Preface

The 25th annual conference of the Gesellschaft für Klassifikation e.V. (German Classification Society) took place at the University of Munich between March 14-16, 2001. Within the scope of **Exploratory Data Analysis in Empirical Research** scientists and practitioners discussed recent developments in this field and established cross-disciplinary cooperations in their own fields of research. The scientific program of the conference included 18 plenary or semiplenary lectures and about 130 presentations in more than 20 special sections. Furthermore, publishing companies informed the participants about current software products and books at several exhibition booths.

The peer–reviewed papers are presented in 4 chapters as follows:

1. Classification, Data Analysis and Statistics
2. Web Mining, Data Mining and Computer Science
3. Medicine, Biological Sciences, Health Care
4. Marketing, Finance, and Management Science

In contrast to the alphabetical order in each of the four chapters, the following overview is arranged according to textual considerations.

Several problems of **Classification**, **Data Analysis** and **Statistics** are considered in 23 papers of chapter 1.

The paper of F. DOMENACH and B. LECLERC deals with the notion of Galois connection which is proved to be a fundamental tool in various situations of classification theory. A new clustering method based on the notion of biclosed relation is presented by the authors. E. GODEHARDT and J. JAWORSKI describe two models which use random bipartite graphs to provide clusters and classifications. The authors derive mathematical properties, characterizations and asymptotic results for their models. P. BRITO and F.A.T. DE CARVALHO introduce an approach useful for clustering purposes, taking into account constraints on probabilistic data. In this context they propose new generality measures in order to compare couples of constrained probabilistic objects, together with an appropriate extension of generalization operators to get new constrained probabilistic objects. The main focus of the contribution of A. GRAUEL, I. RENNERS and E. SAAVEDRA is the presentation of a general methodology for structure optimization of fuzzy classifiers in classification techniques. The paper of M. CSERNEL and F.A.T. DE CARVALHO is embedded in the field of symbolic data analysis. The authors propose a special decomposition method of a symbolic data table which splits the table into subtables. It is shown that the decomposition reduces computation time from exponential to polynomial with bounded factors for memory space. M. KREUSELER, T. NOCKE and H. SCHUMANN investigate the combination of numerical and visual exploration techniques focused on cluster analysis of multidimensional data. They describe new visualization approaches and selected clustering techniques and discuss the major features of them. The paper of D. WISHART addresses practical issues in k-means cluster analysis or segmentation with mixed types of variables and missing values. A more general k-means clustering procedure is developed that is suitable for use with very large datasets.

The paper of C. WEIHS and U.M. SONDHAUSS considers the interpretability of partitions generated by classification rules with severe concern for the adequacy for human understanding, the mental fit. The authors discuss various criteria introduced in relation to mental fit in literature and derive a general criterion for the interpretability of partitions generated by classification rules. Using hierarchical methods, the main focus of S. UEDA'S and Y. ITOH'S contribution is to show that languages can be mainly divided into prepositional and postpositional or adpositional languages with numerals and nouns as second most important clustering variables.

J. KRAUTH considers clusters resulting from connecting the nearest neighbours of points generated on the real line by a Poisson process. His particular interest is the distribution of the number of nearest neighbour clusters which can be used to test the hypothesis of no clustering. Furthermore, an approximation of the exact distribution of the test statistics is given. The paper of M.T. GALLEGOS is dealing with clustering objects in the presence of outliers. The author develops an estimator which simultaneously detects outliers and partitions. Another algorithm is proposed which approximates the estimator. G.J. MCLACHLAN, S.K. NG and D. PEEL consider several problems using mixture models for clustering multivariate data. They report some recent results on speeding up the fitting process by an EM–algorithm. Furthermore, the problem of clustering high–dimensional data by use of the factor analyzers model is considered. A. DOUGARJAPOV and G. LAUSEN study the problem of matching time series. They present a technique for evaluating a classifier which allows a reasonable choice of attributes to describe time points. As an evaluation function of the classifier they present the domain error.

Based on the generalized correlation coefficient introduced by Kendall K. JAJUGA, M. WALESIAK and A. BAK present a general distance measure for ordinal, interval and ratio scaled data. Y. TANAKA, F. ZHANG and W. YANG discuss a sensitivity analysis for principal component and canonical correlation analyses based on Cook's local influence. They compare their results with those obtained by procedures based on the influence function approach. Using an additive conjoint measurement

A. Okada analyzes social network data. In contrast to many other approaches his procedure applies not only to binary but also to discrete or continuously valued data as well as to asymmetric relations. In this context, an application to interpersonal attraction among students is presented. P.J.F. Groenen and J. Poblome propose a constrained correspondence analysis for seriation of archaeological artefactual assemblages. The method is applied to sagalassos ceramic tablewares data.

Using computer intensive Bayesian methods such as Gibbs sampling or Markov Chain Monte Carlo R. Winkelmann discusses in a survey paper advances in analyzing complex count data models. The paper of L. Fahrmeir, C. Gössl and A. Hennerfeind considers alternatives of spatial magnetic resonance imaging (MRI) priors in functional MRI's which are expected to have better edge preserving properties. The authors study the performance of functional MR priors with random weights using simulated and real data. They show improved edge preserving properties for the fitted functional MRI, however the edge preserving properties are lost for the fitted standardized functional MRI. An extension of the so–called Sliced Inverse Regression (SIR) to dynamic models is discussed by C. Becker and R. Fried. They discuss the general application of dynamic SIR, the detection of relevant directions and relations among the variables as well as the comparison with other methods in this context like graphical modeling. Th. Otter, R. Tüchler and S. Frühwirth–Schnatter present a Bayesian analysis of the latent class model using a new approach towards MCMC estimation in the context of mixture models. Conjoint data from the Austrian mineral water market serves to illustrate the method. S. Lang, P. Kragler, G. Haybach and L. Fahrmeir provide a Bayesian semiparametric approach which permits to simultaneously incorporate effects of space, time and further covariates within a joint model for analyzing the claims process in non–life insurances. The authors apply the approach to analyze costs of hospital treatment and accommodation using a large data set from a German health insurance company.

Chapter 2 includes 9 contributions dealing with **Web Mining**, **Data Mining** and **Computer Science** in a wide sense.

In a survey paper J.R. Punin, M.S. Krishnamoorthy and M.J. Zaki propose the use of a graph description language, XCMML and LOGML, especially developed as a tool for representing web navigation for discovering navigational patterns. F. Säuberlich and K.-P. Huber suggest a combination of different data mining techniques to analyze common log files of web users to find out structural problems in web design, typical navigation paths and to improve the web design. Th. Liehr discusses several methods for the imputation of missing values in the course of data preparation in large real–world data mining projects.

Before clustering T.A. Runkler and J.C. Bezdek transform considered text–data to relational data using Levenstein distance. These relational data can be clustered by relational ACS (RACE) providing keywords as the RACE cluster centers for further use. In a survey paper D. Saad uses methods adopted from statistical physics to analyze the dynamics of online learning in multilayer neuronal networks. The analysis is based on monitoring a set of macroscopic variables from which the generalisation error can be calculated. Two approaches for document retrieval which group pre–processed documents are presented in the paper of A. Nürnberger, A. Klose, R. Kruse, G. Hartmann and M. Richards. The authors apply self–organizing maps which help the user to navigate through similar documents.

One of the major problems in developing educational software is the appropriate way to visualize complex data structures for learners; the paper of K. Friesen and H. Schmitz deals with the semantic modelling issues in cost accounting and describes the software environment of the solution.

Finally the consumption patterns for information goods are studied by W. Böhm, A. Geyer–Schulz, M. Hahsler and M. Jahn in the context of an information broker at a virtual university. The authors investigate whether Ehrenberg's repeat–buying theory can also be applied in electronic markets for information goods. In the field of automatic documentation and the combination of different documentation systems U. Rist presents a metadata management system with mapping tables to link the different documentation systems.

Applications of data analysis and data mining in **Medicine**, **Biological sciences** and **Health Economics** are discussed in 9 contributions of chapter 3.

Several applications of unsupervised neural networks or self–organizing maps in medicine are considered in the paper of G. Guimarães and W. Urfer. Especially, the authors focus on sleep apnea discovery, protein sequence analysis and tumor classification. For the analysis of cancer mortality data U. Schach makes an intrapolation from coarse time sequential data to the belonging small area data. The small area estimation uses special and temporal dependence structures based on a Bayesian hierarchical modelling approach. C. Vogel and O. Gefeller discuss the effects of misclassification on the estimation of measures of association in epidemiologic studies. The effects on the attributable risk and on the relative risk are compared. The paper of T. Hothorn, I. Pal, O. Gefeller, B. Lausen, G. Micholson and D. Paulus presents first results of the development of an automatic classification scheme of optic nerve head topography images for Glaucoma screening. Different linear and tree based discriminant techniques as well as stabilized methods are evaluated for the classification tasks. C. Ernst, G. Ernst and A. Szczesny estimate a learning curve related to knee replacement surgery from the data of surgeries performed in a German hospital. For this problem a regression model is applied. For constructing survival trees M. Radespiel–Tröger, T. Rabenstein, L. Höpfner and H.T. Schneider compare various split criteria for a special case of censored data. They use a likelihood model and Magee's R^2 as measures for the explained variation and estimate bootstrap confidence intervals of the explained variation.

The paper of A. Ziegler, O. Hartmann, I.R. Böddecker and H. Schäfer discusses recently developed techniques which might be alternatives for the identification of disease susceptibility genes and their function. They illustrate the absolute need for new statistical methods by means of two areas of current research, genome wide association analysis and gene expression groups.

To detect quality deficiencies in health care, J. Stausberg and Th. Albrecht propose the use of a common data mining tool. To receive useful results they found out that the underlying data have to be especially prepared for the OLAP–model. M. Staat considers the question to provide physicians with appropriate information which improves the cost–efficiency relation of their work. He uses a bench marking method based on data envelopment analysis.

Many problems of **Marketing**, **Finance** and **Management Science** make use of models and methods of numerical and statistical data analysis. The 12 papers of chapter 4 are dealing with these issues.

M. Meyer considers the problem of sequence mining from a computational and a marketing point of view. He explains differences between these and points out the necessity to develop new algorithms, focusing stronger on marketing aspects.

C. Bornemeyer and R. Decker identify key success factors of city marketing projects by using structural equation modeling and discriminant analysis. The paper of H.H. Bauer, M. Staat and M. Hammerschmidt offers an analytical framework for an integrated treatment of market positioning and benchmarking based on data envelopment analysis. Furthermore, an exploratory data mining approach is used. For predicting market shares of competing products in assumed market scenarios D. Baier and W. Polasek compare empirical–traditional estimation procedures of conjoint analysis to a Bayesian approach. M. Löffler presents a modification of the wellknown „tandem approach“ for segmentation which takes into account certain shortcomings. Subsequently he applies his approach to the German automobile market to reveal additional insights.

Within the framework of the „default–mode“ credit portfolio risk model R. Kiesel and U. Stadtmüller analyze the importance of factor correlation and granularity for the total risk of credit risky portfolios. S. Höse and S. Huschens deal with the problem of estimating stable default probabilities from individual credit scores without loosing information by aggregating individual to rating grades. For this purpose they propose a maximum likelihood estimator first for the binomial, second for the poisson model. The paper of N. Wagner focuses on the estimation of extreme parts of some special distributions with fat tails which exist in financial return distributions. In this context the problem of optimal subsamples is discussed. M. Pojarliev and W. Polasek use a VAR–GARCH model to predict the monthly returns of a portfolio on a stock market to find an appropriate portfolio as well as the kind of forecasts which improves the portfolio performance.

An optimization method is presented by R. Bennert and M. Missler–Behr to find promising portfolio positions for a reorganizing company. The method is based on a genetic algorithm using a cashflow index and a potential of success index for the overall fitness criterion. D. Kwiatkowska–Ciotucha, U. Zaluska and J. Dziechciarz study the attractiveness of manufactoring branches in Poland from the investor's point of view. They analyze several descriptive variables obtained from statistical reports.

The main results of the investigation on the cognitive representation of trait–descriptive terms, considered by S. Krolak–Schwerdt and B. Ganter clearly show that the choice of a particular measurement technique determines which organizational aspects of cognitive relations between traits become visible and whether subjects attention is directed either to the semantic components of the trait terms or to stereotypes and self concept features.

The editors of these proceedings are very indebted to all colleagues who chaired a session during the conference and/or reviewed some papers for this volume. We gratefully acknowledge the help and support given by the members of the scientific program committee as well as the active cooperation of all participants and authors. Finally we would like to emphasize the excellent work of all assistants and secretaries involved in the organization of the conference and preparation of this proceedings volume.

We hope that the presented volume will find interested readers and may encourage further research.

Augsburg and Munich, June 2002

O. Opitz, M. Schwaiger

Acknowledgements

Among our principal supporters we would like to highlight the role of **Bain & Company**, who had a major stake in ensuring the success of our Munich conference. **Bain & Company** is known to be one of the leading strategic management consultancies worldwide. Since its foundation nearly thirty years ago, it has successfully focused on helping its diverse client base of over 2000 organizations formulating and implementing business strategies. The capacity to output relevant, functioning and lasting business solutions demands thorough and continuous development of knowledge, which **Bain & Company** successfully complied with by generously supporting our conference. The company's patronage underlines its strong and credible interest in both the development of innovative approaches in business research as well as the exchange of ideas between theory and practice. In respect of the above we would once more like to express our true appreciation and gratitude to **Bain & Company**.

Furthermore, we gratefully take the opportunity to acknowledge the support by:

- Bayerische Hypo-und Vereinsbank AG
- SAS Deutschland
- NFO Infratest
- Booz · Allen & Hamilton
- DaimlerChrysler AG
- Siemens AG, München
- HUK–COBURG Versicherungen Bausparen
- H.F. & Ph.F. Reemtsma GmbH
- EADS Deutschland GmbH
- Gesellschaft von Freunden und Förderern der Universität München
- Bavarian State Ministry of Sciences, Research and the Arts

Contents