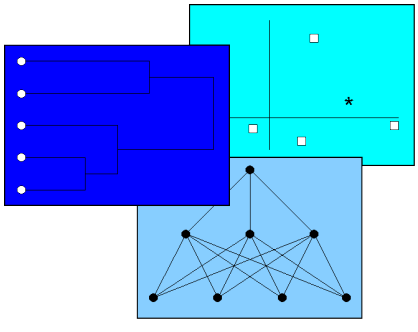
4th Japanese-German Symposium on Classification, Kyoto, 2012 (JGSC2012)

Abstracts Book



JCS

date : march 9-10, 2012

venue : Kambai-Kan, Muromachi Campus,
Doshisha University, Kyoto.

Preface

It is a real pleasure and honor to host the fourth Japanese-German Symposium on Classification from March 9 to 10 at Kyoto. The first Japanese-German Symposium on Classification was held in 2005 at Tokyo. The second German-Japanese Symposium on Classification was held at 2006 in Berlin. The third German-Japanese Symposium on Classification was held at 2010 in Karlsruhe. While the present symposium was originally planned to be held in August 2011, we had to postpone the symposium because of the earthquake in the eastern part of Japan in March 2011 and of the accident at the nuclear power plant caused by the earthquake. It is a great pleasure to welcome all of you to this symposium.

Kyoto had been the capital of Japan for about 1,100 years from the end of eighth century through the end of nineteenth century, and is a large city. All streets of Kyoto city are along either west-east direction or along south-north direction, and furthermore each street has its own name that is exceptional among Japanese cities. This makes it easy to go around in Kyoto, if you know where you are. The conference site is very close to the Imperial Residence of Kyoto, and is located north-west direction from the Imperial Residence. We hope that you will have a pleasant and comfortable stay in Kyoto, and that the symposium is interesting and stimulating scientifically for you.

March 2012 in Kyoto, JAPAN

Akinori Okada
Hiroshi Yadohisa

On behalf of the Organizing Committee and the Local Organizing Committee

Organization Committee

Japanese Side

- Prof. Dr. Yasumasa Baba (The Institute of Statistical Mathematics)
- Prof. Dr. Tadashi Imaizumi (Tama University)
- Prof. Dr. Akinori Okada (Tama University)
- Prof. Dr. Hiroshi Yadohisa (Doshisha University)

Germany Side

- Prof. Dr. Daniel Baier (Brandenburgische Technische Universität Cottbus)
- Prof. Dr. Wolfgang Gaul (Karlsruhe Institut für Technologie)
- Prof. Dr. Andreas Geyer-Schulz (Karlsruher Institut für Technologie)
- Prof. Dr. Claus Weihs (Universität Dortmund)

Friday March 9

			Page
7:50–18:00	Registration		
8:30–8:45	Opening		
8:45–10:25	Session1 (The chairperson: Hans-Hermann Bock)		
	8:45–9:10	**Comparison of two Distribution Valued Dissimilarities and its Application for Symbolic Clustering** Yusuke Matsui | Yuriko Komiya | Hiroyuki Minami | Masahiro Mizuta (Hokkaido University)	1
	9:10–9:35	**Graph Symmetries and Tests for Formal Cluster Stability in Modular Clustering** Geyer-Schulz Andreas | Michael Ovelgönne | Stein Martin (KIT-Campus Süd)	2
	9:35–10:00	**Fold Change Classifier for the Analysis of Gene Expression Profiles** Hans A. Kestler (Ulm University)	3
	10:00–10:25	**Model-Based Clustering Methods for Time Series** Hans-Hermann Bock (RWTH Aachen University)	4
10:25–10:40	Tea Break		
10:40 – 12:20	Session2 (The chairperson: Akinori Okada)		
	10:40–11:05	**On Finding Unique Clusters of Individuals in 'Pick r/n' Data Matrix** Tadashi Imaizum (Tama University), Mitsuaki Huzii | Toshinari Kamakura (Chuo University)	5
	11:05–11:30	**Agglomerative Clustering Using Assymetric Measures without Reversals in Dendrograms** Satoshi Takumi | Sadaaki Miyamoto (University of Tsukuba)	6
	11:30–11:55	**Examination of the Necessity of Using One-mode Three-way Asymmetric Model** Atsuho Nakayama (Tokyo Metropolitan University), Hiroyuki Tsurumi (Yokohama National University), Akinori Okada (Tama University)	7
	11:55–12:20	**Least Squares Permutation and Its Applications to Factor Rotation and Fixed Size Clustering** Kohei Adachi (Osaka University)	8
12:20 – 13:30	Lunch Break		

			Page
13:30 – 15:10	Session3 (The chairperson: Keiji Yajima)		
	13:30–13:55	**The Effects of Q-matrix Specification and Misspecification in Multiple-choice DINA Models** Koken Ozaki (The Institution of Statistical Mathematics), Ikko Kawahashi (Japan Foundation), Tomoko Takahashi | Yuan Sun | Sumio Kakinuma (National Institute of Informatics)	9
	13:55–14:20	**Automatic Regularization of Factorization Models** Steffen Rendle (University of Konstanz)	10
	14:20–14:45	**Thresholding Loadings in Factor Analysis** Yusuke Miyamoto (Osaka University)	11
	14:45–15:10	**Assessment of the Relationship between Native Thoracic Aortic Curvature and Endoleak Formation after TEVAR Based on Linear Discriminant Analysis** Kuniyoshi Hayashi | Fumio Ishioka (Okayama University, CREST), Bhargav Raman | Daniel Y. Sze (Stanford University School of Medicine), Hiroshi Suito (Okayama University, CREST), Takuya Ueda (St. Luke's International Hospital, CREST), Koji Kurihara (Okayama University, CREST)	12
15:10 – 15:25	Tea Break		
15:25 – 17:30	Session4 (The chairperson: Shizuhiko Nishisato)		
	15:25–15:50	**Three-way Data Analysis for Multivariate Spatial Time Series** Mitsuhiro Tsuji (Kansai University), Toshio Shimokawa (University of Yamanashi)	13
	15:50–16:15	**Three-Mode Subspace Clustering for Considering Effects under Noise Variables and Occasions** Kensuke Tanioka | Hiroshi Yadohisa (Doshisha University)	14
	16:15–16:40	**Classification, Clustering and Visualisation Based on Dual Scaling** Hans-Joachim Mucha (Weierstrass Institute)	15
	16:40–17:05	**Structural Representation of Categorical Data and Cluster Analysis through Filters** Shizuhiko Nishisato (University of Toronto)	16
	17:05–17:30	**Variable Selection in K-means Clustering via Regularization Method** Yutaka Nishida (Osaka University)	17
19:00 –21:00	Conference Dinner		

Saturday March 10

			Page
8:20 – 15:00	Registration		
9:00–10:40	Session5 (The chairperson: Tadashi Imaizumi)		
	9:00–9:25	**On the Stress Function of Asymmetric von Mises Scaling** Kojiro Shojima (The National Center for University Entrance Examinations)	18
	9:25–9:50	**Bayesian Analysis of Asymmetry by the Slide-Vector Model** Kensuke Okada (Senshu University)	19
	9:50–10:15	**A New Multidimensional Scaling Methodology for Analysis of Asymmetric Citation Data in Scientific Publications** Yuan Sun (National Institute of Informatics), Tadashi Imaizumi (Tama University)	20
	10:15–10:40	**Asymmetric Multidimensional Scaling with Generalized Hyperellipse Model** Yoshikazu Terada (Osaka University), Hiroshi Yadohisa (Doshisha University)	21
10:40–10:55	Tea Break		
10:55 – 12:35	Session6 (The chairperson: Claus Weihs)		
	10:55–11:20	**Statistical Process Modelling for Machining of Inhomogeneous Mineral Subsoil** Claus Weihs (Technische Universität Dortmund)	22
	11:20–11:45	**Zone Detection Process for Rainfall Inflow into Sewage Pipe** Ken Wada	Tomoyuki Hasegawa (STONEGATE Co.), Keiji Yajima (Independent)	23
	11:45–12:10	**The Utility of Smallest Space Analysisfor the Cross-National Survey Data Analysis: The Structure of Religiosity** Kazufumi Manabe (Aoyama Gakuin University)	24
	12:10–12:35	**Analysis of Changes of Brand Categories Using Purchase History Data and Eigenvalue to Find New Category** Yuki Toyoda (Tama University)	25
12:35 – 13:45	Lunch Break		

			Page
13:45–15:25	Session7 (The chairperson: Yasumasa Baba)		
	13:45–14:10	**An Automatic Extraction of Academia-Industry Collaborative Research and Development Documents on the Web** Kei Kurakawa | Yuan Sun (National Institute of Informatics), Nagayoshi Yamashita (Japan Society for the Promotion of Science), Yasumasa Baba (The Institute of Statistical Mathematics)	26
	14:10–14:35	**Ancient Population Dynamics Estimation from Archaeological data 'Nuzi personal names'** Sumie Ueda (The Institution of Statistical Mathematics), Kumi Makino (Kamakura Women's University), Yoshiaki Itoh (The Institution of Statistical Mathematics), Takashi Tsuchiya (National Graduate Insitute for Policy Studies)	27
	14:35–15:00	**Classification of Literature by Analyzing Figure-Ground Relationship of Characters** Tetsuya Matsui | Yukio-pegio Gunji | Eugenio-Schneider Kitamura (Kobe University)	28
	15:00–15:25	**Non-Additive Utility Functions: Choquet Integral versus logic-based Querying** Ingo Schmitt (Brandenburgische Technische Universitat Cottbus)	29
15:25–15:40	Tea Break		
15:40 – 17:20	Session8 (The chairperson: Wolfgang Gaul)		
	15:40–16:05	**Feature Selection and Clustering of Digital Images Versus Questionnaire Based Grouping of Consumers: A Comparison** Ines Daniel | Daniel Baier (Brandenburg University of Technology Cottbus)	30
	16:05–16:30	**Analysis of Asymmetric Relationships Among Soft Drink Brands** Akinori Okada (Tama University)	31
	16:30–16:55	**How to use Willingness-to-Pay Data for Product Bundling** Wolfgang Gaul (KIT-Campus Süd)	32
	16:55–17:20	**From Online Customer Reviews to New Marketing Insights Methodological Issues and Challenges** Reinhold Decker (Bielefeld University)	33
17:20 – 17:30	Closing		

Comparison of two Distribution Valued Dissimilarities and its Application for Symbolic Clustering

Yusuke Matsui[1], Yuriko Komiya[2], Hiroyuki Minami[2] and Masahiro Mizuta[2]

[1] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
[2] Information Initiative Center, Hokkaido University, Sapporo, Japan

Abstract. Symbolic Data Analysis is a new approach for data analysis. We are interested in dissimilarities represented by distributions. Let symbolic objects be $\{ O_1, O_2, \ldots, O_n \}$. We define distribution valued dissimilarities between the objects O_i and O_j as

$$d\left(O_i, O_j\right) \equiv \ S_{ij},$$

where $\{ S_{ij} \}$ are random variables.

We give a comparison rule for ordering $\{ S_{ij} \}$ as follows.

$$S_{ij} > S_{kl} \quad \Longleftrightarrow \quad Pr(S_{ij} > S_{kl}) > \frac{1}{2},$$

we put

$$Pr(S_{ij} > S_{kl}) = P_{ij,kl}.$$

Note that $P_{ij,kl} = 1 - P_{kl,ij}$. Here, we apply the rule to Symbolic Clustering as follows.

1. Initialize $cls = \{ C_1, C_2, \ldots, C_n \}$, $C_i = \{ O_i \}$, $K := n$.
2. Select the pair of clusters (i, j) to be merged such that $(i, j) = \arg\min_{i,j} P_{ij,..}$, where $P_{ij,..} = \sum_{k<l}^{n} P_{ij,kl}$.
3. Merge the pair such that $C_i \cup C_j$.
4. Update $P_{ij,kl}$ such that $P_{ij,kl} \leftarrow \ P_{\alpha k,\beta l}$, where $\alpha = \arg\min_{\alpha} \left(P_{\alpha k,ik}, P_{\alpha k,jk}\right)$ and $\beta = \arg\min_{\beta} \left(P_{\beta l,il}, P_{\beta l,jl}\right)$.
5. Update the indices of cls such that $cls = \{ C_1, C_2, \cdots, C_{K-1} \}$.
6. $K \leftarrow \ K - 1$.
7. Repeat from 2 to 6 until $K = 1$.

We will show the results of simulations and of the case of real data.

Keywords

SYMBOLIC DATA ANALYSIS (SDA), SYMBOLIC DISSIMILARITY, DISTRIBUTION VALUED DATA, HIERARCHICAL CLUSTERING

Graph Symmetries and Tests for Formal Cluster Stability in Modular Clustering

Geyer-Schulz, Andreas[1], Ovelgonne, Michael[1] and Stein, Martin[1]

KIT-Campus Süd andreas.geyer-schulz@kit.edu

Abstract. The analysis of resolution limits in community detection by Fortunato and Barthelemy (2007) shows that modularity as a formal cluster criterion does not solve the problem of determining the number of clusters and detecting a proper cluster structure simultaneously. The counter-examples given are symmetric graphs (cyclic graphs and complete cliques). In addition, we show how we can use these counter-examples to construct test cases for almost all well-known measures for comparing clusterings (e.g. the RAND index or mutual information measures) fail with arbitrary large errors.

In our contribution we investigate the construction of measures of graph symmetries as tests of formal cluster stability in modular clustering. Our approach applies results from graph enumeration and Polya's combinatorial approach as well as from automorphic forms and representations.

References

Fortunaty, Santo and Barthelemy, Marc (2007): Resolution Limit in Community Detection. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 104(1), 36-41.

Fold Change Classifiers for the Analysis of Gene Expression Profiles

Hans A. Kestler

Reasearch Group Bioinformatics and Systems Biology Institute of Neural Information Processing, Ulm University, Germany. hans.kestler@uni-ulm.de

Abstract. The classification of gene expression data is based on high-dimensional profiles containing expression levels of thousands of RNA molecules. Most state of the art algorithms in this field (e.g. RandomForests, Boosting ensembles, ...) can be seen as combining strategies of single threshold classifiers (rays). The structure of these classifiers is beneficial in this high-dimensional settings as it performs an implicit feature reduction and allows an easy semantic and syntactic interpretation.

A single ray compares a single expression value of the profile towards a global threshold t. Although this base classifier is adequate in many applications, it is questionable if the single threshold classifier is the best choice for expression data. Its dependency on a fixed threshold makes a ray susceptible to global multiplicative or additive effects.

In this work an alternative base classifier, the fold change classifier, is discussed. In a fold change classifier the global threshold t is replaced by a individual measurement taken from the query sample; the classifier compares two expression values of the same sample. This relative decision criteria makes the fold change classifier more independent against global assumptions. It is invariant against global scaling or transition.

We analyze the influence of fold change classifiers on unweighted ensembles of type majority vote or unaminity vote (logical conjunction). A sample compression bound for unweighted ensemble of fold change classifiers is shown.

Model-based clustering methods for time series

Hans-Hermann Bock

This paper considers the problem of clustering n observed time series $\mathbf{x}_k = \{ x_k(t) \mid t \in \mathscr{T}\}$, $k = 1,...,n$, with a observation interval $\mathscr{T} = [0,T]$ or a set of discrete time points $\mathscr{T} = \{t_1,...,t_p\}$ and $x_k(t) \in R^d$, into a suitable number m of clusters $C_1,...,C_m \subset \{1,...,n\}$ each one comprizing time series with a 'similar' structure. A first approach proceeds by computing dissimilarities between the observed time series and then applying classical, possibly hierarchical clustering methods. Here we will present some alternatives and provide a survey on probabilistic clustering approaches and clustering algorithms that are based on class-specific probability models for the time series within each class. In particular we will consider:

– Mixture models with class-specific Gaussian processes
– Mixture models with class-specific Markov chains
– A fixed-partition approach with normal and t-distributions.

Finally, we will briefly consider clustering methods that are related to martingale theory and financial time series.

On Finding Unique Clusters of Individuals in 'Pick r/n' Data Matrix

Tadashi Imaizumi[1], Mituaki Huzii[2], and Toshinari Kamakura[2]

[1] Tama University, 4-1-1 Hijirigaoka,Tama-shi, , Tokyo, 206-0022,Japan
imaizumi@tama.ac.jp
[2] Chuo University,1-13-27 Kasuga,Bunkyo-ku,Tokyo 112-8551, Japan
huzii@indsys.chuo-u.ac.jp,kamakura@indsys.chuo-u.ac.jp

Abstract. When we want to gather some type of sensory data, for example, movie reviews,wine tasting, etc, from expert reviewers, we need to analyze this type of data using with the individual difference model.

The expert reviewers will have different criteria on comparison. So we need to pay attention on how to represent them as points in lower dimensional space. And other hand, we will adopt the 'pick r/n' data collection method when the number of goods is too large, for example, over 30s. This data collection method is a versatile method, but, has some difficulty on analyzing data.

To analyze data collected on these two conditions, we propose a new method to analyze a sensory data gathered 'pick r/n' data collection, specially, Japanese movies data(Aoki,2007;Kamakura, Shoji, and Watanabe,2007).

In this method, we try to assign scores to represent each reviewer's data with many missing data, and classify them as a member of mini-clusters.

References

AOKI, S(2007) : *Kinema Junpo Besuto Ten 80 Kai Zenshi 1924-2006*(in Japanese). Kinena Junpo Sha, Tokyo.

KAMAKURA, T, SHOJI, H, and WATANABE, N.(2007):Eizou Kontentu no Kati Hyoka Shisutemu no Koutiku(in Japanese). KAKEN FY2007 Annual Research Report, Tokyo .

Keywords

MINI-CLUSTERS, 'PICK r/n DATA', SENSORY DATA,UNIQUENESS

Agglomerative Clustering Using Assymetric Measures without Reversals in Dendrograms

Satoshi Takumi[1] and Sadaaki Miyamoto[2]

[1] Master's Program in Risk Engineering, University of Tsukuba, Ibaraki 305-8573, Japan takumi@soft.risk.tsukuba.ac.jp
[2] Department of Risk Engineering, University of Tsukuba, Ibaraki 305-8573, Japan miyamoto@risk.tsukuba.ac.jp

Abstract. It is well-known that different linkage methods in agglomerative hierarchical clustering (abbreviated AHC) are obtained from different updating schemes between a pair of clusters. Generally, they are defined by a bottom-up scheme: first we define dissimilarity $d(x, y)$ between two individuals, then inter-cluster dissimilarity, say $d(G, G')$ is defined in terms of $d(x, y)$. In contrast, the centroid method has inter-cluster dissimilarity from the first: $d(G, G') = \|v(G) - v(G')\|^2$ where $v(G)$ is the centroid of cluster G. It thus is based on a top-down scheme and inter-cluster dissimilarity is dependent on the Euclidean model.

There have been studies in AHC using asymmetric similarity measures (Okada and Iwamoto, 1996; Saito and Yadohisa, 2005). They use the bottom-up scheme such as an asymmetric average linkage. In contrast, we study two new and top-down linkage methods that use specific application-dependent models. They are called a *citation probability model* and a *multiset-theoretical model for text analysis*. We prove that they have no reversals in dendrograms. We also prove that asymmetric average linkage and another bottom-up linkage method have no reversals in dendrograms as well. A number of application examples are shown to see the effectiveness of the proposed methods.

References

OKADA, A. and IWAMOTO, T. (1996): A Comparison before and after the Joint First Stage Achievement Test by Asymmetric Cluster Analysis. *Behaviormetrika, 23, 169–185.*

SAITO, T. and YADOHISA, H. (2005): *Data Analysis of Asymmetric Structures*, Marcel Dekker, New York.

Keywords

AGGLOMERATIVE HIERARCHICAL CLUSTERING, ASSYMETRIC SIMILARITY MEASURES, REVERSALS IN DENDROGRAMS

Examination of the Necessity of Using One-mode Three-way Asymmetric Model

Atsuho Nakayama[1], Hiroyuki Tsurumi[2], and Akinori Okada[3]

[1] Graduate School of Social Sciences, Tokyo Metropolitan University, 1-1 Minami-Ohsawa, Hachioji-shi, Tokyo 192-0397, Japan atsuho@tmu.ac.jp
[2] Faculty of Business Administration, Yokohama National University, 79-1 Tokiwadai, Hodogayaku, Yokohama 240-8501 Japan tsurumi@ynu.ac.jp
[3] Graduate School of Management and Information Sciences, Tama University, 4-4-1 Hijirigaoka, Tama-shi Tokyo 206-0022, Japan okada@rikkyo.ac.jp

Abstract. Some models have been proposed to analyze one-mode three-way data. These models usually assume triadic symmetric relationships except for the model that was proposed by De Rooij and Heiser (2000). Therefore, Nakayama and Okada (2010) proposed a method that extended Harshman, Green, Wind, and Lundy (1982)'s reconstructed method to one-mode three-way asymmetric proximity data. So, the method reconstructs one-mode three-way asymmetric data so that the overall sum of the rows, columns and depths is made equal over all objects. However, the need for the analysis of asymmetric model has never been examined in one-mode three-way asymmetric data. In one-mode two-way asymmetric data, the need for the analysis of asymmetric model has been examined by test for symmetry and so on. Therefore, the present study proposes the method to evaluate the need for one-mode three-way asymmetric model.

References

De ROOIJ, M. and HEISER, W. J. (2000): Triadic Distances Models for the Analysis of Asymmetric Three-way Proximity Data.*British Journal of Mathematical and Statistical Psychology, 53, 99–119.*

HARSHMAN, R.A., GREEN, P.E., WIND, Y., and LUNDY, M.E. (1982): A Model for the Analysis of Asymmetric Data in Marketing Research. *Marketing Science, 1, 205–242.*

NAKAYAMA, A. and OKADA, A. (2010): Reconstructing One-mode Three-way Asymmetric Data for Multidimensional Scaling. In GFKL 2010 printed proceedings. Manuscript submitted for publication.

Keywords

MDS, ASYMMETRIC DATA, TRIADIC RELATIONSHIPS

Least Squares Permutation and Its Applications to Factor Rotation and Fixed Size Clustering

Kohei Adachi

Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka 565-0871, Japan. k-adachi@lt.ritsumei.ac.jp

Abstract. Least squares permutation (LSP) refers to obtaining the permutation matrix optimally matching a permuted matrix to a target matrix in a least squares sense. For LSP I present a simple iterative algorithm, which is applied for two multivariate classification problems. One is to rotate factor loading matrices toward simple structures. I propose a new rotation method PERMUTIMIN using LSP, which is viewed as a relaxed version of Procrustes methods (Gower and Dijksterhuis, 2004) and a constrained one of Simplimax (Kiers, 1994). The other application is fixed size clustering, that is, to classify observations while keeping their numbers allocated to each cluster fixed at prescribed constants. I present a fixed size clustering procedure with LSP, which can be formulated as a constrained K-means method (MacQueen, 1967).

References

GOWER, J.C. and DIJKSTERHUIS, G.B. (2004): *Procrustes problems*, Oxford: Oxford University Press.

KIERS, H.A.L. (1994): Oblique rotation to an optimal target with simple structure, *Psychometrika*, **59**, 567–579.

MACQUEEN, J.B. (1967): Some methods for classification and analysis of multivariate observations, *Proceedings of the 5th Berkeley Symposium, Vol. 1*, 281–297.

Keywords

LEAST SQUARE PERMUTATION, PERMUTIMIN ROTATION, FIXED SIZE CLUSTERING

The Effects of Q-matrix Specification and Misspecification in Multiple-choice DINA Models

Koken Ozaki[1], Ikko Kawahashi[2], Tomoko Takahashi[3], Yuan Sun[3] and Sumio Kakinuma[3]

[1] The Institution of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo, Japan koken@ism.ac.jp
[2] Japan Foundation, 4-4-1 Yotsuya, Shinjuku-ku, Tokyo, Japan
[3] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

Abstract. Cognitive diagnosis models are ones for estimating examinees' mastery or non-mastery of skills that are needed to answer items correctly. Skills can be thought of as binary latent variables. Therefore, cognitive diagnosis models are one of the latent class models. To estimate latent classes, binray response data (correct or incorrect) by examainees, some cognitive diagnosis model such as the deterministic input noisy "and" gate (DINA) model, and a Q-matrix are needed. The size of a Q-matrix is the number of items by the number of skills. And its element is 1 if an item needs a skill, otherwise it is 0. The elements of a Q-matrix are not the parameters to be estimated but have to be specified before estimating latent classes. However, it is sometimes difficult to provide an appropriate Q-matrix. Therefore, the studies examining the effects of Q-matrix specification and misspecification on the estimated latent classes such as Rupp and Templin (2008) are important in this field. Recently, cognitive diagnosis models for multiple-choice items are developed. However, the effects of Q-matrix specification and misspecification for multiple-choice models have never been examined. In this study, simulation studies are perfomed to examine these points. And the results are compared with those of the models for binary data.

References

Rupp, A.A., and Templin, J. (2008): The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model, *Educational and Psychological Measurement, 68, 78–96.*

Keywords

ACADEMIC APTITUDE TEST, COGNITIVE DIAGNOSIS MODEL, MULTIPLE-CHOICE ITEM, Q-MATRIX, SIMULATION STUDY

Automatic Regularization of Factorization Models

Steffen Rendle

University of Konstanz steffen.rendle@uni-konstanz.de

Abstract. Many recent machine learning approaches for prediction problems over categorical variables are based on factorization models like matrix or tensor factorization. Due to the large number of model parameters, factorization models are prone to overfitting and typically Gaussian priors are applied for regularization. Finding proper values for the regularization parameters is usually done with an expensive grid-search using holdout validation data. In this talk, two alternatives are presented where regularization values are found without increasing computational complexity. The first one is based on interweaving optimization of model parameters and regularization. The second one uses a two-level Bayesian model with MCMC to integrate regularization values into inference. The approaches are studied on applications from recommender systems and educational data mining.

Thresholding loadings in factor analysis.

MIYAMOTO, Yusuke

Graduate School of Human Sciences, Osaka University
Yamadaoka 1-2, Suita, Osaka 565-0871, Japan my@hus.osaka-u.ac.jp

Abstract. An criterion for thresholding and rotation of loadings in factor analysis model is proposed. Factor analysis has been widely used to classify variables into correlating groups. Factor loadings, with certain normalisation, are represented as a stochastic matrix, and could be regarded as fuzzy membership of the variables in each factor. Reformulating factor rotation as a clustering problem, we can take account of thresholding loadings. The basic idea is derived from existing rotation criteria, varimax (Kaiser, 1958), simplimax (Kiers, 1994), and component loss function (Jencrich, 2004, 2006), and it is shown how their performances are affected by the thresholdings.

References

Jenrich, R. I. (2004): Rotation to simple loadings using component loss functions: the orthogonal case. *Psychometrika, 69, 257–273.*

Jenrich, R. I. (2006): Rotation to simple loadings using component loss functions: the oblique case. *Psychometrika, 71, 173–191.*

Kaiser, H. F. (1958): The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 1958, 23, 187–200.*

Kiers, H. A. L. (1994): SIMPLIMAX: Oblique rotation to an optimal target with simple structure. *Psychometrika, 59, 567–579.*

Keywords

thresholding, factor rotation, varimax, simplimax, CLF

Assessment of the relationship between native thoracic aortic curvature and endoleak formation after TEVAR based on linear discriminant analysis

Kuniyoshi Hayashi[1,5], Fumio Ishioka[2,5], Bhargav Raman[3], Daniel Y. Sze[3], Hiroshi Suito[1,5], Takuya Ueda[4,5] and Koji Kurihara[1,5]

[1] Graduate School of Environmental Science, Okayama University
k-hayashi@ems.okayama-u.ac.jp, suito@ems.okayama-u.ac.jp, kurihara@ems.okayama-u.ac.jp
[2] School of Law, Okayama University fishioka@law.okayama-u.ac.jp
[3] Department of Radiology, Stanford University School of Medicine
ramanb@stanford.edu, dansze@stanford.edu
[4] Department of Radiology, St. Luke's International Hospital takueda@luke.or.jp
[5] CREST, Japan Science and Technology Agency

Abstract. Endoleak is one of the adverse clinical side effects after thoracic endovascular aortic repair (TEVAR), a treatment for thoracic aortic disease. Risk prediction of endoleak is essential for pre-operative planning (Nakatamari *et al.*, 2011). In this study, we focus on the two-class discriminant problem of no-endoleak and endoleak, and evaluate the relationship between the native thoracic aortic curvature of a patient and the endoleak formation on the basis of linear discriminant analysis. In addition, we assess the result by applying statistical sensitivity analysis to the estimated discriminant model.

References

NAKATAMARI, H., UEDA, T., ISHIOKA, F., RAMAN, B., KURIHARA, K., RUBIN, G.D., ITO, H. and SZE, D.Y. (2011): Discriminant Analysis of Native Thoracic Aortic Curvature: Risk Prediction for Endoleak Formation After Thoracic Endovascular Aortic Repair. *Journal of Vascular and Interventional Radiology, 22, 974–979.*

Keywords

CURVATURE INDEX, LEAVE-ONE-OUT CROSS-VALIDATION, QUANTITATIVE ANALYSIS OF AORTIC MORPHOLOGY

Three-way Data Analysis for Multivariate Spatial Time Series

Mitsuhiro Tsuji[1] and Toshio Shimokawa[2]

[1] Kansai University, Takatsuki, Osaka, Japan. tsuji@kansai-u.ac.jp
[2] University of Yamanashi, Kofu, Yamanashi, Japan. shimokawa@yamanashi.ac.jp

Abstract. We would discuss methods for three-way (three mode) approaches to clustering INDCLUS and multidimensional scaling INDSCAL which assumed the objects are embedded in a discrete and continuous space common to all data including individual differences by weighting each dimension. We applied some effective dynamic graphical approach with two methods to make the time space structural analysis for multivariate spatial time series. Clustering and scaling of multivariate spatial time series would consider: 1) the spatial nature of the objects to be clustered geometrically (discrete) ; 2) the characteristics of the feature space with time series (continuous) ; 3) the latent structure between space and time. The last aspect is dealt with using the dynamic graphics with matrix type presentation. Then we can look at the spatial nature at the same time, we can move the feature space at the same time, and we can zoom in/out some results into their suitable size. The proposed analysis would be applied to the classification and scaling of the prefectures of Japan, on the basis of the observed dynamics of some safety indicators.

References

ARABIE, P., CAROLL, J. D., and DESARBO, W. S. (1987): *Three-way Scaling and Clustering*. Sage, Newbury Park.

COPPI, R., D'URSO, P. and GIORDANI, P. (2010): A Fuzzy Clustering Model for Multivariate Spatial Time Series. *Journal of Classification, 27, 54–88.*

TSUJI, M. SHIMOKAWA, T. and OKADA, A. (2010): Three-way Scaling and Clustering Approach to Musical Structural Analysis. In Locarek - Junge, H. and Weihs, C. (Eds.): *Classification as a Tool for Research, Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag Berlin Heidelberg, 767-777

Keywords

Individual Clustering, Individual Scaling, Dynamic Graphics, GIS

Three-Mode Subspace Clustering for Considering Effects under Noise Variables and Occasions

Kensuke Tanioka[1] and Hiroshi Yadohisa[2]

[1] Graduate School of Culture and Information Science, Doshisha University, Japan dik0012@mail4.doshisha.ac.jp
[2] Department of Culture and Information Science, Doshisha University,Japan hyadohis@mail.doshisha.ac.jp

Abstract. Recently, improvements in information technology allow us to deal with large and complex data. It is difficult to apply classical clustering methods to such large and complex data since they tend to have noise variables and the results mey be affected by dimensionality curse. Although three-mode data are also affected by such noises and are collected in various domains such as marketing science and web mining, there are a few clustering methods that consider three-mode data with noise (Vichi, et al., 2007).

Vichi, et al. proposed two methods: the T3Clu and 3Fk-means. In these methods, clustering and factorial analyses are simultaneously applied to three-mode data. These methods allow us to easily interpret the results, evenif the three-mode data have a high dimensionality. However, these clustering results that are based on the factorial model are affected by noises (Lance, et al., 2004) such as masking variables, multiple cluster structures and so on. On the other hand, the feature selection methods detect clusters in data subspace to consider such noise problems.

In this paper, we propose new subspace clustering method for three-mode data by using feature selection methods.

References

LANCE, P., EHTEASHAM, H. and HUAN, L. (2004): Subspace Clustering for High Dimensional Data:A Review. *SIGKDD Explorations*, **6**(1), 90–105.

VICHI, M., ROCCI, R. and KIERS, H.A.L. (2007): Simultaneous component and clustering models for three way-data: Within and between approaches. *Journal of Classification*, **24**(1), 71–98.

Keywords

THREE-MODE DATA, CLUSTERING, SUBSPACE, MASKING VARIABLE

Classification, Clustering and Visualisation Based on Dual Scaling

Hans-Joachim Mucha

Weierstrass Institute, Berlin mucha@wias-berlin.de

Abstract. In practice, the statistician is often faced with data already available. In addition, there are often mixed data. The statistician must now try to gain optimal statistical conclusions with most sophisticated methods. But, are the variables scaled optimally? And, what is with missing data? Here an approach is outlined, applicable to both classification and clustering, based on preceding data preparation of each single variable by dual scaling (Nishisato, 1980). As a byproduct, the scores can be used for multivariate visualisation of both data and classes/clusters.

First, a binary classification method is proposed that processes both mixed data (numeric, ordinal and categorical) and missing values by the dual scaling technique (Mucha, 2002). The latter enables us to establish an appropriate multivariate space that gives categorical data a quantitative meaning. Diverse multivariate graphics allow you to communicate - visually and quickly - the numeric story the proposed statistical analysis is telling.

In the second part of the contribution, an unsupervised binary classification method (clustering) is outlined based on the dual scaling technique. Finally, the paper deals with an investigation of the stability of the results by bootstrapping techniques such as data jittering. Throughout the paper, an application of the proposed methods to optical character recognition (OCR) is presented. Also, an application to cancer research will be presented in order to convince you about the practicality.

References

Mucha, H.-J. (2002). An intelligent clustering technique based on dual scaling. In: Measurement and Multivariate Analysis, S. Nishisato, Y. Baba, H. Bozdogan, K. Kanefuji, eds., Springer, Tokyo, 37-46.

Nishisato, S. (1980). Analysis of Categorical Data: Dual Scaling and Its Applications. The University of Toronto Press, Toronto.

Structural Representation of Categorical Data and Cluster Analysis Through Filters

Shizuhiko Nishisato

University of Toronto, Canada shizuhiko.nishisato@utoronto.ca

Abstract. Representation of categorical data by nominal measurement leaves the entire information intact, which is not the case with widely used numerical or pseudo-numerical representation such as Likert-type scoring. This aspect is first explained, and then we turn our attention to analysis on nominally represented data. For analysis of a large number of variables, one typically resorts to dimension reduction, and its necessity is often greater with categorical data than with continuous data. In spite of this, Nishisato and Clavel (2010) proposed an approach which is diametrically opposite to the dimension-reduction approach, for they advocate the use of doubled hyper-space to accommodate both row variables and column variables of two-way data in common space. The rationale of doubled space can be used to vindicate the validity of the Carroll-Green-Schaffer scaling (1986). The current paper will then introduce a simple procedure for analysis of a hyper-dimensional configuration of data, called cluster analysis through filters. A numerical example will be presented to show a clear contrast between the dimension-reduction approach and total information analysis by cluster analysis. There is no doubt that the our approach is preferred to the dimension-reduction approach on two grounds: our results are a factual summary of a multidimensional data configuration, and our procedure is simple and practical.

References

CARROLL, J. D., GREEN, P. E. and SCHAFFER, C. M. (1986). Inter-point distance comparisons in correspondence analysis. *Journal of Marketing Research*, 23, 271-280.

NISHISATO, S. and CLAVEL, J. G.. (2010): Total informaiton analysis: Comprehensive dual scaling. *Behaviormetrika*, 57, 15–32.

Variable Selection in *K*-means Clustering via Regularization Method

Yutaka Nishida

Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka, Japan nishida@bm.hus.osaka-u.ac.jp

Abstract. Recently, high dimensional data is treated in many fields such as text mining, brain science, marketing science and so on. In many cases, both essential and irrelevant variables to the data structure are included in the data set. The K-means algorithm (MacQueen, 1967) can treat such high dimensional data, but can't distinguish which variable is essential to the data structure. In supervised-learning methods such as regression analysis, variable selection is a major topic. However, variable selection in clustering is not an active area. In this study, a new method of K-means clustering is proposed to detect irrelevant variables to the data structure. The proposed method achieves the purpose of calculating variable weights using an entropy regularization method (Miyamoto & Mukaidono, 1997) which is developed to obtain fuzzy memberships in fuzzy clustering. This method allows us to identify the important variable for clustering.

References

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Miyamoto, S. and Mukaidono, M. (1997). Fuzzy c-means as a regularization and maximum entropy approach. *Proceedings of the 7th International Fuzzy Systems Association World Congress*, Vol.2, 86–92.

Keywords

K-MEANS CLUSTERING, REGULARIZATION, VARIABLE SELECTION

On the Stress Function of Asymmetric von Mises Scaling

Kojiro Shojima

The National Center for University Entrance Examinations
2-19-23 Komaba, Meguro-ku, Tokyo 153-8501, Japan
shojima@rd.dnc.ac.jp

Abstract. Asymmetric von Mises Scaling (AMISESCAL; Shojima, 2011) is an asymmetric multidimensional scaling that is used for analyzing an asymmetric proximity data matrix. It can express an asymmetric relationship by using a von Mises (vM) distribution in directional statistics (Mardia, & Jupp, 2000). Consider an example of analyzing data for a sociometric matrix; if Persons A and B like each other, their coordinate estimates in a multidimensional space are located close to each other. Further, if Person C likes Person D but Person D does not like Person C, their coordinates are estimated to be located away from each other, and the mean direction parameter of the vM distribution associated with Person C looks towards the coordinate of Person D.

The objective of this study was to improve the stress function of AMISESCAL proposed by Shojima (2011). This was done as follows. First, a function to prevent the degeneration of coordinate estimates was added to the stress function. Second, a function to penalize the stress function in the case that the mean direction parameter of the vM distribution of each element looks towards where there is no one-sided relation was added to the stress function. We confirmed that addition of these two functions to the stress function improved the readability of the map after AMISESCAL analysis.

References

MARDIA, K. V. & JUPP, P. E. (2000): Directional Statistics. John Wiley and Sons.

SHOJIMA, K. (2011): Asymmetric von Mises scaling. Paper presented in the proceedings of the 39th annual meeting of the Behaviormetric Society of Japan, Okayama University of Science, pp.261-262.

Keywords

ASYMMETRIC MULTIDIMENSIONAL SCALING,
DIRECTIONAL STATISTICS, VON MISES DISTRIBUTION

Bayesian Analysis of Asymmetry by the Slide-Vector Model

Kensuke Okada[1]

[1] Senshu University, 2-1-1, Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8580, Japan ken@psy.senshu-u.ac.jp

Abstract. Since the proposal of the first general Bayesian multidimensional scaling (MDS) by Oh & Raftery (2001), Bayesian analysis of MDS using Markov chain Monte Carlo (MCMC) algorithm has recently received considerable attention. However, most of the existing Bayesian MDS methods assume symmetric data matrix as input. In this paper, we propose Bayesian analysis of asymmetry by the slide-vector model, which was originally proposed by Kruskal in 1973 and elaborated by Zielman & Heiser (1993). In symmetric MDS, the distance function is given by

$$d_{ij}(\mathbf{X}) = \sqrt{\sum_s (x_{is} - x_{js})^2}. \tag{1}$$

On the other hand, the definition of the distance in the slide-vector model is given by

$$d_{ij}(\mathbf{X}, \mathbf{z}) = \sqrt{\sum_s (x_{is} + z_s - x_{js})^2}, \tag{2}$$

where the vector $\mathbf{z}$ is called the slide vector. We use a multivariate normal prior on $\mathbf{z}$; the priors for the rest of the parameters are taken from the existing Bayesian MDS of symmetric data. A Markov chain Monte Carlo (MCMC) algorithm is used to explore the posterior distributions.

References

OH, M-S. and RAFTERY, A. E. (2001): Bayesian Multidimensional Scaling and Choice of Dimension. *Journal of the American Statistical Association, 96, 1031–1044.*

ZIELMAN, B. and HEISER, W. J. (1993): Analysis of Asymmetry by a Slide-Vector. *Psychometrika, 58, 101–114.*

Keywords

BAYESIAN INFERENCE, MULTIDIMENSIONAL SCALING, SLIDE-VECTOR MODEL, ASYMMETRY

A New Multidimensional Scaling Methodology for Analysis of Asymmetric Citation Data in Scientific Publications

Yuan Sun[1] and Tadashi Imaizumi[2]

[1] National Institute of Informatics,2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 yuan@nii.ac.jp
[2] Tama University, 4-1-1 Hijirigaoka, Tama-shi, Tokyo 206-0022 imaizumi@tama.ac.jp

Abstract. Citations analysis is one of the standard methods to study research activities in bibliometrics field. Number of citations are used as indicators to represent the relations among researches, thus relations between the authors, affiliations the authors belong to, and so forth. In this study, we take a look at relations among Japanese academic societies by using citation data of publications published in the journals of the societies. The data is extracted from the Citation Database for Japanese Papers (CJP), produced by the National Institute of Informatics (NII). Journals of 14 Japanese academic societies which have strong citing or cited relations with each other have been chosen for the analysis. Sun and Negishi(2003) analyzed this asymmetric citation data by a classical Multi-Dimensional Scaling(MDS) method with loss of information in data which is undesirable. This paper presents a new MDS methodology which based on Asymmetric MDS by Okada and Imaizumi(1987) ,and is modified as their model is applicable in the bibliometrics field.

References

OKADA, A and IMAIZUMI, T.(1987): Nonmetric Multidimensional Scaling of Asymmetric Proximites.*Behaviormetrika, 21,81-96*

SUN, Y. and NEGISHI, M. (2003): Measuring of the Relationship among the Japanese Academic Societies based on Citation Data from the CJP Database, *Proceeding of 11th conference of Japan Society of Information and Knowledge, 5-8 (in Japanese).*

Keywords

CITATION MATRIX, DEGREE OF SIMILARITY, ASYMMETRIC PROXIMITY, ASYMMETRIC MULTIDIMENSIONAL SCALING

Asymmetric Multidimensional Scaling with Generalized Hyperellipse Model

Yoshikazu Terada[1] and Hiroshi Yadohisa[2]

[1] Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, Japan terada@sigmath.es.osaka-u.ac.jp
[2] Department of Culture and Information Science, Doshisha University, Japan hyadohis@mail.doshisha.ac.jp

Abstract. In this paper, we extend the hyperellipse model of Okada (1990) to represent each object as a more general hyperellipse and we propose a gradient based algorithm for this model. In general, a p-dimensional hyperellipse which has center $\boldsymbol{c}$ and radii $\boldsymbol{r}$ is defined by $(\boldsymbol{x}-\boldsymbol{c})^T\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{c})=1$, where $\boldsymbol{A}$ is a symmetric positive definite matrix which has eigenvalues r_s^2 $(s=1,\ \ldots,\ p)$. Let ξ_{ij} be a given dissimilrity of object i to j $(i,\ j=1,\ \ldots,\ n)$. In our model, each objects is represented by a general hyperellipse which has center $\boldsymbol{x}_i$ and radii $\boldsymbol{r}_i$ in a low dimensional space, in such a way that asymmetric dissimilarities are approximated by

$$\xi_{ij} \approx d_{ij}(1-v_{ij}+v_{ji}),$$

where $d_{ij}=\left(\sum_{s=1}^{p}\left(x_{is}-x_{js}\right)^2\right)^{1/2}$, $v_{ij}=\left(\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)^T\boldsymbol{A}_i^{-1}\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)\right)^{-1/2}$. Fig. 1 illustrates the relationships in terms of an asymmetric dissimilarity between two objects in the model.

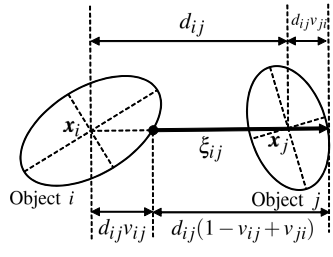


Fig. 1. Relationships in terms of the dissimilarity between objects i and j

References

OKADA, A. (1990): A generalization of asymmetric multidimensional scaling. In: M. Schader & W. Gaul (Eds.): *Knowledge, data and computer-assited decisions.* Springer, Berlin, 127–138.

Keywords

ASYMMETRIC PROXIMITY, MDS, HYPERELLIPSE MODEL

Statistical Process Modelling for Machining of Inhomogeneous Mineral Subsoil

Claus Weihs

Technische Universitat Dortmund weihs@statistik.uni-dortmund.de

Abstract. Machining of concrete with diamond tools is characterized by its very heterogeneous structure of the workpiece as well as the related tools. Complicated engagement conditions arise which can only be monitored by statistical methods. An optimization of the machining process is difficult due to lack of knowledge of the complex interactions between tool and workpiece structure.

Core drills hold cutting segments consisting of randomly distributed grains in a binder phase. In a first phase the influences of the process parameters cutting speed, depth of cut and diameter of the produced hole on the quality measures chip removal rate, disruptions, tool wear and forces in both axial and radial direction are investigated. To be able to gain process knowledge on a micro level, this is obtained by conducting experiments with solely one single grain. These single grain scratch experiments are planned by a Central Composite Design blocked on cutting edge geometry.

In this early phase linear regression models are applied to identify the process parameters with significant influence on the quality measures. Within these models, the two-dimensional grain orientation, which can be measured but not controlled, is used as a covariate. Furthermore the possibility of using the structure borne sound as a proxy for tool wear is investigated. In later study phases the results of these investigations will be generalized to the macro level of multi grain experiments. Density and distribution of the grains will then be taken into account as additional parameters influencing process quality.

Zone Detection Process for Rainfall Inflow into Sewage Pipe

Ken Wada[1], Tomoyuki Hasegawa[1] and Keiji Yajima[2]

[1] STONEGATE Co., Chuoh 3-3-32, Ebina, Kanagawa 243-0432 JAPAN
(s.kenwada, s.hasegawa)@stonegate.co.jp
[2] Independent, Nakayama 1-7-14, Ichikawa, Chiba 272-0813 JAPAN
k2yjm@icnet.ne.jp

Abstract. At present the separate sewer system consists of rainfall pipe and the sewage duct is dominant and for this system big rainfall inflow into sewage pipe causes failure of the sewage plant. For the maintenance of the drainage system rainfall inflow is a big subject next to the decrepitude of the pipe system.

The rise of inflow into the sewage plant related with rainfall causes disorder of the sewer plant and then increase of the operating cost. To cope with the rainfall inflow we need to grasp the overall situation and then to detect the main inflow zone.

Traditional process consists of following operations such as collection of flood information, determination of the survey zone, and planning of continual researches to measure flow flux rainfall. So it takes a couple of years and lot of money for the measurement. It arrises quite naturally the overall processing of the past plant operation data is inevitable.

Zone detection process(mesh data system) comes out and has following steps.

1. Basic survey: collection of operating data including rainfall per 500 meter mesh domain using meteological radar survey.
2. Drawing of the pipeline and street map: by using of the GIS(geographic information system)
3. Predominant rainfall inflow detection: finding of the predominant inflow domain, determination of the inflow estimating spots.
4. A general purpose software tool called as Dr.TCBM is prepared for the mesh data analysis. Started from mesh data it provides mesh-wise functional representation (usually non-linear form) of flux with respect to time, then it supplies time dependent flow flux at the clear weather condition from which additional rainfall flux quantity could be found.

The traditional system and the mesh data system comparison table is shown.

	Traditional System	Mesh Data System
Expenditure	20-30 $\times$ 10^6 Yen, 3 years	10 $\times$ 10^6Yen, 6 months
No. of rainfalls used	1-2	almost 300(3 years data)
Rainfall power effect	great	small(using total drainage)

The Utility of Smallest Space Analysis for the Cross-National Survey Data Analysis: The Structure of Religiosity

Kazufumi Manabe

School of Cultural and Creative Studies, Aoyama Gakuin University, JAPAN
kazufumi.manabe@nifty.com

Abstract. The purpose of this paper is to illustrate the utility of Smallest Space Analysis (SSA) developed by Louis Guttman using the examples of National Religion Surveys conducted in Japan (2007) and Germany (2008) by the research team organized by the author.

As a type of multidimensional scaling, SSA is a method of expressing the relationship between n number of question items shown in a correlation matrix by the size of the distance between n points in an m-dimensional (m¡n) space. The higher the correlation, the smaller the distance, and the lower the correlation, the greater the distance. Usually a 2-dimensional (plane) or 3-dimensional (cube) space is used to visually depict the relationship between question items. This shows that SSA is the most appropriate method of visually depicting the overall structure and relationships among question items.

SSA is to be applicable as a very effective tool in examining equivalence of measurement when conducting cross-national surveys. In cross-national surveys that can yield SSA maps showing the same spatial structures, it is highly likely that "commonality of meanings" can likewise be established in the countries compared. This is an important reason for using SSA.

Thus we illustrate the utility of SSA for the cross-national comparison of religiosity in Japan and Germany.

References

LEVY, S. (Eds.) (1994): *Louis Guttman on Theory and Methodology, Selected Writings*, Dartmouth, Aldershot.

MANABE, K. (2001): *Facet Theory and Studies of Japanese Society*. Bier'sche Verlangsanstalt, Bonn.

Keywords

SMALLEST SPACE ANALYSIS, CROSS-NATIONAL SURVEY, RELIGIOSITY

Analysis of Changes of Brand Categories Using Purchase History Data and Eigenvalue to Find New Category

Yuki, Toyoda,[1]

Tama University toyoda@tama.ac.jp

Abstract. Brand managers looking for a Blue Oceans (in there, demand is created rather than fought over and there is ample opportunity for profit), because brands compete with other brands in a Red Oceans (those are more competitive market with low level of profitability).

To find blue oceans, brand managers need to interpret the composition and changes of brand categories. 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.

The aim of our study (at GJSC2010 and JGSC2011) is to propose a method required for brand management to understand the composition and changes of categories in it as easily as possible.

For this aim, at GJSC2010 Toyoda and Imaizumi propose a method using Cronbach's coefficient alpha, but it is not clear a easiness to interpret output. With this in mind, at JGSC2011 I improved a method with eigenvalue and to applly it to foodstuffs (including beverages) for which consumer selection behavior is diverse. Its effectiveness and problems are thus examined.

Keywords

SEGMENTATION, PURCHASE HISTORY DATA, BRNAD CATEGORY, BRAND MANEGEMENT, EIGENVALUE

An Automatic Extraction of Academia-Industry Collaborative Research and Development Documents on the Web

Kei Kurakawa[1], Yuan Sun[2], Nagayoshi Yamashita[3] and Yasumasa Baba[4]

[1] National Institute of Informatics kurakawa@nii.ac.jp
[2] National Institute of Informatics yuan@nii.ac.jp
[3] Japan Society for the Promotion of Science nagayoshi3@gmail.com
[4] The Institute of Statistical Mathematics baba@ism.ac.jp

Abstract. To make a policy of science and technology research and development, university-industry-government relationship is an important aspect to investigate it. Web document is one of the research targets to clarify the state of the relationship. Our research focuses on automatic extraction method of Japanese web documents describing academia-industry collaboration through web crawling. The method we adopt is natural language processing(morphological analysis, structural dependency) for Japanese and classification techniques of machine learning. We seek structural features of documents from the viewpoint of academia-industry collaboration, and apply them to these general machine learning framework for text categorization.

References

Loet Leydesdorff, Martin Meyer (2003): The Triple Helix of university-industry-government relations *Scientometrics 58 (2) p. 191-203.*

MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/

Taku Kudo, Yuji Matsumoto (2002): Japanese dependency analysis using cascaded chunking. In: Dan Roth and Antal van den Bosch (eds.), *Proceedings of CoNLL-2002, Taipei, Taiwan, pp. 63-69.* http://code.google.com/p/cabocha/

Vladimir N. Vapnik (1995): *The nature of statistical learning theory* Springer-Verlag New York, Inc. New York, NY, USA

Keywords

ACADEMIA INDUSTRY COLLABORATION, WEB DOCUMENTS, TEXT CATEGORIZATION, MACHINE LEARNING

Ancient Population Dynamics Estimation from Archaeological data 'Nuzi personal names'

Sumie Ueda[1], Kumi Makino[2], Yoshiaki Itoh[3] and Takashi Tsuchiya[4]

[1] The Institute of Statistical Mathematics, 10-3 Midori-cho Tachikawa Tokyo
ueda@ism.ac.jp
[2] Kamakura Women's University, 6-1-3 Oofuna Kamakura Kanagawa
kumi@kamakura-u.ac.jp
[3] The Institute of Statistical Mathematics itoh@ism.ac.jp
[4] National Graduate Institute for Policy Studies, 7-22-1 Roppongi Minato-ku Tokyo tsuchiya@grips.ac.jp

Abstract. We estimate population of an ancient Mesopotamia city Nuzi from the database 'Nuzi personal names'. Nuzi is a city existed in the neighbor of Kirkuk, Iraq, around B.C. 15 century. The city lasted about 100 years. Nuzi personal name is an index of individuals appeared in social contracts written on clay tablets (Gelb et al. (1943), Makino (1991)). Population estimation is done by reconstructing family trees from Nuzi personal names (Ueda et al. (2005)) and then by allocating the obtained family trees along the time axis. In this process, we developed two novel approaches based on mathematical and computer sciences. The first method determines the generation of each person by matching two family trees based on the information of documents shared by the two families. The other utilizes the method of mathematical optimization, namely, Nuzi population is estimated by solving an optimization problem formulating the relation between family trees and documents. The both methods give consistent results and we conclude that Nuzi population of those days was around 18,000.

References

Gelb, I.J., Purves, M.P. and Macrae A.A. (1943): *Nuzi Personal Names*. The University of Chicago Press.

Makino, K. (1991): Social changing on the adoption contract (in Japanese). *SHIGAKU, Vol.60, pp.91-119.*

Ueda, S., Itoh, Y. and Makino, K. (2005): Reconstructing family trees of ancient population form Nuzi personal names (in Japanese). *Proceedings of the Institute of Statistical Mathematics, 53-2, pp.285-295.*.

Keywords

Nuzi personal names, family tree, generation estimation, population estimation, convex quadratic programming

Classification of Literature by analyzing Figure-ground relationship of Characters

Tetsuya Matsui Yukio-pegio Gunji Eugenio-Schneider Kitamura

Kobe university, 1-1,Rokkoudai,Nada,Kobe,Hyogo
Matsui@081s416s@stu.kobe-u.ac.jp Gunji@pegioyukio@gmail.com
Kitamura@kitamura@godzilla.kobe-u.ac.jp

Abstract. In this reserch we suggest a method to classify literatuers by figure-ground relation of their characters. Figure-ground relationship is a notion that defined the noticed objects in cognition as "figure" and defined the rest of the objects as "ground". In this research we define the character whose have unique verbs in a particular scene as "figure" and other characters as "ground". We compare two figure-ground relationships in one literary work. One relationship is based on verbs that actually take place in the scene. The other is based on verbs mentioned in speech in the scene. We devide a literary work into some small scenes and estimate these two figure-ground relationships in each scenes. To estimate figure-ground relationship we use a notion of lattice and rough set. We make some set that includes one or some characters that they have identical verb or set of verbs. These sets make one lattice by ordered by inclusion relation. We made two series of lattices based on verbs that actually take place and verbs mentioned in speech. We can classify literary works based on a relation of these two lattices. We use 8 literary works witten by Kenji Miyazawa, a Japanese writter in the early 20th century. We find they can be classified into at least two categories. One category includes literary works that thier two lattices have a few difference as a whole. The other category includes literary works that thier two lattices have a large different as a whole. Most of the literary works of former category is witten by Miyazawa in early years and most of the literary works of later category is witten in later years. This result shows this method is effective to classify literary works witten by a particular writter.

References

Gunji, Y.-P., and Haruna, T., (2010): "A Non-Boolean Lattice Derived by Double Indiscernibility", Transactions on Rough Sets XII, LNCS, Vol. 6190, pp. 211-225.

Keywords

Cultuer, Literature, Figuer-ground relationship, Lattice

Non-Additive Utility Functions: Choquet Integral versus logic-based Querying

Ingo Schmitt

Brandenburgische Technische Universität Cottbus, Germany
schmitt@tu-cottbus.de

Abstract. In the context of conjoint analysis, consumer´s purchase preferences w.r.t. a product selection can be modeled by use of a utility function. Based on the attribute values of a certain product, a utility function returns a utility value (real number). A product with a higher utility value is preferred to a product with a lower value. Simple additive utility functions, however, do not sufficiently reflect real-world decisions where dependencies between product attributes occur. In recent years, different non-additive functions have been developed including an approach based on a Choquet integral and one based on a query of a logic based query language. After introducing both non-additive approaches we will compare them w.r.t. their expressivity. We will develop a conceptual link between both approaches. For a more realistic modeling of preferences, the link allows us to exploit concepts from both approaches simultaneously.

Feature Selection and Clustering of Digital Images Versus Questionnaire Based Grouping of Consumers: A Comparison

Ines Daniel and Daniel Baier

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

Abstract. Clustering algorithms are standard tools for marketing purposes. So, e.g., in market segmentation, they are applied to derive homogeneous customer groups. However, recently, the available resources for this purpose have extended. So, e.g., in social networks potential customers provide images which reflect their activities, interests, and opinions. To compare whether images lead to similiar results as conventional methods for lifestyle analysis, a comparison study was conducted among 495 people. In this paper we discuss the results of the study. We also analyze possible advantages and disadvantages of using images for lifestyle analysis compared to conventional procedures of grouping customers for market segmentation.

References

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.*

VAN HOUSE, N.A. (2009): Collocated Photo Sharing, Story-Telling, and the Performance of Self. *International Journal of Human-Computer Studies, 67(12), 1073–1086.*

WEDEL, M., KAMAKURA, W.A. (2000): Market Segmentation: Conceptual and Methodological Foundations. Kluwer, Dordrecht.

WELLS, W.D., TIGERT, D.J. (1971): Activities, Interests, and Opinions. *Journal of Advertising Research, 11(4), 27–35.*

Keywords

MARKET SEGMENTATION, IMAGE CLUSTERING ALGORITHMS

Analysis of Asymmetric Relationships Among Soft Drink Brands

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 brand switching data among eight soft drink brands were analyzed. The eight brands were characterized by three attribute; cola/lemon-lime, diet/non diet, and coca cola/pepsi. The brand switching data are represented by an 8×8 brand switching matrix. The brand switching matrix is inevitably asymmetric, because the relationship from brand j to k is not necessarily equal to the relationship from brands k to j. The brand switching matrix was normalized so that the sum of each row is unity (Bass, Pessemier, and Lehmann, 1972; DeSarbo, and De Soete, 1984). The brand switching matrix was analyzed by asymmetric multidimensional scaling based on the singular value decomposition (Okada and Tustumi, in press) and asymmetric cluster analysis. The result obtained by the asymmetric multidimensional scaling shows the outward tendency, which represent the strength to be switched from the corresponding brand and the inward tendency, which represent the strength to be switched to the corresponding brand. The result also shows the difference between the diet and non diet brands. The result obtained by the asymmetric cluster analysis confirms the result given by the asymmetric multisimensional scaling.

References

BASS, F. M., PESSEMIER, E. A., and LEHMANN, D. R. (1972): An Experimental study of Relationships Between Attitudes, Brand Preference, and Choice. *Behavioral Science, 17, 532–541.*

DeSARBO, W. S., and De SOETE, G. (1984): On the Use of Hierarchical Clustering for the Analysis of Nonsymmetric proximities. *Journal of Consumer Research, 11, 532–610.*

OKADA, A., and TUSRUMI, H. (in press): Asymmetric Multidimensional Scaling of Brand Switching Among Margarine Brands. *Behaviormetrika.*

Keywords

ASYMMETRY, BRAND SWITCHING, CLUSTER ANALYSIS, MULTIDIMENSIONAL SCALING, SINGULAR VALUE DECOMPOSITION

How to use Willingness-to-Pay Data for Product Bundling

Wolfgang Gaul
Institut für Entscheidungstheorie und Unternehmensforschung
KIT-Campus Süd, 76128 Karlsruhe
e-mail: wolfgang.gaul@kit.edu

Anticipation of consumers' choice behavior and segmentation of possible customers according to their willingness-to-pay are salient tasks within strategies of firms concerning designing and pricing of products. In this context firms have also options to select subsets of products and sell these subsets as so-called "bundles". Here, knowledge about consumers' willingness-to-pay data plays a crucial role for determining profitable firm strategies. A branch & cut algorithm that simultaneously takes into account the principle of customers' surplus maximization and the optimization of firm's profit is presented. Based on product-specific willingness-to-pay data an example (that shows which product bundles should be offered at which prices to optimize the profit of a firm) is used to explain our findings.

From Online Customer Reviews to New Marketing Insights
Methodological Issues and Challenges

Reinhold Decker

Department of Business Administration and Economics, Bielefeld University,
PO Box 10 01 31, 33501 Bielefeld, Germany

Email: rdecker@wiwi.uni-bielefeld.de

The systematic analysis of online customer reviews, as a new access to consumer opinions, is a promising option for marketing research devoted to products that are frequently subject to electronic word-of-mouth. Typically, these reviews (see, e.g., amazon.com, ciao.de or epinions.com) consist of three elements – product ratings and/or recommendations, pros and cons as well as full texts. In the ideal case, linking these elements together should provide a clear and consistent picture of consumers' opinions and preferences regarding the product category of interest.

However, comprehensive investigations of online customer reviews across brands and borders indicate that the conclusions drawn from the relevant elements may diverge. The extent of this phenomenon and the respective patterns are investigated using an econometric framework that takes into account the specific structure of the data considered. The latter particularly concerns the unknown but potentially mattering features appearing in the full texts. As an empirical basis for this research we use a large data set which contains several thousands of customer reviews posted in different countries. The available results explicitly underline the usefulness of a holistic approach and, therewith, complement recent findings in this field of marketing data analysis.

Authors