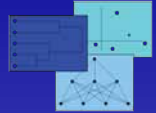


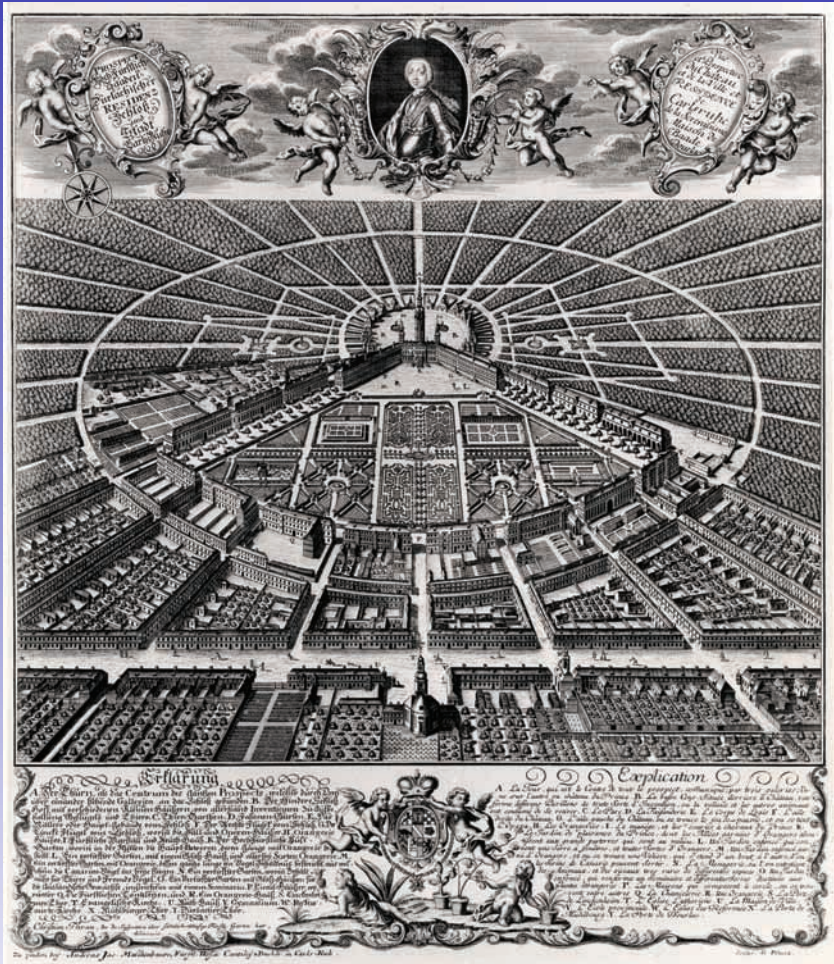


Japanese Classification Society

3rd German-Japanese Workshop



Program and Abstracts



For the Content

Prof. Yasumasa Baba,
The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo
190-8562, Japan.

Prof. Dr. Wolfgang Gaul,
Institut für Entscheidungstheorie und Unternehmensforschung, KIT-Campus
Süd, Postfach 69 80, D-76049 Karlsruhe.

Prof. Dr. Andreas Geyer-Schulz,
Lehrstuhl für Informationsdienste und elektronische Märkte, KIT-Campus
Süd, Postfach 69 80, D-76049 Karlsruhe.

Prof. Akinori Okada,
Graduate School of Management and Information Sciences, Tama University,
4-4-1 Hijirigaoka, Tama-shi, Tokyo 206-0022, Japan.

The support of the teams of Prof. Gaul and Prof. Geyer-Schulz w.r.t. the
preparation of this booklet is gratefully acknowledged.

Welcome!

Welcome to the 3rd German-Japanese Workshop to be held in Karlsruhe from July 20 to July 21, 2010.

After the first meeting in Tokyo in 2005 and the second one in Berlin in 2006 we are glad to have you here in Karlsruhe.

Compared to the previous workshop locations Karlsruhe is rather a small town.

Karlsruhe was founded in 1715 and became soon the official residence of Baden. It is located in the southern Rhine Valley between the mountains of the “Black Forest” on the German and the “Vosges” on the French side of the river. The unique layout of its city centre - radial in shape, similar to a fan - aroused intense international interest. Today, Karlsruhe is home to Germany’s two highest courts - the Federal Constitutional Court and the Federal Supreme Court - and to Universität Karlsruhe (TH), the oldest German technical university dating back to 1825 when a technical college commenced its teaching program. Now, Universität Karlsruhe (TH) and Forschungszentrum Karlsruhe have merged to KIT (Karlsruhe Institute of Technology) and belong to the “elite” universities of Germany.

It is easy to find one’s way in Karlsruhe as the Schloß (and that part of the university where the workshop takes place) is located in the middle of the city centre with its star-shaped designed street map.

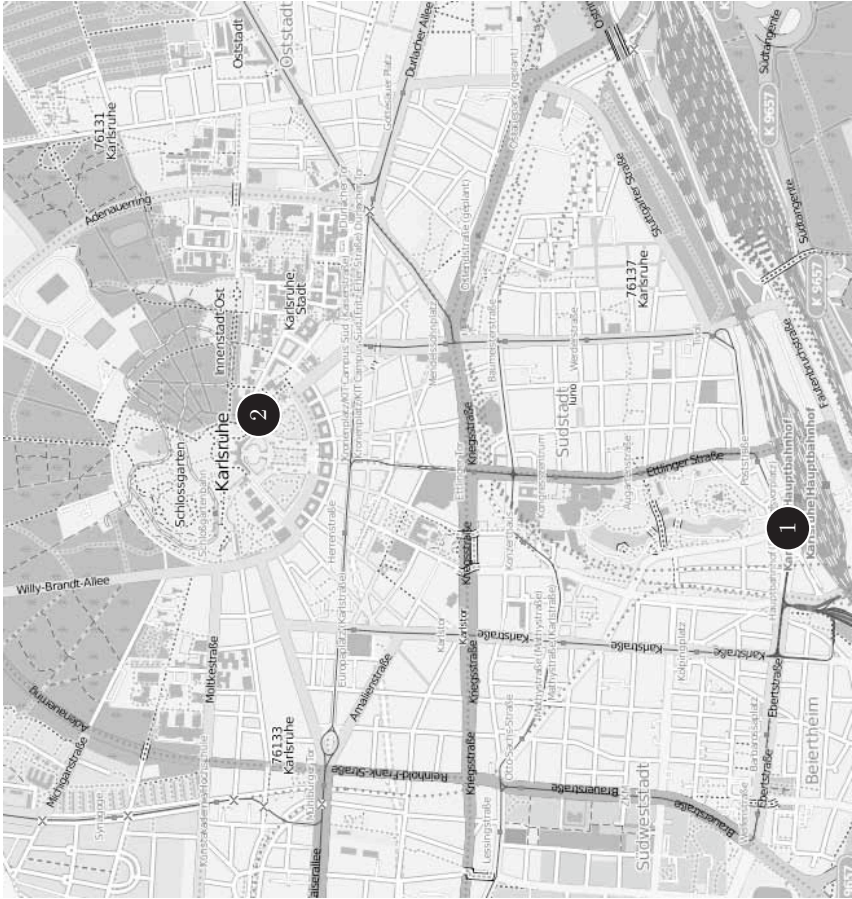
Karlsruhe is also known for its well-planned street car system. Included in your workshop material you will find a free ride ticket for Karlsruhe street cars that allows you to use the system on your own (valid for July 20, 21, 2010, within city boundaries).

Please, take into consideration that the this year’s annual conference of the Gesellschaft für Klassifikation (GfKl 2010) will overlap with our German-Japanese workshop.

We wish you a pleasant stay in Karlsruhe and a scientifically interesting workshop.

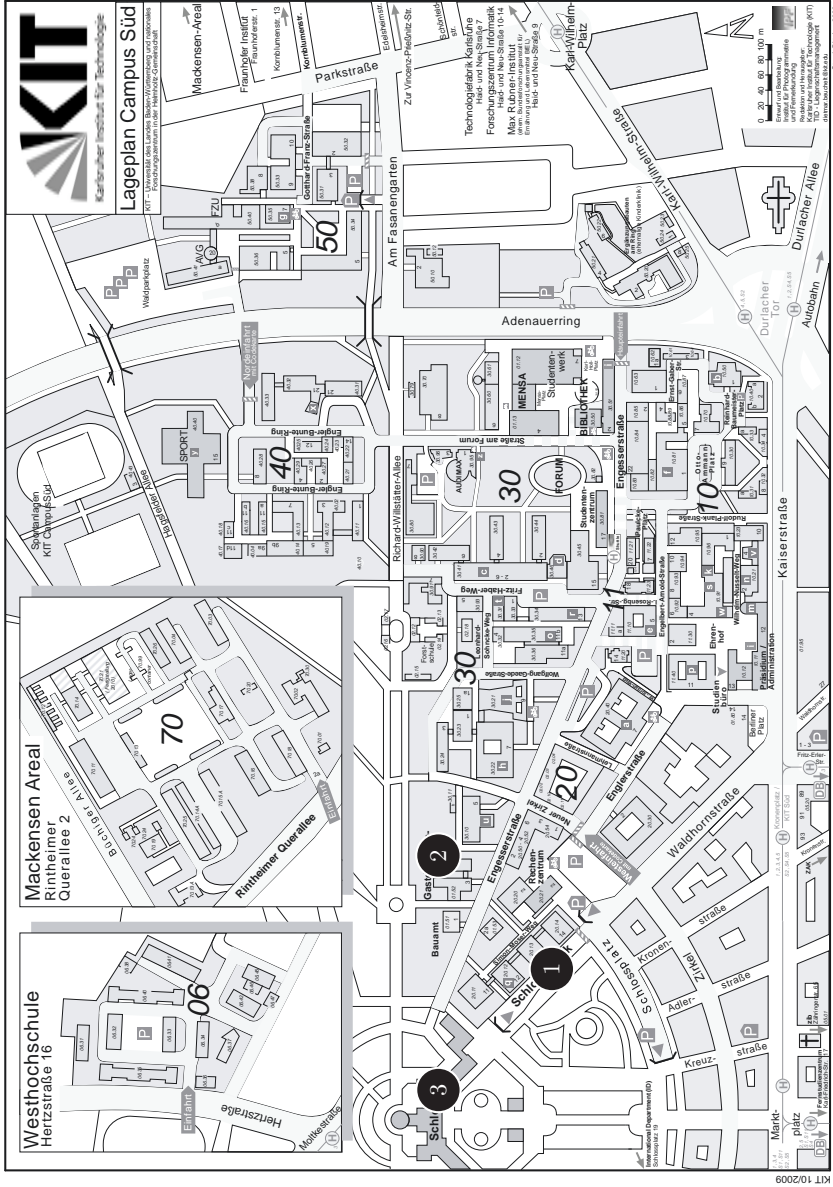
The Organizers

City Centre of Karlsruhe

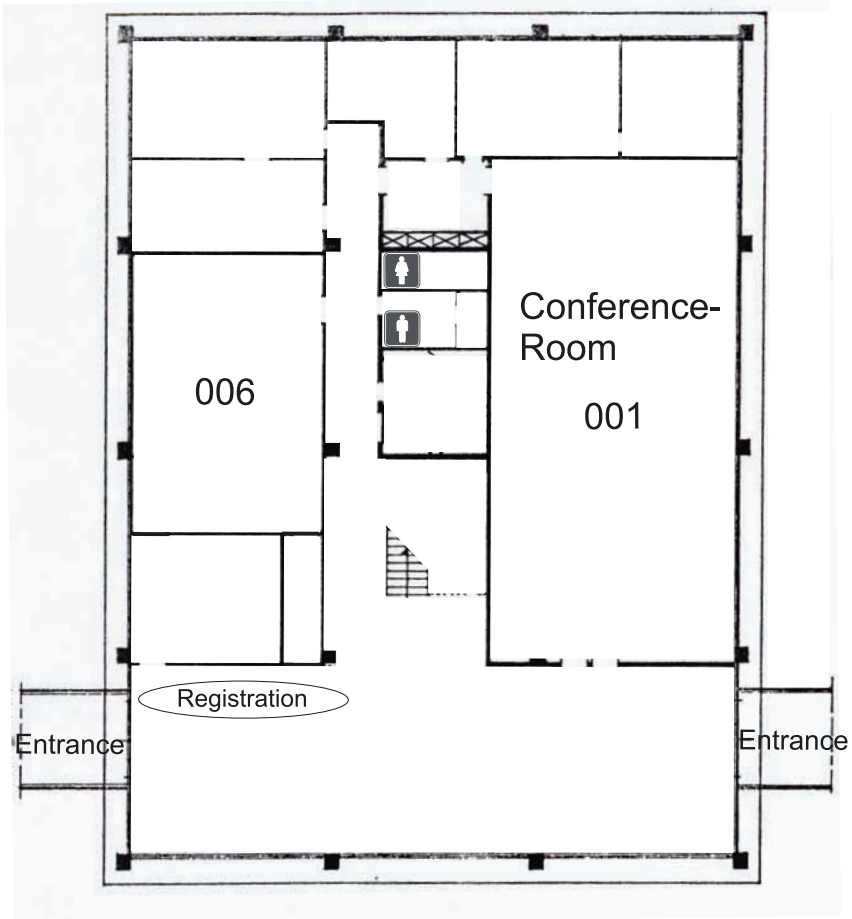


1 Central Railway Station 2 Conference Location

Map of the Campus of the KIT/University of Karlsruhe



- 1 Building 20.13
- 2 Gastdozentenhaus
- 3 Schloßcafe

Workshop Location**Gebäude 20.13 (Bau III)**

Registration starts on Tuesday morning (July 20, 2010) at 8.00 am.

Social Program

Lunch at Gastdozentenhaus (July 20, 2010)
12.00 - 14.00

Get-together at Schloßcafe (July 20, 2010)
together with GfKI 2010
19.00

Contents

Program Schedule	8
-------------------------	----------

Abstracts (in alphabetical order w.r.t. first author's name)	11
---	-----------

Optimal Interval-Valued Scaling of Successive Categories <i>Sayaka Arai, Hisao Miyano</i>	13
Continuous-discrete Transformation in Multivariate Data Analysis <i>Yasumasa Baba</i>	14
Image Clustering Algorithms for Marketing Purposes <i>Daniel Baier, Ines Daniel</i>	15
Clustering Methods for Time Series <i>Hans-Hermann Bock</i>	16
Consistency Improving of Pairwise Comparison AHP Data and Their Impact on Derived Weights <i>Dominic Gastes, Wolfgang Gaul</i>	18
The Randomized Greedy Modularity Clustering Algorithm and Its Relationship with Incomplete SAT-Solvers <i>Andreas Geyer-Schulz, Michael Ovelgönne</i>	19
A Discrete Decomposition of Preference Matrix by ADCLUS-type model <i>Tadashi Imaizumi</i>	20

Comparison of Regression-based Modeling Methods for Judgment Data	
<i>Sabine Krolak-Schwerdt, Thomas Hörstermann</i>	21
Empirical Studies on the Analysis of a Vast Amount of the Internet Traffic Data in Japan	
<i>Hiroyuki Minami</i>	22
Combination of Symbolic Data Analysis and Functional Data Analysis	
<i>Masahiro Mizuta</i>	23
Pairwise Data Clustering Accompanied by Validation and Visualisation	
<i>Hans-Joachim Mucha</i>	24
Socioeconomic and Gender Differences in Voluntary Participation in Japan	
<i>Miki Nakai</i>	26
Analysis of Conditional and Marginal Association in One-Mode Three-Way Proximity Data	
<i>Atsuho Nakayama</i>	27
Centrality of Two-Mode Two-Way Social Network	
<i>Akinori Okada</i>	28
Relational Time Series Classification	
<i>Christine Preisach, Lars Schmidt-Thieme</i>	29
Modal Interval-Valued Dissimilarity between Histogram-Valued Data	
<i>Yoshikazu Terada, Hiroshi Yadohisa</i>	30
Analysis of Brand Categories Using Purchase History Data for Brand Management: the Study of Category Composition and those Trends	
<i>Toyada, Yuki, Imaizumi, Tadashi</i>	31
Three-Way Individual Scaling and Clustering for Musical Structure and Its Application	
<i>Mitsuhiro Tsuji</i>	32
Identifying Consumer Segments from Online Product Reviews Using Finite Mixture Models	
<i>Michael Tuma, Reinhold Decker</i>	33

A Non-Parametric Analysis for a Questionnaire Survey <i>Takahiko Ueno, Shinobu Tatsunami</i>	34
Which Items are Relevant in Customers' Shopping Cart? A Comparison of Insights from Mining Association Rules with a Network Approach <i>Ralf Wagner</i>	35
A Case Study on the Use of Statistical Classification Methods in Particle Physics <i>Claus Weihs, Olaf Mersmann, Bernd Bischl, Arno Fritsch, Heike Trautmann, Till Moritz Karbach, Bernhard Spaan</i>	36
Problems of Fuzzy c-Means and Similar Algorithms With High Dimensional Data Sets <i>Roland Winkler, Frank Klawonn, Rudolf Kruse</i>	37
On the Local Independence Assumption for Classification <i>Kazunori Yamaguchi</i>	38
Index	39

Program Schedule on July 20, 2010

09.00 - 09.15	Opening - Greetings by Prof. Baba, Prof. Gaul, Prof. Okada, Prof. Weihs
	Room 001 Session 1: Chair: Prof. Baba
09.20 - 09.45	<i>Yamaguchi</i> : On the Local Independence Assumption for Classification (Page 38)
09.45 - 10.10	<i>Mucha</i> : Pairwise Data Clustering Accompanied by Validation and Visualisation (Page 24)
10.10 - 10.35	<i>Baier, Daniel</i> : Image Clustering Algorithms for Marketing Purposes (Page 15)
10.35 - 10.50	Coffee Break
	Room 001 Session 2: Chair: Prof. Weihs
10.50 - 11.15	<i>Tsuji</i> : Three-Way Individual Scaling and Clustering for Musical Structure and Its Application (Page 32)
11.15 - 11.40	<i>Winkler, Klawonn, Kruse</i> : Problems of Fuzzy c-Means and Similar Algorithms With High Dimensional Data Sets (Page 37)
11.40 - 12.05	<i>Weihs, Mersmann, Bischl, Fritsch, Trautmann, Karbach, Spaan</i> : A Case Study on the Use of Statistical Classification Methods in Particle Physics (Page 36)
12.05 - 14.00	Lunch at Gastdozentenhaus for Workshop Participants
	Room 001 Session 3: Chair: Prof. Okada
14.00 - 14.25	<i>Geyer-Schulz, Ovelgönne</i> : The Randomized Greedy Modularity Clustering Algorithm and Its Relationship with Incomplete SAT-Solvers (Page 19)
14.25 - 14.50	<i>Ueno, Tatsunami</i> : A Non-Parametric Analysis for a Questionnaire Survey (Page 34)
14.50 - 15.15	<i>Wagner</i> : Which Items are Relevant in Customers' Shopping Cart? A Comparison of Insights from Mining Association Rules with a Network Approach (Page 35)
15.15 - 15.30	Coffee Break
	Room 001 Session 4: Chair: Prof. Krolak-Schwerdt
15.30 - 15.55	<i>Nakai</i> : Socioeconomic and Gender Differences in Voluntary Participation in Japan (Page 26)
15.55 - 16.20	<i>Imaizumi</i> : A Discrete Decomposition of Preference Matrix by ADCLUS-type model (Page 20)
16.20 - 16.45	<i>Krolak-Schwerdt, Hörstermann</i> : Comparison of Regression-based Modeling Methods for Judgment Data (Page 21)
16.45 - 17.00	Coffee Break
	Room 001 Session 5: Chair: Prof. Mizuta
17.00 - 17.25	<i>Arai, Miyano</i> : Optimal Interval-Valued Scaling of Successive Categories (Page 13)
17.25 - 17.50	<i>Terada, Yadohisa</i> : Modal Interval-Valued Dissimilarity between Histogram-Valued Data (Page 30)
17.50 - 18.15	<i>Minami</i> : Empirical Studies on the Analysis of a Vast Amount of the Internet Traffic Data in Japan (Page 22)
19.00	Get-Together at Schloßcafe with GfK1 2010 Participants

Program Schedule on July 21, 2010

	Opening of GfKI 2010
	Room 001 Session 6: Chair: Prof. Imaizumi
11.00 - 11.25	<i>Gastes, Gaul</i> : Consistency Improving of Pairwise Comparison AHP Data and Their Impact on Derived Weights (Page 18)
11.25 - 11.50	<i>Nakayama</i> : Analysis of Conditional and Marginal Association in One-Mode Three-Way Proximity Data (Page 27)
11.50 - 12.15	<i>Toyada, Imaizumi</i> : Analysis of Brand Categories Using Purchase History Data for Brand Management: the Study of Category Composition and those Trends (Page 31)
12.15 - 14.00	Lunch
14.00 - 14.40	Semi-Plenary Presentations of GfKI 2010
	Room 001 Session 7: Chair: Prof. Yamaguchi
14.45 - 15.10	<i>Okada</i> : Centrality of Two-Mode Two-Way Social Network (Page 28)
15.10 - 15.35	<i>Tuma, Decker</i> : Identifying Consumer Segments from Online Product Reviews Using Finite Mixture Models (Page 33)
15.35 - 16.00	<i>Mizuta</i> : Combination of Symbolic Data Analysis and Functional Data Analysis (Page 23)
16.00 - 16.15	Coffee Break
	Room 001 Session 8: Chair: Prof. Bock
16.15 - 16.40	<i>Baba</i> : Continuous-discrete Transformation in Multivariate Data Analysis (Page 14)
16.40 - 17.05	<i>Preisach, Schmidt-Thieme</i> : Relational Time Series Classification (Page 29)
17.05 - 17.30	<i>Bock</i> : Clustering Methods for Time Series (Page 16)
	Closing of the Workshop

**Abstracts of the 3rd German-Japanese
Workshop**

Optimal Interval-Valued Scaling of Successive Categories

Sayaka Arai and Hisao Miyano

The National Center for University Entrance Examinations,
2-19-23 Komaba, Meguro-Ku, Tokyo, Japan.
sayarai@rd.dnc.ac.jp

Abstract. A new optimal scaling method for successive categories (ordered categories) data is proposed. While the classical optimal scaling method determines one scaled value for each category, in this study, we estimate interval-valued score for each category in view of successive categories.

The relationships between the method proposed in this study and the correspondence analysis (CA) are also revealed. That is, as CA involves the singular value decomposition (SVD) of the matrix $R^{-1/2}NC^{-1/2}$, where $N = [n_{ij}]$ is contingency table, $R = \text{diag}(n_{i.})$, and $C = \text{diag}(n_{.j})$, in the method proposed here, the solution is the SVD of the matrix $R^{-1/2}\bar{N}G^{-1/2}$, where $\bar{N} = [\bar{n}_{ij}]$, and $G = [g_{ij}]$, $g_{ii} = \bar{n}_{.i}$, $g_{i(i-1)} = \bar{n}_{.(i-1)}/4$, $g_{i(i+1)} = \bar{n}_{.i}/4$, and otherwise $g_{ij} = 0$.

References

Golub, G. H., & Van Loan, C. F. (1996): *Matrix Computations*. Johns Hopkins: Baltimore.

Keywords

SUCCESSIVE CATEGORIES, SCALING, CORRESPONDENCE ANALYSIS

Continuous-discrete Transformation in Multivariate Data Analysis

Yasumasa Baba¹

The Institute of Statistical Mathematics,
baba@ism.ac.jp

Abstract. We meet sometimes the situation to use data coded from continuous observations. For example, in social research surveys income is asked in a frame of categorized classes and age is coded sometimes. The process of coding or categorizing is same to transform continuous data to discrete ones. The transformation from original continuous observation to discrete data is useful for condensing data amount if information loss caused by such process is small. In multivariate analysis correlation between variables plays a central role to get solutions. Therefore if correlation is not changed or slightly changed by the transformation, we will get a similar solution before transforming the original one.

In this paper we will discuss the influence of transformation from continuous data to discrete ones in multivariate analysis. As an example principal component analysis will be discussed. It will be shown that the effect of transformation is smaller than expectation in results of PCA. Therefore, such transformation is useful for opening data with protection of individual information. If we transform continuous raw data to categorized ones the risk that any individual is identified is reduced.

Suppose that we need a data set for education of statistics. Raw data are not available to be opened but simulated data are available. In such case we can use data that are generated through two processes. The first step is to transform original continuous data to discrete ones. The second step is to generate new continuous data from the discrete data by addition of fluctuation. If new data have similar correlation structure we can use them as if they were original without any risk of individual identification. In this paper it will be shown that such two step transformations keep correlation structure.

References

Baba, Y. (2010): Continuous-discrete Transformation and Multivariate Analysis. *Proceedings of JKCS-2010*. 185-186.

Keywords

CATEGORICAL DATA, SCRAMBLED DATA, DISCLOSURE

Image Clustering Algorithms for Marketing Purposes

Daniel Baier and Ines Daniel

Institute of Business Administration and Economics,
Brandenburg University of Technology Cottbus,
Postbox 101344, 03013 Cottbus, Germany
{daniel.baier, ines.daniel}@tu-cottbus.de

Abstract. Clustering algorithms are standard tools for marketing purposes (see, e.g., Punj, Stewart 1983). So, e.g., in market segmentation, they are applied to sociodemographic, psychographic, preference, or usage descriptions of potential customers in order to derive homogeneous groups of them. However, recently, the available resources for this purpose have extended. So, e.g., in social networks (e.g. facebook, flickr) potential customers provide images - and other information as e.g. profiles, contact lists, music or videos - which reflect their interests (e.g., w.r.t. living conditions, preferred stars, or holiday experiences).

In this paper we discuss, how uploaded images could be used for deriving market segments. However, since the similarity of images is a highly subjective matter (the focus of similarity is different across individuals and situations, see, e.g., Jain 1989), the standard algorithms for image clustering (e.g., Law et al. 2004, Figueiredo 2007) have to be adapted for this purpose. The paper discusses these standard algorithms and their modification for marketing purposes and compares results with results obtained using traditional approaches. Real and simulated social network databases are used for demonstrating the usefulness of the new approach.

References

- FIGUEIREDO, M. (2007): Semi-Supervised Clustering: Application to Image Segmentation. *Studies in Classification, Data Analysis, and Knowledge Organization*, 34, 39–50.
- JAIN, A.K. (1989): *Fundamentals of Digital Image Processing*. Prentice Hall, Upper Saddle River, NJ, USA.
- LAW, M., FIGUEIREDO, M., and JAIN, A.K. (2004): Simultaneous Feature Selection and Clustering Using Mixture Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), 1154–1166.
- PUNJ, G., STEWART, D.W. (1983): Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*, 20 (2, May), 134–148.

Keywords

MARKET SEGMENTATION, IMAGE CLUSTERING ALGORITHMS

Clustering Methods for Time Series

Hans-Hermann Bock

Institute of Statistics, RWTH Aachen University, Aachen, Germany,
bock@stochastik.rwth-aachen.de

Abstract. The paper considers the problem of clustering a given set of time series into a suitable number of clusters each comprising time series with a 'similar' structure. There are many ways of describing the similarity of time series, based either on suitable dissimilarity measures followed by classical clustering techniques, or on probabilistic models for class-specific processes followed by a maximum likelihood or mixture approach.

We provide, on the one hand, a brief survey on several of these approaches and present, on the other hand, a new dynamic clustering model for time series (in analogy to a finite element approach by Horenko 2009). This model comprises (1) a set of m class-specific prototype processes within the m classes, (2) a fuzzy clustering of the n observed time series that is (3) dynamically evolving over time (such that each time series may change its membership degrees in the course of time) with the constraint that (4) the time-dependent membership functions may not vary too fast and chaotically. Within this framework a suitable clustering criterion is proposed. The optimization of the membership functions and the prototypes is possible by approximating the membership functions by a discrete finite series of windows functions. The corresponding optimization algorithm then results in a k -means like iteration process that alternately determines optimum parameter configurations for the prototypes and solves a quadratic optimization problem under linear constraints for the unknown coefficients of the membership series.

References

- Aggarwal, Ch.C. (2006): *Data streams: models and algorithms*. Springer, New York, 2006.
- Boets, J., De Cock, K., Espinoza, M., De Moor, B. (2005): Clustering time series, subspace identification and cepstral distances. *Communications in Information and Systems* 5, 69-96.
- Douzal-Chouakria, A., Naghabushan, P.N. (2007): Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification* 1, 5-22.
- Horenko, I. (2009a): Finite element approach to clustering of multidimensional time series. *SIAM J. on Scientific Computation* (in print)
- Horenko, I. (2009b): On clustering of non-stationary meteorological time series. *Dynamics of Atmospheres and Oceans* (Electr. Journal), DOI:10.1016/j.dynatmoce.2009.04.003
- Kalpakis K., Gada, D., Puttagunda, V. (2001): Distance measures for effective clustering of ARIMA time series. Proc. IEEE International Conference on Data Mining (ICDM'01), San Jose, CA, November 29- December 2, 2001, 273-280.

- Mörchen, F. (2006): *Times series knowledge mining*. Ph.D. thesis, University of Marburg, Germany. Gorich & Weiershauser, Marburg, ISBN 3-89703-670-3.
- Oates, T., Firoiu, L., Cohen, P.R. (2001): Using dynamic time warping to bootstrap HMM-based clustering of time series. In: R. Sun, L. Giles (des.): *Sequence Learning: Paradigms, Algorithms and Applications*. Lecture Notes in Computer Science, vol. 1828. Springer-Verlag, Berlin, 35-52.
- Domenico Piccolo (1990): A distance measure for classifying ARIMA models. *J. of Time Series Analysis* 11, 153-164.
- Pattarin, F., Paterlini, S., Minerva, T. (2004): Clustering financial time series: an application to mutual fund style analysis. *Computational Statistics and Data Analysis* 47, 353-372.
- Scharl, T., Leisch, F. (2006): Jackknife distances for clustering timecourse gene expression data. In: 2006 JSM Proceedings, American Statistical Association, Alexandria, USA, S. 346 - 353.

Consistency Improving of Pairwise Comparison AHP Data and Their Impact on Derived Weights

Dominic Gastes¹ and Wolfgang Gaul²

¹ Institute of Decision Theory and Operations Research, Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, 76131 Karlsruhe, Germany.

`dominic.gastes@kit.edu`

² Institute of Decision Theory and Operations Research, Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, 76131 Karlsruhe, Germany.

`wolfgang.gaul@kit.edu`

Abstract. The Analytic Hierarchy Process (AHP) is an often applied and well researched method in the area of multi attribute decision making. One of the main recurring tasks in AHP is the creation of pairwise comparison matrices, the examination of their consistencies, and the derivation of weights.

The importance of controlling these consistencies and the consequences of subsequent adjustments of comparison matrices are often unvalued issues in AHP applications. We conduct a simulation study and show, that increasing inconsistencies caused by superimposed stochastic error-terms on consistent starting pairwise comparison AHP data result in decreasing correlations of the corresponding weights. Then we compare different approaches for automated consistency adjustments but have to conclude that correlations w.r.t. the weights derived from adjusted matrices show nearly no improvement.

These results emphasize the importance of data collection tasks, which should be organized in a way, that yields consistent matrices right from the beginning.

References

- Fedrizzi, M. and Giove, S. (2007): Incomplete pairwise comparison and consistency optimization. *European Journal of Operational Research*, 183:393–313.
- Lina, C., Wanga, W., and Yub, W. (2008): Improving AHP for construction with an adaptive AHP approach (A3). *Automation in Construction*, 17:180–187.
- Mamat, N. and Daniel, J. (2007): Statistical analyses on time complexity and rank consistency between singular value decomposition and the duality approach in AHP: A case study of faculty member selection. *Mathematical and Computer Modelling*, 46, 1099–1106.
- Saaty, T. L. (2003): Decision-making with the AHP: Why is the principal eigenvector necessary. *European Journal of Operational Research*, 145, 85–91.
- Zeshui, X. and Cuiping, W. A. (1999): Consistency improving method in the analytic hierarchy process. *European Journal of Operational Research*, 116, 443–449.

Keywords

ANALYTIC HIERARCHY PROCESS, CONSISTENCY, COMPARISON MATRICES

The Randomized Greedy Modularity Clustering Algorithm and Its Relationship with Incomplete SAT-Solvers

Andreas Geyer-Schulz¹ and Michael Ovelgönne²

¹ Information Services and Electronic Markets, IISM, KIT, Kaiserstrasse 12, D-76128 Karlsruhe andreas.geyer-schulz@kit.edu

² Information Services and Electronic Markets, IISM, KIT, Kaiserstrasse 12, D-76128 Karlsruhe michael.ovelgoenne@kit.edu

Abstract. Newman and Girvan (2004) introduced modularity clustering as an efficient way of clustering these networks by maximizing the modularity measure. Brandes et al. (2008) have given an integer linear programming formulation for modularity clustering and established that the formal problem is NP-hard. An efficient randomized greedy modularity clustering algorithm has been introduced by the authors in Ovelgönne et al. (2010). In this contribution the similarities of the randomized greedy algorithm with randomized GSAT-solvers are explored. It is conjectured that Smale's complexity results (Smale (1983)) on the average behavior of the simplex algorithm are applicable too.

References

- NEWMAN, M. E. J. and GIRVAN, M. (2004): Finding and Evaluating Community Structure in Networks. *Physical Review E*, 69(2), 026113.
- BRANDES, U. and DELLING, D. and GAERTLER, M. and GORKE, R. and HOEFER, M. and NIKOLOSKI, Z. and WAGNER, D. (2008): On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2), 172-188.
- OVELGÖNNE, M. and GEYER-SCHULZ, A. (2010): *Randomized Greedy Modularity Optimization for Group Detection in Huge Social Networks*. To be published: Workshop on Social Network Mining and Analysis.
- SMALE, S. (1983): On the average number of steps of the simplex method of linear programming. *Mathematical Programming*, 27(3), 241-262.

Keywords

MODULARITY CLUSTERING, GSAT SOLVERS, COMPLEXITY, SIMPLEX ALGORITHM

A Discrete Decomposition of Preference Matrix by ADCLUS-type model

Tadashi Imaizumi

Tama University,
 imaizumi@tama.ac.jp

Abstract. Preference data of an individual to n objects is a popular data collected in Marketing. The unfolding distance models have been used to analyze these type of data matrix. It is difficult to understand what attributes contribute on preference evaluation from using these continuous mapping models. On the other hand, the overlapping cluster models and methods such as ADCLUS (Shepard and Arabie, 1979) have interesting features to find the attributes in similarity data. Wiedenbeck and Krolak-Schwerdt (2010) investigated the properties of two-mode ADCLUS model. I will propose ADCLUS-type model for preference data. We need to include the term representing the difference from "ideal object" to real object in modeling,

$$pref_{ij} \approx \sum_{p=1}^t w_p f_{ip} g_{jp} - \sum_{p=1}^t u_t f_{ip} (1 - g_{jp}) + c_i$$

where $pref_{ij}$ is preference for object o_j of individual i , f_{ip} and g_{jp} is a binary value representing discrete feature. An optimization method and an application to real data set are also shown.

References

- SHEPARD, R.N. and ARABIE, P.(1979). Additive clustering representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86, 87-123
- Wiedenbeck, M. and Krolak-Schwerdt, S. (2010). ADCLUS:A data model for the comparison of two-mode clustering methods by Monte Carlo simulation. In Okada, A., Imaizumi, T. Bock, H.-H. and Gaul W. *Cooperation in Classification and Data Analysis, 2010, Springer, 41-51*

Keywords

OVERLAPPING CLUSTER, ADCLUS, PREFERENCE FEATURE, OPTIMIZATION

Comparison of Regression-based Modeling Methods for Judgment Data

Sabine Krolak-Schwerdt and Thomas Hörstermann

University of Luxembourg, Route de Diekirch, L-7220 Walferdange, www.uni.lu

Abstract. Research on judgment and decision making in applied settings, e.g. education or marketing, generally deals with situations in which multiple information is available and has to be integrated into a mostly unidimensional judgment. For example, teachers have to grade students based on continuous observation of performance and customers have to integrate several aspects of a product to judge its desirability. Much effort has been invested by research to investigate the structure of the decision rules that are used to transform available information into a judgment. Judgment theories consider its structure as either compensatory or noncompensatory. Compensatory structures can be approximated by (weighted) additive models. These models may be formally represented by multiple regression and, vice versa, multiple regression is applied to model these decision rules (Dawes & Corrigan, 1974, Dhami & Harries, 2001). Nevertheless, it might be disputable if the specification of the regression design influences its appropriateness to model judgment data. Two possible regression designs are compared: (a) individual regression analysis and (b) hierarchical linear modeling (HLM). The comparison is based on simulated data and investigates the superiority of one of the two designs under varying assumptions to the data.

References

- DHAMI, M. and HARRIES, C. (2001): Fast and frugal versus regression models of human judgment. *Thinking and Reasoning*, 7, 5-27.
- DAWES, R. M. and CORRIGAN, B. (1974): Linear models in decision making. *Psychological Bulletin*, 81, 95-106.

Keywords

JUDGMENT FORMATION, JUDGMENT MODELING, REGRESSION-BASED DESIGNS

Empirical Studies on the Analysis of a Vast Amount of the Internet Traffic Data in Japan

Hiroyuki Minami

Information Initiative Center, Hokkaido University, JAPAN
min@iic.hokudai.ac.jp

Abstract. In the paper, we discuss several results on the analysis for tons of the Internet traffic collected in Japan.

Next Generation IX Consortium (<http://www.distix.net/index.html>) is organized to realize and encourage a new traffic model in the next age in the Internet from the viewpoint of an Internet exchange. We took part in Resilient Project in the consortium, whose objective was to monitor and visualize its reliability and quality on end-to-end transmissions in the framework of the Internet, and to seek much robust and reliable connections. The project placed around 30 servers in many prefectures in Japan and collected ICMP RTT (Round Trip Time) data every 5 minutes over years.

We have reported a part of them (e.g. Minami & Mizuta, 2008; Minami, 2009). Now, the project is over and had stopped recording the huge data. It is time to analyze and discuss the total overview.

In this paper, we show their analytic results and try to offer some useful interpretation.

References

- Minami, H. (2009). Comparative Studies on Real Applications for Internet Control Management Protocol Data with Functional and Symbolic Data Analysis. *Proceedings of IFCS2009 / GFKL*, 115.
- Minami, H. and Mizuta, M. (2008). An Analysis of Layer 2 Network Monitoring Data with Huge-Data Oriented Statistical Techniques. *Proceedings of IASC2008*, 258.

Keywords

ICMP, PACKET TRANSMISSION, HUGE DATA

Combination of Symbolic Data Analysis and Functional Data Analysis

Masahiro Mizuta

Information Initiative Center, Hokkaido University, Sapporo 060-0811, JAPAN
 mizuta@iic.hokudai.ac.jp

Abstract. I discuss Functional Data Analysis (FDA) and Symbolic Data Analysis (SDA), and propose a method for clustering.

Most methods for data analysis assume that the data are sets of numbers with structure. For example, typical multivariate data are identified as a set of n vectors of real numbers. However, requests for analysis with new models become higher, as kinds and quantities of data are increased. In accordance with the requests, Ramsay *et al.* proposed FDA, which treats data as functions. Another great approach to deal with such complex data is SDA proposed by Diday. He said “We define Symbolic Data Analysis as the extension of standard analysis to (symbolic) tables”. Symbolic table may contain data of different types, including quantitative values, categorical values, interval values.

From the viewpoints of FDA and SDA, we propose a clustering method for functional symbolic data.

References

- Billard,L. and Diday,E.(2006): *Symbolic Data Analysis*, Wiley.
- Bock,H.H. and Diday,E. eds(2000): *Analysis of Symbolic Data*. Exploratory Methods for Extracting Statistical Information from Complex Data, Series: Studies in Classification, Data, and Knowledge Organization, 15. Springer-Verlag.
- Diday,E. and Noirhomme-Fraiture,M. eds.(2008): *Symbolic Data Analysis and the SODAS Software*, Wiley.
- Ramsay, J.O. and Silverman, B.W.(2002): *Applied Functional Data Analysis – Methods and Case Studies –*. New York: Springer-Verlag.
- Ramsay, J.O. and Silverman, B.W.(2005): *Functional Data Analysis*. 2nd Edition. New York: Springer-Verlag.

Keywords

INTERVAL VALUED DATA, DERIVATION

Pairwise Data Clustering Accompanied by Validation and Visualisation

Hans-Joachim Mucha

Weierstraß-Institut für Angewandte Analysis und Stochastik, D-10117 Berlin,
mucha@wias-berlin.de

Abstract. Pairwise proximities (distances, similarities, . . .) are often the starting point for finding clusters by applying hierarchical cluster analysis techniques and partitional clustering methods. We refer to such clustering methods as pairwise data clustering methods (Mucha, 2009). Proximities are in some sense more general compared to a data matrix. First we focus on Gaussian model-based cluster analysis of observations in its simplest setting that results in the sum of squares and logarithmic sum of squares methods. We formulate the corresponding partitional K -means-like cluster analysis as pairwise data clustering. Obviously, the usual estimation of expectation values of clusters is no longer necessary. These simple methods can become more general by weighting the observations. By doing so, for instance, clustering the rows and columns of a contingency table will be performed based on pairwise chi-square distances. Another main focus is on validation based on bootstrapping, for example, by random weighting the observations (Mucha, 2007, Mucha, 2009). The proposed built-in validation techniques can verify the results of the two most important families of methods, the hierarchical and partitional cluster analysis. The finding of the appropriate number of clusters, as the main task of model selection, is the ultimate aim here. The built-in validation evaluates additionally both the stability of each cluster based on measures of correspondence between clusters (Hennig, 2007) and the degree of membership of each observation to its cluster. The new Excel “Big Grid” spreadsheet is both the distinguished repository for data/distances/clusters/hierarchies and the perfect plotting board for multivariate graphics that can be composed in VBA-code. Examples are dendrograms, plot-dendrograms (Mucha et al., 2005), scatterplot matrices, density plots, principal components analysis plots, correspondence analysis plots, . . . In cell-based graphics (“pixel graphics” like the heat plot of a proximity matrix), “playing” with the properties/content of cells is possible. Cell-sited graphics, such as sparklines, are suitable for showing the behavior of the stability of clusters of a hierarchy. Applications are presented throughout the paper. They come from quite different fields such as archaeometry and economics.

References

- HENNIG, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, 52, 258–271.
- MUCHA, H.-J. (2007). On Validation of Hierarchical Clustering. In: R. Decker and H.-J. Lenz, editors, *Advances in Data Analysis*. Springer, Berlin, 115–122.

- MUCHA H.-J. (2009). Cluscorr98 for Excel 2007: Clustering, Multivariate Visualization, and Validation. In: H.-J. Mucha and G. Ritter, editors, *Classification and Clustering: Models, Software and Applications*, Report no. 26, WIAS, Berlin, 14–41.
- MUCHA, H.-J., BARTEL H.-G. and DOLATA, J. (2005). Techniques of Rearrangements in Binary Trees (Dendrograms) and Applications. *Match* 54 (3), 561–582.

Socioeconomic and Gender Differences in Voluntary Participation in Japan

Miki Nakai

Department of Social Sciences, College of Social Sciences, Ritsumeikan University,
56-1 Toji-in Kitamachi, Kyoto 603-8577 Japan mnakai@ss.ritsumei.ac.jp

Abstract. The aim of the present reserach paper is to examine the relationship among association membership, socioeconomic position, and gender. Participation in voluntary association has been seen as an agent by which weak individuals become strong and has been regarded as a form of social capital. This 'social capital' concept has been argued by a number of social scientists; involvement in associational activity is conducive to the prosperity of democratic institutions (Putnam, 2000). However, how different types of association are associated have yet to be elucidated fully (Li et al. 2003, Nakai 2005). To clarify the structural pattern of voluntary participation, multiple correspondence analysis was adopted so that we get a multidimensional graphical visualization of the pattern of relationship among the categorical variables. Based upon a national sample in Japan in 2005 (N=2827), we analyze composition of cultural capital and the relation between social class and participation in the Japanese context. We show socioeconomic and gender differences in participation. We discuss the role of voluntary association membership.

References

- LI, Y., SAVAGE, M. and PICKLES, A. (2003): Social Capital and Social Exclusion in England and Wales (1972-1999). *British Journal of Sociology*, 54, 497-526.
- NAKAI, M. (2005) Social Stratification and Social Participation: the Role of Voluntary Association Membership. In: F. Ojima (Ed.): *Research on Gender and Social Stratification in Contemporary Japan*.53-63.
- PUTNAM, R. (2000): *Bowling Alone: the Collapse and Revival of American Community*. New York: Simon and Schuster.

Keywords

PARTICIPATION, SOCIAL CAPITAL, VOLUNTARY ASSOCIATION

Analysis of Conditional and Marginal Association in One-Mode Three-Way Proximity Data

Atsuhō Nakayama

Faculty of Economics, Nagasaki University, 4-2-1 Katafuchi, Nagasaki, Japan
850-8506

Abstract. The purpose of present study is to examine whether the triadic distance model is necessary or not. Triadic distance model would be used to analyze conditional and marginal one-mode three-way proximity data in the present study. The results obtained from marginal association would be compared that of conditional association in one-mode three-way proximity data. Then, it examined whether the triadic distance model is necessary or not. Gower and De Rooij (2003) concluded that the results of a one-mode three-way MDS are similar to that of a one-mode two-way MDS. The reason for such similarity is that strong influences of dyadic relationships on triadic relationships would be included in one-mode three-way proximity data. If, the results of a one-mode three-way MDS are similar to that of a one-mode two-way MDS, dyadic relationships strongly influences triadic relationships. On the other hand, if the results of a one-mode three-way MDS are not similar to that of a one-mode two-way MDS, dyadic relationships weakly influences triadic relationships. Under weak influence of the dyadic relationships, the triadic and dyadic relationships would be separately analyzed. Therefore, the present study compares the results obtained from marginal association with that of conditional association. It examined whether the triadic distance model is necessary or not.

References

GOWER, J. C., and De ROOIJ, M. (2003). A comparison of the multidimensional scaling of triadic and dyadic distances. *Journal of Classification*, 20, 115-136.

Centrality of Two-Mode Two-Way Social Network

Akinori Okada

Graduate School of Management and Information Sciences Tama University, 4-1-1
Hijirigaoka Tama-shi Tokyo Japan 206-0022 okada@tama.ac.jp,
okada@rikkyo.ac.jp

Abstract. The centrality represents the strength, dominance, or importance of the actor in a social network. The two-mode two-way social network consists of relationships between two different sets of actors. Relationships of two-mode two-way social network are represented by a rectangular matrix, where the row and the column correspond to two different sets of actors respectively. Okada (2008) introduced a procedure to derive the centrality along two or more dimensions by the singular value decomposition. Okada (2010) extended the procedure to deal with the asymmetric social network. The procedure used in the present study is extended from that of Okada (2010), and it derives the centrality of the actor corresponding to the row and the centrality of the actor corresponding to the column along two or more dimensions. The centrality given to row j represents the strength of actor j , and the centrality of given to column k represents the strength of actor k in the relationships between actors corresponding to the row and those corresponding to the column. The present procedure is applied to analyze the relationships between chief executive officers and the clubs and boards (Wasserman & Faust, 1994).

References

- OKADA, A. (2008): Two-Dimensional Centrality of a Social Network. In: C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker (Eds.): *Data Analysis, Machine Learning and Applications*. Springer, Heidelberg-Berlin, 381–388.
- OKADA, A. (2010): Two-Dimensional Centrality of Asymmetric Social Network. In: F. Palumbo, C.N. Lauro and M.J. Greenacre (Eds.): *Data Analysis and Classification*. Springer, Heidelberg, 93–100.
- WASSERMAN, S. and FAUST, K. (1994): *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

Keywords

CENTRALITY, MDS, SINGULAR VALUE DECOMPOSITION, SOCIAL NETWORK, TWO-MODE TWO-WAY PROXIMITIES

Relational Time Series Classification

Christine Preisach¹ and Lars Schmidt-Thieme¹

Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim,
31141 Hildesheim, Germany

preisach@ismll.uni-hildesheim.de,
schmidt-thieme@ismll.uni-hildesheim.de

Abstract. In many application domains such as health and engineering a huge amount of time dependent data is recorded and used. For easy browsing and searching it usually is categorized in a set of classes. Formally, this problem can be described as time series classification. While state-of-the-art methods in time series classification use elaborated distance measures such as dynamic time warping (DTW), the applied classification models are often very simple 1-Nearest Neighbor models.

In this paper we show how the time series classification problem can be cast into a relational learning problem, where a neighborhood graph built from similarities is used for collective inference. We propose a new method called relational time series classification (*RTSC*) and show empirically on a set of 20 benchmark datasets that our method outperforms existing state-of-the-art algorithms. Furthermore, we apply relational ensemble classification to multivariate time series classification problems and show that it outperforms more complex state-of-the-art methods on two real-life datasets.

References

- KEOGH, E. and KASSETY, S. (2002): On the need for time series data mining benchmarks: a survey and empirical demonstration. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY.
- PREISACH, C. and SCHMIDT-THIEME, L. (2006): Relational Ensemble Classification. In: *ICDM '06 Proceedings of the Sixth International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 499–509.

Keywords

TIME SERIES CLASSIFICATION, MULTIVARIATE TIME SERIES, RELATIONAL CLASSIFICATION

Modal Interval-Valued Dissimilarity between Histogram-Valued Data

Yoshikazu Terada¹ and Hiroshi Yadohisa²

¹ Graduate School of Culture and Information Science, Doshisha University, Kyoto 610-0394, JAPAN

² Department of Culture and Information Science, Doshisha University, Kyoto 610-0394, JAPAN

Abstract. Representation of interval-valued data and histogram-valued data come in very useful, when we represent a variation of data or aggregate large amounts of data. Histogram-valued data is expressive of more information regarding a distribution of the data. A variety of methods for analyzing the histogram-valued data have been proposed (See, e.g. Box and Diday, 2000; Billard and Diday, 2006). However, the distribution of the data was not considered in those methods. For example, most dissimilarities between two histograms don't represent the distribution of the data, i.e., dissimilarities have been represented as a single value. If we focus on the distribution of the original data, we must analyze for histogram-valued data with that. Then, we consider dissimilarities, which represent the distribution of the data, between two histograms, that is, these dissimilarities are represented as modal interval values.

In this paper, we define a new dissimilarity measure, which is represented as a modal interval value, between histogram valued data. A new multidimensional scaling method by using the dissimilarity is also proposed.

Analysis of Brand Categories Using Purchase History Data for Brand Management: the Study of Category Composition and those Trends

Toyada, Yuki¹ and Imaizumi, Tadashi²

¹ Tama University toyoda@tama.ac.jp

² Tama University imaizumi@tama.ac.jp

Abstract. The aim of this study is to propose a method required for brand management to understand the composition of categories and analyze trends in it as easily as possible. In particular, it seeks to respond to the practical demand for useful insights into brand management based on purchase history data accumulated from day to day.

The important point in this is to be able to do so as easily as possible. The interpretation of brand categories complicated by diversified consumer selection behavior requires on-the-ground knowledge. Hence, a method of analysis is needed for marketers who are not necessarily specialized in data analysis to be able to interpret the data by themselves through trial and error.

This study also seeks to establish a method to analyze what sort of customers favor the sampled categories. Thus, it seeks not only to inquire into the composition of categories but also to determine trends regarding them.

In order to achieve that aim, this study establishes an analytical method, or procedure and applies it to foodstuffs (including beverages) for which consumer selection behavior is diverse. Its effectiveness and problems are thus examined.

Keywords

SEGMENTATION, PURCHASE HISTORY DATA, BRAND CATEGORY, BRAND MANAGEMENT

Three-Way Individual Scaling and Clustering for Musical Structure and Its Application

Mitsuhiro Tsuji¹

Kansai University, tsuji@kansai-u.ac.jp

Abstract. The field of statistics in music research is very broad. I investigate three-way individual scaling and clustering to musical structural analysis of emotion expression. The affective values of music were measured from subjects who were interested in theater and in instrumental music as well. I expected to find interesting structure in the affective expression of music from three view points : affective values, individuals, and music.

I analyzed the three-way structure model by applying INDSCAL and INDCLUS. I expected that the INDSCAL model would show a geometrical structure which offers interesting insights about the characteristics of affective values. Furthermore the INDCLUS model would reveal further real geometrical structure.

References

- ARABIE, P., CAROLL, J. D., and DESARBO, W. S. (1987): *Three-way Scaling and Clustering*. Sage, Newbury Park.
- GABRIELSSON, A. and JUSLIN, P.N. (2009): Emotional expression in music. In: R. J. Davidson, H. H. Goldsmith, and K. R. Scherer (Eds): *Handbook of affective sciences*, New York: Oxford University Press.
- OKADA, A. (2001): A Review of Cluster Analysis Research in Japan. In: P. Arabie, L. J. Hubert and G. De Sorte (Eds): *Cluster and Classification*. River Edge, NJ: World Scientific.
- WEIHS, C., LIGGES, U., MORCHEN, F., and MULLENSIEFEN, D. (2007): Classification in music research. *Advances in Data Analysis and Classification*, Vol. 1, No. 3, 255-291.

Keywords

SCALING, CLUSTERING, GRAPHICS, MUSIC, EMOTION

Identifying Consumer Segments from Online Product Reviews Using Finite Mixture Models

Michael Tuma¹ and Reinhold Decker²

¹ Department of Business Administration and Economics, Bielefeld University, D-33615 Bielefeld, Germany mctuma@googlemail.com

² Department of Business Administration and Economics, Bielefeld University, D-33615 Bielefeld, Germany rdecker@wiwi.uni-bielefeld.de

Abstract. Online product reviewing is an emerging phenomenon that plays an increasingly important role in purchase decisions (Chen and Xie, 2008). Recent empirical surveys show that more and more people rely on opinions posted on blogs, online forums and opinion portals when making a variety of decisions ranging from what movies to watch to what products to purchase.

Despite this importance of opinion analysis, to the best of our knowledge, there has been no attempt by marketing researchers to identify different types of consumers who post their opinions online. This study seeks to fill this gap. Using appropriate text mining techniques (Feinerer et al. 2008), we develop and empirically evaluate a novel approach - related to that of Decker and Gribba (2009) - to identify those variables which, at least partly, explain the articulated opinions. These variables are then used in a model-based segmentation approach (Wedel and Kamakura, 2000) to identify homogeneous segments of consumers that can be targeted with the same marketing measures.

References

- CHEN, Y. and XIE, J. (2008): Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix. *Management Science*, 54(3), 477-491.
- DECKER, R. and GNIBBA, K. (2009): Konsumentenforschung im Web 2.0 - Analyse von Online-Rezensionen zur kundenorientierten Produktgestaltung. *Marketing ZFP*, 31(2), 117-136.
- FEINERER, I., HORNIK, K. and MEYER, D. (2008): Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5), 1-54.
- WEDEL, M. and KAMAKURA, W. A. (2000): *Market Segmentation: Conceptual and Methodological Foundations*. Kluwer Academic Publishers, Dordrecht

Keywords

MARKET SEGMENTATION, ONLINE REVIEWS, FINITE MIXTURE MODELS, SEGMENTATION VARIABLES, MODEL-BASED CLUSTERING

A Non-Parametric Analysis for a Questionnaire Survey

Takahiko Ueno¹ and Shinobu Tatsunami²

¹ Medical Statistics, St. Marianna University School of Medicine, Kawasaki, Japan
216-8511 t2ueno@marianna-u.ac.jp

² Medical Statistics, St. Marianna University School of Medicine, Kawasaki, Japan
216-8511 s2tatsu@marianna-u.ac.jp

Abstract. The questionnaire survey is used frequently that aims at investigation of quality of life (QOL) as well as other social problems. Then multivariate analyses such as the multiple regression analysis, the principal component analysis and the factor analysis, will be used for the interpretation of the responses. We consider the survey form composed of items each of that requires a selection from five prepared answers such as from 1 to 5, that is the answers of queries graded in 5 levels. In this case, if a responder gives one response for all queries such as $(1, 1, \dots, 1)$ or $(5, 5, \dots, 5)$, it will be less meaningful as information. However, the correlation coefficient between two items would become large if such responses were included frequently. This will make an inappropriate influence to an interpretation of results from a factor analysis. In this context, we will propose an analytical method that is not based on the correlation matrix.

We consider data from n responders on the questionnaire instrument consisted of m items. Let X_i be the column of responses from n responders to the i -th item, X_i is express as a point in an m -dimensional Euclidean space. Instead of the correlation matrix, we proposed the methods based on the Euclidean distance between X_i and X_j . Our method will be free of problematic influence arising from less meaningful responses.

References

- AGRESTI, A. (2002): *An Introduction to Categorical Data Analysis, 2nd Edition*. Wiley, England.
- FAYERS, P. M. and MACHIN, D. (2000): *Quality of Life, Assessment, Analysis and Interpretation*. Wiley, England.

Keywords

CATEGORICAL DATA, MULTIVARIATE ANALYSIS, NON-PARA-METRIC, QUESTIONNAIRE SURVEY

Which Items are Relevant in Customers' Shopping Cart? A Comparison of Insights from Mining Association Rules with a Network Approach

Ralf Wagner¹

SVI Endowed Chair for International Direct Marketing
DMCC - Dialog Marketing Competence Center, University of Kassel, Germany
rwagner@wirtschaft.uni-kassel.de

Abstract. Analyzing the choice decisions of identified customers is dominated by the application of association rules (Decker & Wagner (2002); Hahsler, Hornik, and Reutterer (2006)). This methodology is implemented in most common statistical software packages und adapted by many brick and mortar as well as online retailers. However, there are recent attempts to gain additional insights from market basket data (Decker (2005)) or more precise information (Boztuğ & Reuterer (2008)). In this study the interrelations of products bought by one and the same customer are investigated by means of a social network analysis approach (Batagelj & Mrvar (2002); Handcock, Hunter, Butts, Goodreau, and Morris (2008)).

References

- BATAGELJ, V. and MRVAR, A. (2002): Pajek- Analysis and Visualization of Large Networks. In: P. Mutzel, M. Jünger, and S. Leipert (Eds.): *Graph Drawing*. Springer, Berlin, LNCS Vol. 2265, 8–11.
- BOZTUĞ, Y. and REUTTERER, T. (2008): A combined approach for segment-specific market basket analysis. *European Journal of Operational Research*, Volume 187, Issue 1, 16 May 2008, 294–312.
- DECKER, R. (2005): Market Basket Analysis by Means of a Growing Neural Network. *The International Review of Retail, Distribution and Consumer Research*, Volume 15, No. 2, 151–169.
- DECKER, R. and WAGNER, R. (2002): *Marketingforschung*. Moderne Industrie, München.
- HANDCOCK, M. S., HUNTER, D. R., BUTTS, C.T., GOODREAU, S. M., and MORRIS, M. (2008): statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software* 24(1): 1548–7660.
- HAHSLER, M., HORNIK, K., REUTTERER, T. (2006): Implications of Probabilistic Data Modeling for Mining Association Rules. In: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (Eds.): *Studies in Classification, Data Analysis, and Knowledge Organization. From Data and Information Analysis to Knowledge Engineering*, Springer, Berlin-New York, 598–605.

Keywords

ASSOCIATION RULES, MARKET BASKET, SOCIAL NETWORK ANALYSIS

A Case Study on the Use of Statistical Classification Methods in Particle Physics

Claus Weihs¹, Olaf Mersmann¹, Bernd Bischl¹, Arno Fritsch¹, Heike Trautmann¹,
Till Moritz Karbach², and Bernhard Spaan²

¹ Statistics Department, TU Dortmund University, Germany,
{fweihs,olafmg}@statistik.tu-dortmund.de

² Physics Department, TU Dortmund University, Germany

Abstract. Current research in experimental particle physics is largely driven by the analysis of massive datasets collected by detectors at particle accelerators worldwide. One of the major tasks in this analysis is the selection of interesting or relevant events. This process is divided into several stages. Initially, a trigger system has to decide if the event should be recorded for reconstruction. This is done using time efficient rules. Next, during the reconstruction phase the data is divided into several streams. Each stream contains one category of events as well as different types of background noise. Finally the events of interest are extracted from one of these streams. In this paper we propose to use statistical classification algorithms for this task. To illustrate our method we apply it to an MC dataset from the BABAR experiment. One of the major obstacles in constructing a classifier for this task is the imbalanced nature of the dataset. Only about 0.5% of the data are interesting events. The rest are background or noise events. We show how ROC curves can be used to find a suitable cutoff value to select a reasonable subset of a stream for further analysis. Finally, we estimate the CP asymmetry of the $B \rightarrow DK$ decay using the samples extracted by the classifiers.

Problems of Fuzzy c-Means and Similar Algorithms With High Dimensional Data Sets

Roland Winkler¹, Frank Klawonn², and Rudolf Kruse²

¹ German Aerospace Center roland.winkler@dlr.de

² Ostfalia University of Applied Sciences f.klawonn@ostfalia.de

³ Otto von Guericke Universitaet Magdeburg kruse@iws.cs.uni-magdeburg.de

Abstract. Fuzzy c-Means and its derivatives work very well on most clustering problems. However, FcM and many similar algorithms have their problems with high dimensional data sets and a large number of prototypes. Similar algorithms in this context are those, which generate fuzzy membership values by using a ratio of distances to ensure a sum of membership values of 1. Possibilistic clustering is explicitly of no concern because the degrees of possibility are computed for each cluster individually. In this paper, we exploit some structural problems using the ratio of distances as normalisation method in high dimensional spaces. We also show that a high number of prototypes influences the clustering procedure in a similar way as a high number of dimensions. Both effects are not entirely independent since the number of dimensions can be effectively reduced if the number of prototypes is smaller than the number of dimensions.

Keywords

FCM, HIGH DIMENSIONALITY, DIMENSION PROBLEMS

On the Local Independence Assumption for Classification

Kazunori Yamaguchi

Rikkyo Universtiy, Tokyo 171-8501, Japan kyamagu@rikkyo.ac.jp

Abstract. The latent class analysis (LCA) was introduced by Lazarsfeld and Henry (1968) for dichotomous survey data. This methodology, in which estimation of unobserved grouping variables provides usable segments for demonstrating individual differences, has since been expanded for practical use in marketing and other behavioral fields, where both quantitative methods and substantive research are very important. The LCA method has been proven as an effective tool for clustering categorical typed data into several segments.

The local independence assumption is common for the LCA. When we use a tool for classifications, we usually have to define distances among objects.

In this paper, we give graphical interpretations of local independence assumptions and discuss about means of them using the information geometry (Amari 1985) of independent models of contingency table.

References

- AMARI, S.(1985) : *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics, Springer-Verlag, Berlin.
- LAZARSELD, P. F. and HENRY, N. W.(1968) : *Latent Structure Analysis*, Houghton Mufflin, Boston.

Keywords

LATENT CLASS ANALYSIS, LOCAL INDEPENDENCE, DISTANCE

Index

- Arai, 13
- Baba, 14
Baier, 15
Bischl, 36
Bock, 16
- Daniel, 15
Decker, 33
- Fritsch, 36
- Gastes, 18
Gaul, 18
Geyer-Schulz, 19
- Hörstermann, 21
- Imaizumi, 20, 31
- Karbach, 36
Klawonn, 37
Krolak-Schwerdt, 21
Kruse, 37
- Mersmann, 36
Minami, 22
Miyano, 13
Mizuta, 23
- Mucha, 24
- Nakai, 26
Nakayama, 27
- Okada, 28
Ovelgönne, 19
- Preisach, 29
- Schmidt-Thieme, 29
Spaan, 36
- Tatsunami, 34
Terada, 30
Toyada, 31
Trautmann, 36
Tsuji, 32
Tuma, 33
- Ueno, 34
- Wagner, 35
Weihs, 36
Winkler, 37
- Yadohisa, 30
Yamaguchi, 38

Picture Credits

The image rights, especially Publishing and Distributing, remain with the copyright owners.

- Logo of JCS, Cover Page: Japanese Classification Society, wwwsoc.nii.ac.jp/jcs/en/index_e.html
- Logo of GfKl, Cover Page: German Classification Society, www.gfkl.de
- Picture of the fan of Karlsruhe, Cover Page: Stadtarchiv Karlsruhe 8/PBS XVI 15. The image rights, especially Publishing and Distributing, remain with Stadtarchiv Karlsruhe.
- Logo of KIT, Cover Page: Karlsruhe Institute of Technology, www.kit.edu
- Logo of ETU, Cover Page: Institute of Decision Theory and Management Science, <http://marketing.wiwi.uni-karlsruhe.de/institut/index.jsp>
- Map of Karlsruhe, Page 2: www.openstreetmap.org
- Map of the Campus of KIT, Page 3: Karlsruhe Institute of Technology, www.uni-karlsruhe.de/img/KIT_campus_sued_200910.pdf
- Workshop Location, Page 4: KIT Infrastructure and Services, www.tid.kit.edu

