke (Freiburg), A. Busche (Hildesheim), K. Tso (Hildesheim), L. Marinho (Hildesheim) and the student support team for their hard work in the preparation
of this conference, for the support during the event and the post-processing of the conference.

The GfKl Conference 2007 would not have been possible in the way it took place without the financial and/or material support of the following institutions and companies (in alphabetical order): Albert-Ludwigs-University Freiburg – Faculty of Applied Sciences, Gesellschaft für Klassifikation e.V., Microsoft München and Springer Verlag. We express our gratitude to all of them. Finally, we would like to thank Dr. Martina Bihn from Springer Verlag, Heidelberg, for her support and dedication to the production of this volume.

Hildesheim, Freiburg and Bielefeld, February 2008

Christine Preisach
Hans Burkhardt
Lars Schmidt-Thieme
Reinhold Decker
Contents

Part I Classification

Distance-based Kernels for Real-valued Data
Lluís Belanche, Jean Luis Vázquez, Miguel Vázquez ............................. 3

Fast Support Vector Machine Classification of Very Large Datasets
Janis Fehr, Karina Zapieán Arreola, Hans Burkhardt ......................... 11

Fusion of Multiple Statistical Classifiers
Eugeniusz Gatnar ........................................................................ 19

Calibrating Margin–based Classifier Scores into Polychotomous
Probabilities
Martin Gebel, Claus Weihs .......................................................... 29

Classification with Invariant Distance Substitution Kernels
Bernard Haasdonk, Hans Burkhardt ............................................... 37

Applying the Kohonen Self-organizing Map Networks to Select Variables
Kamila Migdal Najman, Krzysztof Najman ....................................... 45

Computer Assisted Classification of Brain Tumors
Norbert Röhrl, José R. Iglesias-Rozas, Galia Weidl ......................... 55

Model Selection in Mixture Regression Analysis – A Monte Carlo
Simulation Study
Marko Sarstedt, Manfred Schwaiger ............................................ 61

Comparison of Local Classification Methods
Julia Schiffner, Claus Weihs ........................................................ 69

Incorporating Domain Specific Information into Gaia Source
Classification
Kester W. Smith, Carola Tiede, Coryn A.L. Bailer-Jones ................... 77
Identification of Noisy Variables for Nonmetric and Symbolic Data in Cluster Analysis
Marek Walesiak, Andrzej Dudek .................................................. 85

Part II Clustering

Families of Dendrograms
Patrick Erik Bradley ................................................................. 95

Mixture Models in Forward Search Methods for Outlier Detection
Daniela G. Calò ........................................................................ 103

On Multiple Imputation Through Finite Gaussian Mixture Models
Marco Di Zio, Ugo Guarnera ..................................................... 111

Mixture Model Based Group Inference in Fused Genotype and Phenotype Data
Benjamin Georgi, M.Anne Spence, Pamela Flodman, Alexander Schliep .... 119

The Noise Component in Model-based Cluster Analysis
Christian Hennig, Pietro Coretto ............................................... 127

An Artificial Life Approach for Semi-supervised Learning
Lutz Herrmann, Alfred Ultsch ................................................... 139

Hard and Soft Euclidean Consensus Partitions
Kurt Hornik, Walter Böhm ......................................................... 147

Rationale Models for Conceptual Modeling
Sina Lehrmann, Werner Esswein ............................................... 155

Measures of Dispersion and Cluster-Trees for Categorical Data
Ulrich Müller-Funk ................................................................ 163

Information Integration of Partially Labeled Data
Steffen Rendle, Lars Schmidt-Thieme ......................................... 171
### Part III Multidimensional Data Analysis

**Data Mining of an On-line Survey - A Market Research Application**  
*Karmele Fernández-Aguirre, María I. Landaluce, Ana Martín, Juan I. Modroño*  
...
  
**Nonlinear Constrained Principal Component Analysis in the Quality Control Framework**  
*Michele Gallo, Luigi D’Ambra*  
...
  
**Non Parametric Control Chart by Multivariate Additive Partial Least Squares via Spline**  
*Rosaria Lombardo, Amalia Vanacore, Jean-François Durand*  
...
  
**Simple Non Symmetrical Correspondence Analysis**  
*Antonello D’Ambra, Pietro Amenta, Valentin Rousson*  
...
  
**Factorial Analysis of a Set of Contingency Tables**  
*Amaya Zárraga, Beatriz Goitisolo*  
...

### Part IV Analysis of Complex Data

**Graph Mining: Repository vs. Canonical Form**  
*Christian Borgelt and Mathias Fiedler*  
...
  
**Classification and Retrieval of Ancient Watermarks**  
*Gerd Brunner, Hans Burkhardt*  
...
  
**Segmentation and Classification of Hyper-Spectral Skin Data**  
*Hannes Kazianka, Raimund Leitner, Jürgen Pilz*  
...
  
**FSMTree: An Efficient Algorithm for Mining Frequent Temporal Patterns**  
*Steffen Kempe, Jochen Hipp, Rudolf Kruse*  
...
  
**A Matlab Toolbox for Music Information Retrieval**  
*Olivier Lartillot, Petri Toiviainen, Tuomas Eerola*  
...
  
**A Probabilistic Relational Model for Characterizing Situations in Dynamic Multi-Agent Systems**  
*Daniel Meyer-Delius, Christian Plagemann, Georg von Wichert, Wendelin Feiten, Gisbert Lawitzky, Wolfram Burgard*  
...
  
**Applying the $Q_n$ Estimator Online**  
*Robin Nunkesser, Karen Schettlinger, Roland Fried*  
...
A Comparative Study on Polyphonic Musical Time Series Using MCMC Methods
Katrin Sommer, Claus Weihs ........................................ 285

Collective Classification for Labeling of Places and Objects in 2D and 3D Range Data
Rudolph Triebel, Óscar Martínez Mozos, Wolfram Burgard ..................... 293

Lag or Error? - Detecting the Nature of Spatial Correlation
Mario Larch, Janette Walde ............................................ 301

Part V Exploratory Data Analysis and Tools for Data Analysis

Urban Data Mining Using Emergent SOM
Martin Behnisch, Alfred Ultsch ........................................ 311

KNIME: The Konstanz Information Miner
Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, Bernd Wiswedel ........................................................ 319

A Pattern Based Data Mining Approach
Boris Delibašić, Kathrin Kirchner, Johannes Ruhland ......................... 327

A Framework for Statistical Entity Identification in R
Michaela Denk ............................................................. 335

Combining Several SOM Approaches in Data Mining: Application to ADSL Customer Behaviours Analysis
Francoise Fessant, Vincent Lemaire, Fabrice Clérot ........................... 343

On the Analysis of Irregular Stock Market Trading Behavior
Markus Franke, Bettina Hoser, Jan Schröder .................................. 355

A Procedure to Estimate Relations in a Balanced Scorecard

The Application of Taxonomies in the Context of Configurative Reference Modelling
Ralf Knackstedt, Armin Stein ............................................. 373

Two-Dimensional Centrality of a Social Network
Akinori Okada ............................................................... 381

Benchmarking Open-Source Tree Learners in R/RWeka
Michael Schauerhuber, Achim Zeileis, David Meyer, Kurt Hornik ........ 389
From Spelling Correction to Text Cleaning – Using Context Information
Martin Schierle, Sascha Schulz, Markus Ackermann .......................... 397

Root Cause Analysis for Quality Management
Christian Manuel Strobel, Tomas Hrycej ........................................ 405

Finding New Technological Ideas and Inventions with Text Mining and
Technique Philosophy
Dirk Thorleuchter ................................................................. 413

Investigating Classifier Learning Behavior with Experiment Databases
Joaquin Vanschoren, Hendrik Blockeel .......................................... 421

Part VI Marketing and Management Science

Conjoint Analysis for Complex Services Using Clusterwise Hierarchical
Bayes Procedures
Michael Brusch, Daniel Baier .................................................... 431

Building an Association Rules Framework for Target Marketing
Nicolas March, Thomas Reutterer ................................................. 439

AHP versus ACA – An Empirical Comparison
Martin Meißner, Sören W. Scholz, Reinhold Decker .......................... 447

On the Properties of the Rank Based Multivariate Exponentially
Weighted Moving Average Control Charts
Amor Messaoud, Claus Weihs .................................................... 455

Are Critical Incidents Really Critical for a Customer Relationship? A
MIMIC Approach
Marcel Paulssen, Angela Sommerfeld ............................................. 463

Heterogeneity in the Satisfaction-Retention Relationship – A
Finite-mixture Approach
Dorian Quint, Marcel Paulssen ..................................................... 471

An Early-Warning System to Support Activities in the Management of
Customer Equity and How to Obtain the Most from Spatial Customer
Equity Potentials
Klaus Thiel, Daniel Probst ......................................................... 479

Classifying Contemporary Marketing Practices
Ralf Wagner ............................................................................. 489
Part VII Banking and Finance

Predicting Stock Returns with Bayesian Vector Autoregressive Models
Wolfgang Bessler, Peter Lückoff ........................................... 499

The Evaluation of Venture-Backed IPOs – Certification Model versus Adverse Selection Model, Which Does Fit Better?
Francesco Gangi, Rosaria Lombardo ..................................... 507

Using Multiple SVM Models for Unbalanced Credit Scoring Data Sets
Klaus B. Schebesch, Ralf Stecking ........................................... 515

Part VIII Business Intelligence

Comparison of Recommender System Algorithms Focusing on the New-item and User-bias Problem
Stefan Hauger, Karen H. L. Tso, Lars Schmidt-Thieme .................. 525

Collaborative Tag Recommendations
Leandro Balby Marinho and Lars Schmidt-Thieme .......................... 533

Applying Small Sample Test Statistics for Behavior-based Recommendations
Andreas W. Neumann, Andreas Geyer-Schulz ............................... 541

Part IX Text Mining, Web Mining, and the Semantic Web

Classifying Number Expressions in German Corpora
Irene Cramer, Stefan Schacht, Andreas Merkel ............................. 553

Non-Profit Web Portals - Usage Based Benchmarking for Success Evaluation
Daniel Delić, Hans-J. Lenz ...................................................... 561

Text Mining of Supreme Administrative Court Jurisdictions
Ingo Feinerer, Kurt Hornik ...................................................... 569

Supporting Web-based Address Extraction with Unsupervised Tagging
Berenike Loos, Chris Biemann .................................................. 577

A Two-Stage Approach for Context-Dependent Hypernym Extraction
Berenike Loos, Mario DiMarzo .................................................. 585

Analysis of Dwell Times in Web Usage Mining
Patrick Mair, Marcus Hudec ..................................................... 593
New Issues in Near-duplicate Detection
Martin Potthast, Benno Stein .................................................. 601

Comparing the University of South Florida Homograph Norms with Empirical Corpus Data
Reinhard Rapp ................................................................. 611

Content-based Dimensionality Reduction for Recommender Systems
Panagiotis Symeonidis ....................................................... 619

Part X Linguistics

The Distribution of Data in Word Lists and its Impact on the Subgrouping of Languages
Hans J. Holm ....................................................................... 629

Quantitative Text Analysis Using L-, F- and T-Segments
Reinhard Köhler, Sven Naumann ........................................... 637

Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering
John Nerbonne, Peter Kleiweg, Wilbert Heeringa, Franz Manni .......... 647

Structural Differentiae of Text Types – A Quantitative Model
Olga Pustylnikov, Alexander Mehler ...................................... 655

Part XI Data Analysis in Humanities

Scenario Evaluation Using Two-mode Clustering Approaches in Higher Education
Matthias J. Kaiser, Daniel Baier ............................................. 665

Visualization and Clustering of Tagged Music Data
Pascal Lehwark, Sebastian Risi, Alfred Ultsch .......................... 673

Effects of Data Transformation on Cluster Analysis of Archaeometric Data
Hans-Joachim Mucha, Hans-Georg Bartel, Jens Dolata ................ 681

Fuzzy PLS Path Modeling: A New Tool For Handling Sensory Data
Francesco Palumbo, Rosaria Romano, Vincenzo Esposito Vinzi .......... 689

Automatic Analysis of Dewey Decimal Classification Notations
Ulrike Reiner ....................................................................... 697
A New Interval Data Distance Based on the Wasserstein Metric
Rosanna Verde, Antonio Irpino ......................................................... 705

Keywords ................................................................. 713

Author Index ............................................................. 717