Studies in Classification, Data Analysis, and Knowledge Organization

---

Titles in the Series

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 (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. Sokołowski, 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

Claus Weihs · Wolfgang Gaul
Editors

Classification – the Ubiquitous Challenge

Proceedings of the 28th Annual Conference
of the Gesellschaft für Klassifikation e.V.
University of Dortmund, March 9–11, 2004

With 181 Figures and 108 Tables

Springer

Professor Dr. Claus Weihs
Universität Dortmund
Fachbereich Statistik
44221 Dortmund
weihs@statistik.uni-dortmund.de

Professor Dr. Wolfgang Gaul
Universität Karlsruhe (TH)
Institut für Entscheidungstheorie
und Unternehmensforschung
76128 Karlsruhe
wolfgang.gaul@wiwi.uni-karlsruhe.de

ISSN 1431-8814
ISBN 3-540-25677-6 Springer-Verlag Berlin Heidelberg New York

Library of Congress Control Number: 2005927145

Springer · Part of Springer Science+Business Media

springeronline.com

Printed in Germany

Softcover-Design: Erich Kirchner, Heidelberg

SPIN 11419365 43/3153 – 5 4 3 2 1 0 – Printed on acid-free paper

Preface

This volume contains revised versions of selected papers presented during the 28th Annual Conference of the Gesellschaft für Klassifikation (GfKl), the German Classification Society. The conference was held at the Universität Dortmund in Dortmund, Germany, in March 2004. Wolfgang Gaul chaired the program committee, Claus Weihs and Ernst-Erich Doberkat were the local organizers. Patrick Groenen, Iven van Mechelen, and their colleagues of the Vereniging voor Ordinatie en Classificatie (VOC), the Dutch-Flemish Classification Society, organized special VOC sessions.

The program committee recruited 17 notable and internationally renowned invited speakers for plenary and semi-plenary talks on their current research work regarding classification and data analysis methods as well as applications. In addition, 172 invited and contributed papers by authors from 18 countries were presented at the conference in 52 parallel sessions representing the whole field addressed by the title of the conference "Classification: The Ubiquitous Challenge". Among these 52 sessions the VOC organized sessions on Mixture Modelling, Optimal Scaling, Multiway Methods, and Psychometrics with 18 papers. Overall, the conference, which is traditionally designed as an interdisciplinary event, again provided an attractive forum for discussions and mutual exchange of knowledge.

Besides the results obtained in the fundamental subjects Classification and Data Analysis, the talks in the applied areas focused on various application topics. Moreover, along with the conference a competition on "Social Milieus in Dortmund", co-organized by the city of Dortmund, took place. Hence the presentation of the papers in this volume is arranged in the following parts:

I. (Semi-)Plenary Presentations
II. Classification and Data Analysis
III. Applications, and
IV. Contest: Social Milieus in Dortmund.

The part on applications has sub-chapters according to the different application fields Archaeology, Astronomy, Bio-Sciences, Electronic Data and Web, Finance and Insurance, Library Science and Linguistics, Macro-Economics, Marketing, Music Science, and Quality Assurance. Within (sub-)parts papers are mainly arranged in alphabetical order with respect to (first) author's names.

I.

Plenary and semi-plenary lectures enclose both conceptual and applied papers. Among the conceptual papers Erosheva and Fienberg present a fully

Bayesian approach to soft clustering and classification within a general framework of mixed membership, Friendly introduces the Milestones Project on documentation and illustration of historical developments in statistical graphics, Hornik discusses consensus partitions particularly when applied to analyze the structure of cluster ensembles, Kiers gives an overview of procedures for constructing bootstrap confidence intervals for the solutions of three-way component analysis techniques, Pahl argues that a classification framework can organize knowledge about software components' characteristics, and Uter and Gefeller define partial attributable risk as a unique solution for allocating shares of attributable risk to risk factors. Within the applied papers Beran presents preprocessing of musical data utilizing prior knowledge from musicology, Fischer et al. introduce a method for the prediction of spatial properties of molecules from the sequence of amino acids incorporating biological background knowledge, Grzybek et al. discuss how far word length may contribute to quantitative typology of texts, and Snoek and Worring present the Time Interval Multimedia Event framework as a robust approach for classification of semantic events in multimodal soccer video.

II.

The second part of this volume is concerned with methodological progress in classification and data analysis and methods presented cover a variety of different aspects.

In the **Classification** part, more precise confidence intervals for the parameters of latent class models using the bootstrap method are proposed (Dias), as well as a method of feature selection for ensembles that significantly reduces the dimensionality of subspaces (Gatnar), and a sensitive two-stage classification system for the detection of events in spite of a noisy background in the processing of thousands of images in a few seconds (Hader and Hamprecht). Variants of bagging and boosting are discussed, which make use of an ordinal response structure (Hechenbichler and Tutz), a methodology for exploring two quality aspects of cluster analyses, namely separation and homogeneity of clusters (Hennig), and a comparison of Adaboost to Arc-x(h) for different values of h in the subsampling of binary classification data is carried out (Khanchel and Limam). The method of distance-based discriminant analysis (DDA) is introduced finding a linear transformation that optimizes an asymmetric data separability criterion via iterative majorization and the necessary number of discriminative dimensions (Kosinov et al.), an efficient hybrid methodology to obtain CHAID tree segments based on multiple dependent variables of possibly different scale types is proposed (Magidson and Vermunt), and possibilities of defining the expectation of p-dimensional intervals (Nordhoff) are described. Design of experiments is introduced into variable selection in classification (Pumplün et al.), as well as the KMC/EDAM method for classification and visualization as an alternative to Kohonen Self-Organizing Maps (Raabe et al.). A clustering of variables approach extended to situations with missing data based on different imputation methods (Sahmer et al.), a method for binary online-classification incorporating temporal distributed information (Schäfer et al.), and a concept of characteristic regions and a new method, called DiSCo, to simultaneously classify and visualize data (Szepannek and Luebke) are described. The part concludes with two papers discussing multivariate Pareto Density Estimation (PDE), based on information optimality, for data sets containing clusters (Ultsch) and an extension of standard latent class or mixture models that can be used for the analysis of multilevel and repeated measures data (Vermunt and Madgison).

The part on **Data Analysis** starts with papers proposing a robust procedure for estimating a covariance matrix under conditional independence restrictions in graphical modelling (Becker) and a new approach to find principal curves through a multidimensional, possibly branched, data cloud (Einbeck et al.). A three–way multidimensional scaling approach developed to account for individual differences in the judgments about objects, persons or brands (Krolak-Schwerdt), and the Time Series Knowledge Mining (TSKM) framework to discover temporal structures in multivariate time series based on the Unification-based Temporal Grammar (UTG) (Mörchen and Ultsch) are introduced. A framework for the comparison of the information in continuous and categorical data (Nishisato) and an external analysis of two-mode three-way asymmetric multidimensional scaling for the disclosure of asymmetry (Okada and Imaizumi) are presented. Finally, nonparametric regression with the Relevance Vector Machine under inclusion of covariate measurement error (Rummel) is described.

III.

In the third part of this volume all contributions are also related to applications of classification and data analysis methods but structured by their application field.

Two papers deal with applications in **Archaeology**. The first is a historical overview (Ihm) over early publications about formal methods on seriation of archaeological finds, in the second article some cluster analysis models including different data transformations in order to differentiate between brickyards of different areas on the basis of chemical analysis are investigated (Mucha et al.).

Another two papers (both by Bailer-Jones) discuss applications in **Astronomy**. A brief overview of the upcoming Gaia astronomical survey mission, a major European project to map and classify over a billion stars in our Galaxy, and an outline of the challenges are given in the first paper while in the second a novel method based on evolutionary algorithms for designing filter systems for astronomical surveys in order to provide optimal data on stars and to determine their physical parameters is introduced.

The articles with applications in the **Bio-Sciences** all deal with enzyme, DNA, microarray, or protein data, except the presentation of results of a systematic and quantitative comparison of pattern recognition methods in the analysis of clinical magnetic resonance spectra applied to the detection of brain tumor (Menze et al.). The Generative Topographic Mapping approach as an alternative to SOM for the analysis of microarray data (Grimmenstein et al.) and a finite conservative test for detecting a change point in a binary sequence with Markov dependence and applications in DNA analysis (Krauth) are proposed as well as a new algorithm for finding similar substructures in enzyme active sites with the use of emergent self-organizing neural networks (Kupas and Ultsch). How the feature selection procedure "Significance Analysis of Microarrays" (SAM) and the classification method "Prediction Analysis of Microarrays" (PAM) can be applied to "Single Nucleotide Polymorphism" (SNP) data is explained (Schwender) as well as that using relative differences (RelDiff) instead of LogRatios for cDNA microarray analysis solves several problems like unlimited ranges, numerical instability and rounding errors (Ultsch). Finally, a novel method, PhyNav, to reconstruct the evolutionary relationship from really large DNA and protein datasets is introduced applying the maximum likelihood principle (Vinh et al.).

Among the contributions on applications to **Electronic Data and Web** one paper discusses the application of clustering with restricted random walks on library usage histories in large document sets containing millions of objects (Franke and Thede). In the other four papers different aspects of web-mining are tackled. A tool is described assisting users of online news web-sites in order to reduce information overload (Bomhardt and Gaul), benchmarks are offered with respect to competition and visibility indices as predictors for traffic in web-sites (Schmidt-Mänz and Gaul), an algorithm is introduced for fuzzy two-mode clustering that outperforms collaborative filtering (Schlecht and Gaul), and visualizations of online search queries are compared to improve understanding of searching, viewing, and buying behavior of online shoppers and to further improve the generation of recommendations (Thoma and Gaul).

Two of the articles on **Finance and Insurance** deal with insurance problems: A strategy based on a combination of support vector regression and kernel logistic regression to detect and to model high-dimensional dependency structures in car insurance data sets is proposed (Christmann) and support vector machines are compared to traditional statistical classification procedures in a life insurance environment (Steel and Hechter). Applications in Finance deal with evaluation of global and local statistical models for complex data sets of credit risks with respect to practical constraints and asymmetric cost functions (Schwarz and Arminger), show how linear support vector machines select informative patterns from a credit scoring data pool serving as inputs for traditional methods more familiar to practitioners (Stecking and Schebesch), analyze the question of risk budgeting in continuous time (Straßberger), and formulate a one-factor model for the correlation between probabilities of default across industry branches, comparing it to more traditional methods on the basis of insolvency rates for Germany (Weißbach and Rosenow).

Besides one contribution on **Library Science** where it is argued that the history of classification is intensively linked to the history of library science (Lorenz) the volume encloses five papers on applications in **Linguistics**. It is shown that one meta-linguistic relation suffices to model the concept structure of the lexicon making use of intensional logic (Bagheri), that improvements of the morphological segmentation of words using classical distributional methods are possible (Benden), and that in Russian texts (letters and poems by three different authors) word length is a characteristic of genre, rather than of authorship (Kelih et al.). A validation method of cluster analysis methods concerning the number and stability of clusters is described with the help of an application in linguistics (Mucha and Haimerl), clustering of word contexts is used in a large collection of texts for word sense induction, i.e. automatic discovery of the possible senses for a given ambiguous word (Rapp), and formal graphs that structure a document-related information space by using a natural language processing chain and a wrapping procedure are proposed (Rist).

There are three papers with applications in **Macro-Economics**, two of them dealing with the comparison of economic structures of different countries. The sensitivity of economic rankings of countries based on indicator variables is discussed (Berrer et al.), structural variables of the 25 member European Union are analyzed and patterns are found to be quite different between the 15 current and the 10 new members (Sell), while the question whether methods measuring (relative) importance of variables in the context of classification allow interpretation of individual effects of highly correlated economic predictors for the German business cycle (Enache and Weihs) is tackled in a more methods-based contribution.

Within the **Marketing** applications one article shows by means of an intercultural survey (Bauer et al.) that the cyber community is not a homogeneous group since online consumers can be classified into the three clusters: “risk avers doubters”, “open minded online-shoppers” and “reserved information seekers”. Two papers deal with reservation prices. A novel estimation procedure of reservation prices combining adaptive conjoint analysis with a choice task using individually adapted price scales is proposed (Breidert et al.), and an explicit evaluation of variants of conjoint analysis together with two types of data collection is described for the detection of reservation prices of product bundles applied to a seat system offered by a German car manufacturer (Stauß and Gaul).

Music Science is an application field that is present at GfKl conferences for the first time. In this volume one paper deals with time series analysis, the other five papers apply classification methods. A new algorithm structure is introduced for feature extraction from time series, its efficiency is proofed, and illustrated by different classification tasks for audio data (Mierswa). Classification methods are used to show that the more the musical sound is unstable in time domain the more pitch bending is admitted to the musician expressing emotions by music (Fricke). Classification rules for quality classes of “sight reading” (SR) are derived (Kopiez et al.) based on indicators of piano practice, mental speed, working memory, inner hearing etc. as well as the total SR performance of 52 piano students. Classification rules are also found for digitized sounds played by different instruments based on the Hough-transform (Röver et al.). Finally, classifications of possibly overlapping drum sounds by linear support vector machines (Van Steelant et al.) and of singers and instruments into high or low musical registers only by means of timbre, i.e. after elimination of pitch information, are proposed (Weihs et al.).

Applications in **Quality Assurance** include one methodological paper (Jessenberger and Weihs) which proposes the use of the expected value of the so-called desirability function to assess the capability of a process. The other papers discuss different statistical aspects of a deep hole drilling process in machine building. The Lyapunov exponent is used for the discrimination between well-predictable and not-well-predictable time series with applications in quality control (Busse). Two multivariate control charts to monitor the drilling process in order to prevent chatter vibrations and to secure production with high quality are proposed (Messaoud et al.) as well as a procedure to assess the changing amplitudes of relevant frequencies over time based on the distribution of periodogram ordinates (Theis and Weihs).

IV.

The fourth part of this volume starts with an introduction to the competition on “Social Milieus in Dortmund” (Sommerer and Weihs). Moreover, the best three papers of the competition by Scheid, by Schäfer and Lemm, and by Röver and Szepannek appear in this volume. We would like to thank the head of the “dortmund-project”, Udo Mager, and the head of the Fachbereich “Statistik und Wahlen” of the City of Dortmund, Ernst-Otto Sommerer, for their kind support.

The conference owed much to its sponsors (in alphabetical order)

- Deutsche Forschungsgemeinschaft (DFG), Bonn,
- dortmund-project, Dortmund,
- Fachbereich Statistik, Universität Dortmund, Dortmund,
- Landesbeauftragter für die Beziehungen zwischen den Hochschulen in NRW und den Beneluxstaaten,
- Novartis, Basel, Switzerland,
- Roche Diagnostics, Penzberg,
- sas Deutschland, Heidelberg,
- Sonderforschungsbereich 475, Dortmund,
- Springer-Verlag, Heidelberg,
- Universität Dortmund, and
- John Wiley and Sons, Chicester, UK.

who helped in many ways. Their generous support is gratefully acknowledged.

Additionally, we wish to express our gratitude to 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. Furthermore, we thank the reviewers for their careful reviews of the originally submitted papers, and in this way, for their support in selecting the best papers for this publication.

We would like to emphasize the outstanding work of Uwe Ligges and Nils Raabe who did an excellent job in organizing the program of the conference and the refereeing process as well as in preparing the abstract booklet and this volume, respectively. We also wish to thank our colleague Prof. Dr. Ernst-Erich Doberkat, Fachbereich Informatik, University Dortmund, for co-organizing the conference, and the Fachbereich Statistik of the University Dortmund for all the support, in particular Anne Christmann, Dr. Daniel Enache, Isabelle Grimmenstein, Dr. Sonja Kuhnt, Edelgard Kürbis, Karsten Luebke, Dr. Constanze Pumplün, Oliver Sailer, Roland Schultze, Sibylle Sturtz, Dr. Winfried Theis, Magdalena Thöne, and Dr. Heike Trautmann as well as other members and students of the Fachbereich for helping to organize the conference and making it a big success, and Alla Stankjawitschene and Dr. Stefan Dißmann from the Fachbereich Informatik for all they did in organizing all financial affairs.

Finally, we want to thank Christiane Beisel and Dr. Martina Bihn of Springer-Verlag, Heidelberg, for their support and dedication to the production of this volume.

Dortmund and Karlsruhe, *Claus Weihs, Wolfgang Gaul*
April 2005

Contents

Part I. (Semi-) Plenary Presentations

Part II. Classification and Data Analysis

Classification

Data Analysis

Part III. Applications

Archaeology

Astronomy

Bio-Sciences

Electronic Data and Web

Finance and Insurance

Library Science and Linguistics

Macro-Economics

Marketing

Music Science

Quality Assurance

Part IV. Contest: Social Milieus in Dortmund

Part I

(Semi-) Plenary Presentations