Data Science and Classification

10th Jubilee Conference of the
International Federation of Classification Societies

Program and Abstracts

July 25 – 29, 2006

Ljubljana, Slovenia

Edited by
Vladimir Batagelj, Anuška Ferligoj and Aleš Žiberna

Program Chairs

Vladimir Batagelj (University of Ljubljana)
Anuška Ferligoj (University of Ljubljana)

Organizing Committee

Vladimir Batagelj
Andrej Blejec
Matevž Bren
Anuška Ferligoj
Nataša Kejžar
Simona Korenjak Černe
Gregor Petrič
Vincenzo Esposito Vinzi
Aleš Žiberna

Conference Page: `http://ifcs06.org/` or `http://vlado.fmf.uni-lj.si/info/ifcs06/`
Webmaster: Vladimir Batagelj

Published by: Center of Methodology and Informatics
Institute of Social Sciences at Faculty of Social Sciences
University of Ljubljana, Slovenia
Edited by: Vladimir Batagelj, Anuška Ferligoj, and Aleš Žiberna
Designed by: Bojan Senjur
Printed by: Birografika BORI, d.o.o., Ljubljana

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

001.8(063)

INTERNATIONAL Federation of Classification Societies.
Jubilee Conference (10 ; 2006 ; Ljubljana)
Data science and classification : program and abstracts /
10th Jubilee Conference of the International Federation of
Classification Societies, July 25-29, 2006, Ljubljana, Slovenia ;
[edited by Vladimir Batagelj, Anuška Ferligoj, and Aleš Žiberna]. -
Ljubljana : Center of Methodology and Informatics, Institute of
Social Sciences at Faculty of Social Sciences, 2006

ISBN 961-235-234-8
1. Batagelj, Vladimir

227611392

Contents

Preface

We are honored to welcome you to the 10th Jubilee Conference of the International Federation of Classification Societies (IFCS-2006): Data Science and Classification, organized by the Slovenian Statistical Society and Faculty of Social Sciences, University of Ljubljana. The Conference is held at the Faculty of Social Sciences (Kardeljeva ploščad 5, Ljubljana).

This book gives to the conference participants the conference program and abstracts of all 143 talks: plenary and invited lectures, contributed and invited session talks. The abstracts are given by the alphabetical order, by surname. There is also the index of authors and the index of keywords. Some useful information about the conference (e.g., conference venue, information for presenters and chairpersons, social program) are also given.

We wish to thank IFCS council members, especially Hans Hermann Bock, the former IFCS president Henk Kiers, the present IFCS president David Hand, and the IFCS secretary Vincenzo Esposito Vinzi for many hints how to organize the IFCS conference. We are grateful to the members of the local organizing committee for their help, and to several colleagues at the Faculty of Social Sciences that helped us with the organization of the conference, especially Darinka Kovačič, Izidora Kunstelj, and Sabina Otoničar.

The organizers of the 2006 IFCS conference gratefully thank the Slovenian Research Agency, the Ministry of Higher Education, Science and Technology, and the Center of Methodology and Informatics at the Faculty of Social Sciences for their support. We are grateful for the promotion material given by Public Relations and media Office and Slovenian Tourist Board. Financial support given by the sponsors Mercator and SPSS Division, Slovenia is especially appreciated.

We hope that the conference participants will have a great and fruitful time at the IFCS 2006 conference in Ljubljana and that will find some time to visit some nice places in Slovenia.

Ljubljana, July 2006

Vladimir Batagelj,
Anuška Ferligoj,
Aleš Žiberna

Organization of the IFCS 2006 Conference Data Science and Classification

10th Jubilee Conference of the International Federation of Classification Societies

The conference takes place under the auspices of the International Federation of Classification Societies (IFCS) and is organized by the Slovenian Statistical Society and the Faculty of Social Sciences, University of Ljubljana. The conference is hosted by the Faculty of Social Sciences, Ljubljana, Slovenia. The IFCS is a non-profit and non-political scientific organization which promotes the dissemination of technical and scientific information concerning data analysis, classification, related methods, and their applications.

Previous conferences were held in Aachen (Germany, 1987), Charlottesville (USA, 1989), Edinburgh (UK, 1991), Paris (France, 1993), Kobe (Japan, 1996), Rome (Italy, 1998), Namur (Belgium, 2000), Cracow (Poland, 2002), and Chicago (USA, 2004).

The member societies participating in the IFCS are the Associaçao Portuguesa de Classificaçao e Análise de Dados (CLAD), British Classification Society (BCS), Central American and Carribean Society of Classification and Data Analysis (SoCCCAD), Classification Society of North America (CSNA), Gesellschaft für Klassifikation (GfKl), Irish Pattern Recognition and Classification Society (IPRCS), Japanese Classification Society (JCS), Korean Classification Society (KCS), Société Francophone de Classification (SFC), Società Italiana di Statistica (SIS), Sekcja Klasyfikacji i Analizy Danych PTS (SKAD), Vereniging voor Ordinatie en Classificatie (VOC).

2006 Recipients of the Chikio Hayashi Travel Award

- **Alessio Farcomeny**, Italy
- **Gentian Gusho**, France
- **Daniel Kosiorowski**, Poland
- **Marielle Linting**, The Netherlands
- **Michelangelo Misuraca**, Italy
- **Georgi I. Nalbantov**, The Netherlands
- **A. Banire Diallo**, Canada
- **Mario A. Villalobos**, Carribean

Program Chairs

- **Vladimir Batagelj**, *University of Ljubljana*, Slovenia
- **Anuška Ferligoj**, *University of Ljubljana*, Slovenia

Scientific Program Committee

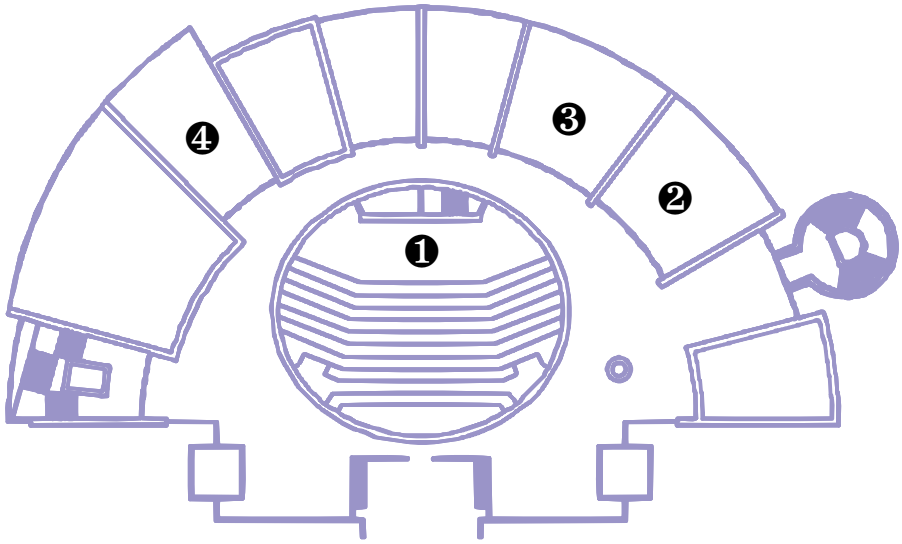
- **Phipps Arabie**, *Rutgers University*, USA
- **Helena Bacelar-Nicolau**, *University of Lisbon*, Portugal
- **David Banks**, *Duke University*, USA
- **Hans-Hermann Bock**, *RWTH Aachen University*, Germany
- **Reinhold Decker**, *University of Bielefeld*, Germany
- **Edwin Diday**, *University Paris Dauphine*, France
- **Bernard Fichet**, *University of Aix Marseille*, France
- **Alain Guenoche**, *CNRS*, France
- **Patrick Groenen**, *Erasmus University Rotterdam*, The Netherlands
- **Pierre Hansen**, *University de Montreal*, Canada
- **Krzysztof Jajuga**, *Wroclaw University of Economics*, Poland
- **Henk Kiers**, *University of Groningen*, The Netherlands
- **Wojtek Krzanowski**, *University of Exeter*, UK
- **Carlo Lauro**, *University of Naples*, Italy
- **Nada Lavrač**, *Institute Jožef Stefan*, Slovenia
- **Yonggoo Lee**, *Chung-Ang University*, Korea
- **Taerim Lee**, *Korea National Open University*, Korea
- **Buck McMorris**, *Illinois Institute of Technology*, USA
- **Jacqueline Meulman**, *University of Leiden*, The Netherlands
- **Clive Moncrieff**, *Natural History Museum*, UK
- **Fionn Murtagh**, *Queen's, University of Belfast*, Northern Ireland
- **Noboru Ohsumi**, *Institute of Statistical Mathematics*, Japan
- **Akinori Okada**, *Rikkyo University*, Japan
- **Jean-Paul Rasson**, *University of Namur*, Belgium
- **Gunter Ritter**, *University of Passau*, Germany
- **Alfredo Rizzi**, *University of Rome*, Italy
- **James F. Rohlf**, *State University of New York*, USA
- **Andrzej Sokolowski**, *Cracow University of Economics*, Poland
- **Javier Trejos**, *University of Costa Rica*, Costa Rica
- **Iven Van Mechelen**, *University of Leuven*, Belgium
- **Maurizio Vichi**, *University of Rome 'La Sapienza'*, Italy
- **Claus Weihs**, *University of Dortmund*, Germany

Local Organizing Committee

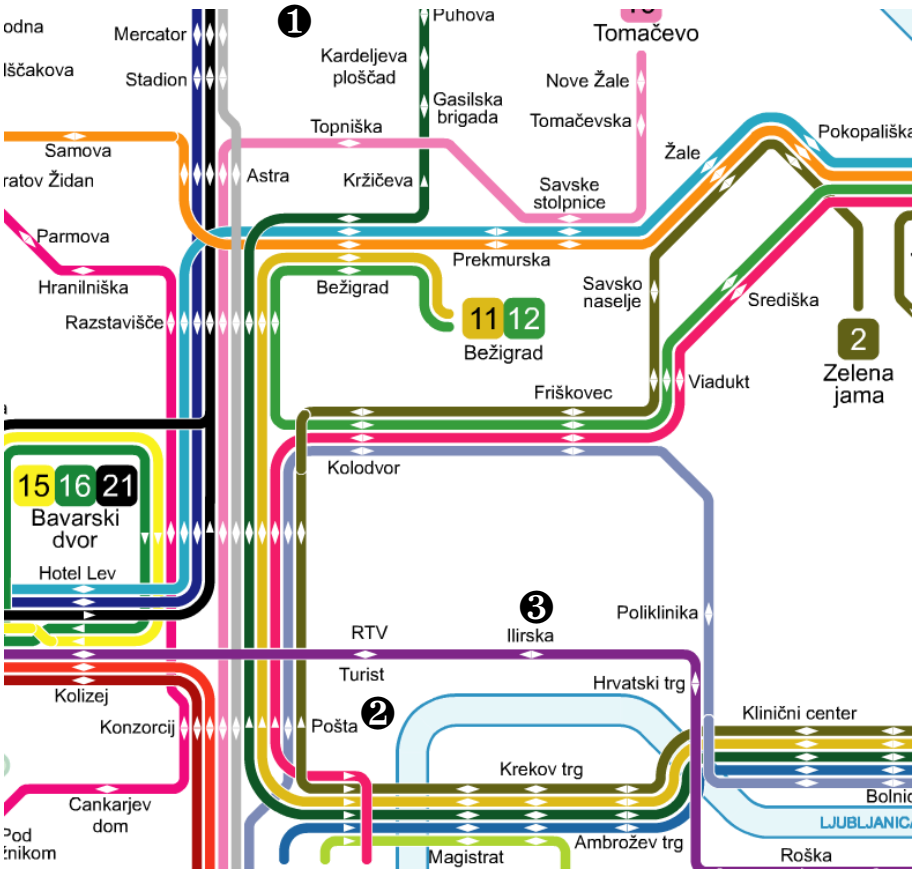
- **Vladimir Batagelj**, *University of Ljubljana*, Slovenia
- **Andrej Blejec**, *University of Ljubljana*, Slovenia
- **Matevž Bren**, *University of Maribor*, Slovenia
- **Anuška Ferligoj**, *University of Ljubljana*, Slovenia
- **Nataša Kejžar**, *University of Ljubljana*, Slovenia
- **Simona Korenjak Černe**, *University of Ljubljana*, Slovenia
- **Gregor Petrič**, *University of Ljubljana*, Slovenia
- **Vincenzo Esposito Vinzi**, *University of Naples*, Italy
- **Aleš Žiberna**, *University of Ljubljana*, Slovenia

Conference Location

The Conference will be held at the Faculty of Social Sciences, University of Ljubljana (Kardeljeva ploščad 5, Ljubljana). It is one of the University's largest faculties, located in a spacious campus relatively close to the town center (10 minutes by bus number 6, 8, or 21; 25 minutes on foot). The premises of the faculty offer adequate facilities and conditions for plenary as well as for working group sessions. See the enclosed maps.



City buses to Faculty of Social Sciences



❶ Faculty of Social Sciences (Conference Venue)
❷ Grand Hotel Union ❸ Hotel Park

The suggested city bus connections are (see also the map on page 12):

From Hotel Union: line number 6 from Pošta to Mercator

From Hotel Park: line number 5 from Ilirska to Hotel Lev, cross the street to Kolizej, line number 8 or 21 to Mercator.

All registered IFCS 2006 Conference participants and registered accompanying persons can travel **free of charge** on Ljubljana city buses during the days of the conference upon presentation of their Conference badge.

❶ Faculty of Social Sciences (Conference Venue)

❷ Grand Hotel Union ❸ Hotel Park

Information for Presenters

Audio/Visual Information. It is planned to have each room used for conference talks equiped with a PC for PowerPoint and PDF presentations as well as an overhead projector. The PC will be connected to the Internet. The map of conference rooms is enclosed.

The alloted times for talks are:

- 50 minutes for ***plenary*** and ***special invited*** talks, with 5-10 minutes for discussion;
- 15 minutes for ***invited session*** and ***contributed*** talks, with 2-3 minutes for discussion.

Information for Session Chairs

Please, contact the registration desk at least 30 minutes before the session.

Conference Proceedings

Conference Proceedings are published by Springer

> Data Science and Classification. Series: Studies in Classification, Data Analysis, and Knowledge Organization Batagelj, V.; Bock, H.-H.; Ferligoj, A.; Žiberna, A. (Eds.) 2006, XII, 358 p., 67 illus., Softcover. ISBN: 3-540-34415-2

`http://www.springer.com/dal/home?SGWID=1-102-22-172424298-0`
`http://www.amazon.com/gp/product/3540344152`

Conference Address

Local Organizing Committee IFCS-2006
Anuška Ferligoj
Faculty of Social Sciences
Kardeljeva ploščad 5
1000 Ljubljana, SLOVENIA

phone: +386 1 5805 281
fax: +386 1 5805 102
e-mail: `ifcs06@fdv.uni-lj.si`
web page: `http://www.ifcs06.org`

Part I

Program

Tuesday July 25

8:00-19:00	**Registration**	
	Tutorials	
Tuesday	**Hall 1**	
July 25	**Tutorial on Symbolic Data Analysis**	
9:00-13:00	**Lectures**	
13:00-14:00	**Lunch**	
14:00-18:00	**Lectures**	
Tuesday	**Hall 2**	
July 25	**Tutorial on Blockmodeling**	
10:00-13:00	**Lectures**	
13:00-14:30	**Lunch**	
14:30-17:30	**Lectures**	
Tuesday		**Senat Room**
July 25		
18:00-	**Welcome Reception**	**Executive Committee Meeting**

Wednesday July 26

8:00-9:00	**Registration**		
Wednesday	**Hall 1**		
July 26			
9:00-9:20	**Opening Ceremony**		
	Plenary session		
Wednesday	**Hall 1 / chair: Fionn Murtagh**		
July 26	**Plenary Speaker**		
9:20-10:20	*Adrian E. Raftery*: Recent Advances in Model-Based Clustering: Variable Selection and Social Networks		
10:20-10:40	**Coffee Break**		
	Invited Sessions		
Wednesday	**Hall 1 / chair: C. Hennig**	**Hall 2 / chair: I. van Mechelen**	**Hall 3 / chair: K. Jajuga**
July 26	**Design and choice of similarity and dissimilarity measures (org. C. Hennig)**	**Clustering and Classification of Microarray Gene Expression Data (1) (org. I. van Mechelen)**	**Classification and Data Analysis in Economics (org. K. Jajuga)**
10:40-11:00	*Daniel Müllensiefen, Klaus Frieler* Evaluating different approaches to measuring the similarity of melodies	*Thomas Dhollander, Qizheng Sheng, Yves Moreau* Biclustering of microarray data in a Bayesian framework	*Eugeniusz Gatnar* Combining classifiers of different types
11:00-11:20	*Remco C. Veltkamp, Longin Jan Latecki* Properties and Performances of Shape Similarity Measures	*Sara C. Madeira, Arlindo L. Oliveira* Discovering Modules in Time-Series Gene Expression Data Using Biclustering	*Sören W. Scholz, Ralf Wagner, Reinhold Decker* Classification in Marketing Science
11:20-11:40	*John C. Gower* Similarity in Retrospect	*Ulrich Moeller* Resampling-based class discovery from microarray data: Guidelines from benchmarking studies	*Krzysztof Jajuga* Copula Methods in Financial Data Analysis
11:40-12:00	*Theresa Scharl, Friedrich Leisch* A Comparison of Distance Measures for Clustering Time–Course Microarray Data	*Marco Alfo, Francesca Martella, Maurizio Vichi* Double Hierarchical Mixture model for microarray data	*Daniel Baier, André Bruder* Model Selection in Latent Class Metric Conjoint Analysis: Sequential and Simultaneous Approaches

Wednesday		Senat Room
July 26		
12:00-13:30	**Lunch**	**Council Meeting**

Contributed Sessions

Wednesday	Hall 1 / chair: M. Vichi	Hall 2 / chair: S. Winsberg	Hall 3 / chair: V. Batagelj	Hall 4 / chair: A. Guénoche
July 26	**Cluster Analysis**	**Multidimensional Scaling**	**Network Analysis**	**Optimization**
13:30-13:50	*Slavka Bodjanova* Crisp Partitions Induced by a Fuzzy Set	*François Bavaud* Spectral clustering and multidimensional scaling: a unified view	*Vladimir Batagelj, Nataša Kejžar, Simona Korenjak-Černe, Matjaž Zaveršnik* Analyzing the Structure of U.S. Patents Network	*Mario Villalobos-Arias, Javier Trejos-Zelaya* Partitioning by particle swarm optimization
13:50-14:10	*Karina Gibert, Alejandra Pérez-Bonilla* Revised boxplot based discretization as the kernel of automatic interpretation of classes using numerical variables	*Sugnet Gardner, Niël J. le Roux* Sub-species of Homopus Areolatus? Biplots and small class inference with Analysis of Distance	*Hiroyuki Minami, Masahiro Mizuta* Empirical Study on the ICMP traffic data via the Internet	*Farid Beninel, Michel Grun Rehomme* Evaluation of allocation rules under some cost constraints
14:10-14:30	*Hans-Joachim Mucha* Finding Meaningful and Stable Clusters Using Local Cluster Analysis	*Patrick J.F. Groenen, Georgi Nalbantov, Cor Bioch* Multigroup SVM through an Optimal Scaling Approach	*Barbara Japelj Pavešič* Comparison of teaching mathematics: An application of Adapted Leaders Clustering Method	*Ralf Wagner* Patterns of Associations in Finite Sets of Items
14:30-14:50	*Bart Jan van Os, Jacqueline J. Meulman* Approximation of Globally Optimized Trees	*Ana Alexandra A.F. Martins, Margarida G.M.S. Cardoso* A Preliminary Approach to the Evaluation of MDS Unfolding Fit Measures	*Petra Ziherl, Hajdeja Iglič, Anuška Ferligoj* Research group social capital: A Clustering Approach	*Philippe Nemery, Yves De Smet* Multicriteria Ordered Clustering
14:50-15:10	**Coffee Break**			

Invited Speakers

Wednesday	Hall 1 / chair: J.-P. Rasson	Hall 2 / chair: D. Banks
July 26		
15:10-16:10	***Carlo Lauro***, *Federica Gioia* Dependence and interdependence analysis for interval-valued variables	***Jacqueline J. Meulman***, *Jerome H. Friedman* New Developments in COSA: Clustering Objects on Subsets of Attributes
16:10-16:30	**Coffee Break**	

	Invited Sessions		
Wednesday	**Hall 1 / chair: F. de Carvalho**	**Hall 2 / chair: I. van Mechelen**	**Hall 3 / chair: C. Weihs**
July 26	**Exploratory and confirmatory analysis of interval data (org. P. Brito, F. de Carvalho, R. Verde)**	**Clustering and Classification of Microarray Gene Expression Data (2) (org. I. van Mechelen)**	**Musicology (org. C. Weihs)**
16:30-16:50	*Lynne Billard* Dependence in Interval-valued Observations	*Berthold Lausen* Machine learning for microarray prediction in clinical research	*Katrin Sommer, Claus Weihs* Using MCMC as a stochastic optimization procedure for musical time series
16:50-17:10	*Jean-Paul Rasson, François Roland* Some New Criteria for Hierarchical Agglomerative Clustering and for Discriminant Analysis. Applications to Interval Data	*Eytan Domany* Mining High-Throughput Biological Data: Methods, Algorithms and Applications	*Claus Weihs, Gero Szepannek, Uwe Ligges, Karsten Luebke, Nils Raabe* Local Models in Register Classification by Timbre
17:10-17:30	*Yves Lechevallier, Rosanna Verde, Francisco de A.T. de Carvalho* Symbolic Clustering of Large Datasets	*Iven van Mechelen* Biclustering Methods for Microarray Gene Expression Data: Towards a Unifying Taxonomy	
18:30-	**Reception**		
20:00-	**Council Dinner**		

Thursday July 27

	President's Invited Session
Thursday	**Hall 1 / chair: David J. Hand**
July 27	**Organizer: David J. Hand**
9:00-10:30	*Fionn Murtagh*: Ultrametrics in data analysis, quantum physics and computational logic *Henk A.L. Kiers*: Comparing what we have rather than developing something new? *Friedrich Leisch*: Benchmarking cluster algorithms
10:30-10:50	**Coffee Break**

	Invited Sessions		
Thursday	**Hall 1 / chair: R. Rocci**	**Hall 2 / chair: B. Fichet**	**Hall 3 / chair: C. Lauro**
July 27	**Classification of complex data (org. R. Rocci)**	**Dissimilarity analysis and clustering (org. B. Fichet)**	**Latent Class Analysis and Classification Issues (org. C. Lauro, V.E. Vinzi)**
10:50-11:10	*Mohamed Nadif, Gerard Govaert* A review on Block clustering under the mixture approach	*L. Denoud, A. Guénoche* Comparison of distance indices between partitions	*José G. Dias* Model Selection for the Binary Latent Class Model. A Monte Carlo Simulation.
11:10-11:30	*Giuliano Galimberti, Gabriele Soffritti* Clustering Data with Multiple Cluster Structures: a Model-Based Perspective	*Patrice Bertrand* Paired-hierarchies and associated dissimilarities	*Alessio Farcomeni, Maurizio Vichi* Model based two-mode partitioning with dependence between row and column clusters
11:30-11:50	*Christophe Biernacki* Simultaneous Model-Based Clustering of Data Arising from Different Populations	*Melvin F. Janowitz* Dissimilarities Taking Values in a Poset	*Vincenzo Esposito Vinzi, Laura Trinchera* Capturing Unobserved Heterogeneity in PLS Path Modeling
11:50-12:10	*Hans-Hermann Bock* Class prototypes for complex (interval) data	*Jean Pierre Barthélémy, Marc Le Pouliquen, Patrice Bertrand* Phylogenetic Trees, Additive Tree Metrics and Filiation of Manuscripts: The Case of the Benares Glose.	*Attilio Gardini, Michele Costa, Stefano Iezzi* Latent Class Analysis for Financial Data
12:10-12:30	**Coffee Break**		

	Contributed Sessions			
Thursday	**Hall 1 / chair: B. Lausen**	**Hall 2 / chair: F.R. McMorris**	**Hall 3 / chair: J. Gower**	**Hall 4 / chair: M. Bren**
July 27	**Data Analysis in Medicine**	**Graphs**	**Similarities for mixed variables**	**Poster Session**
12:30-12:50	*Andrei Gagarin, Dmytro Kevorkov, Vladimir Makarenkov, Pablo Zentilli* Comparison of two methods for detecting and correcting systematic error in high-throughput screening	*Willem J. Heiser, Laurence E. Frank* Network Representations of City-Block Models	*Lan Umek, Luka Bresciani, Luka Kronegger* An overview of mixed data analysis based on Gower's coefficient of similarity	*Iwona Roeske-Slomka* Objects grouping based on entrophy
12:50-13:10	*Shinobu Tatsunami, Rie Kuwabara, Masashi Taki, Junichi Mimaya, Akira Shirahata* Application of Categorical Principal Component Analysis to the Classification of Antiretroviral Drug Usage	*Gentian Gusho* Properties of Minimum Rigidity Graphs Associated with a Clustering System	*Petra Bastjančič, Nejc Bergant, Renata Blatnik, Miran Juretič, Mojca Omerzu, Jure Željko* Comparison of clustering approaches for mixed data: a simulation study	*Deirdre Toher, Gerard Downey, Thomas Brendan Murphy* A One-sided View of Classification: A food science perspective
13:10-13:30	*Vesna Zadnik, Tina Žagar, Maja Primic Žakelj* The Geographical Variation of Cervical Cancer Screening Efficiency in Slovenia	*Zenel Batagelj, Vladimir Batagelj, Bojan Korenini, Vanja Govednik* Revealing hidden relevance in complex laddering networks	*Andreja Bandelj, Mateja Budin, Tomi Deutch, Andrej Kastrin, Ana Kolar, Polona Kramar* Comparison of clustering approaches for mixed variables	*Masaki Nishizawa, Yuan Sun* Investigation into Genome-related and Nanotechnology Research at Grants-in-Aid in JAPAN
				Yuan Sun, Masaki Nishizawa The classification of journals in Citation Database of Japanese Papers (CJP) by keyword analysis
14:00-	**Trip to Bled**			

Friday July 28

	Invited Speakers	
Friday	**Hall 1 / chair: A. Ferligoj**	**Hall 2 / chair: M. Janowitz**
July 28		
09:00-10:00	***Patrick Doreian*** Some Open Problem Sets for Generalized Blockmodeling	*James E. Wicker,* ***Hamparsum Bozdogan****, Halima Bensmail* A Novel Mixture-Model Cluster Analysis With Genetic EM Algorithm and Information Complexity as the Fitness Function
10:00-10:20	**Coffee Break**	

	Invited Sessions		
Friday	**Hall 1 / chair: P. Doreian**	**Hall 2 / chair: A. Okada**	**Hall 3 / chair: Y. Lechevallier**
July 28	**Blockmodeling (org. P. Doreian)**	**Multidimensional scaling of assymmetric relationships data (1) (org. A. Okada)**	**Symbolic data analysis (org. by the IFCS Transversal Group on Symbolic Data Analysis: R. Verde, Y. Lechevallier)**
10:20-10:40	*Matteo Roffilli, Alessandro Lomi* Identifying and classifying social groups: A machine learning approach	*Tadashi Imaizumi* A Method for Local Representation the Asymmetric Similarity Data Matrix	*Renata M.C.R. de Souza, Francisco de A.T. de Carvalho, Daniel Ferrari Pizzato* A Dynamic Clustering Method for Mixed Feature-Type Symbolic Data
10:40-11:00	*Aleš Žiberna* Evaluation of Generalized blockmodeling and REGE on Regular equivalence	*Miki Nakai* Career Mobility over the Life Course among Women in Japan	*P.J.F. Groenen, S. Winsberg* Multidimensional Scaling of Histogram Dissimilarities
11:00-11:20	*Vladimir Batagelj, Patrick Doreian, Anuška Ferligoj* Three Dimensional Blockmodeling	*Akinori Okada* An Asymmetric Cluster Analysis Study of Relationships among Japanese Prefectures in Marriage	*Antonio Irpino, Rosanna Verde* A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data
11:20-11:40	**Coffee Break**		

	Contributed Sessions			
Friday	**Hall 1 / chair: T. Imaizumi**	**Hall 2 / chair: B. Mirkin**	**Hall 3 / chair: H. Bozdogan**	**Hall 4 / chair: G. Petrič**
July 28	**(Dis)similarities**	**Nonhierarchical Clustering**	**Statistics (1)**	**Text Analysis**
11:40-12:00	*Emmanuel Blanchard, Pascale Kuntz, Mounira Harzallah, Henri Briand* A tree-based similarity for evaluating concept proximities in an ontology	*Yoshiharu Sato* Clustering for Mixed Data Using Sphereical Representation	*Hans J. Vos, Ruth Ben-Yashar, Shmuel Nitzan* Comparing optimal individual and collective assessment procedures	*Michelangelo Misuraca* How factorial techniques can support translation processes in the Web era
12:00-12:20	*Fabrice Rossi, Francisco De Carvalho, Yves Lechevallier, Alzennyr Da Silva* Dissimilarities for Web Usage Mining	*Bruno Leclerc* A consensus approach based on frequent groupings	*François-Joseph Lapointe* A Statistical Framework for Assessing the Congruence Among Overlapping Trees and Detecting Common Spatial Patterns in Comparative Phylogeography	*Joseph Rudman* Assumptions Behind the Statistics In Authorship Attribution Studies: A Search for Valid Tests
12:20-12:40	*Véronique Campbell, Pierre Legendre, François-Joseph Lapointe* Assessing congruence among ultrametric distance matrices	*David Banks, Natesh Pillai* Partition-Valued Random Variables	*Nathalie Cheze, Jean-Michel Poggi* Iterated Boosting for Outlier Detection	*David Dubin* Reframing Author Cocitation Analysis
12:40-13:00	*A.M. Shurygin* Interpoint distances	*Masahiro Mizuta* Evaluation of the Results of Functional Clustering	*Matevž Bren* The data sets from Aitchison's book in the 'compositions' package	*Primož Jakopin* Classification of Slovenian words from a search engine index
13:00-14:30	**Lunch**			
	Invited Speakers			
Friday	**Hall 1 / chair: V. Batagelj**		**Hall 2 / chair: W. Heiser**	
July 28				
14:30-15:30	***John Shawe-Taylor*** Generalisation Analysis as a Foundation for Classification		***Christian Hennig***, *Bernhard Hausdorf* Design of Dissimilarity Measures: a New Dissimilarity between Species Distribution Areas	
15:30-15:50	**Coffee Break**			

	Contributed Sessions			
Friday	**Hall 1 / chair: Y. Sato**	**Hall 2 / chair: P. Groenen**	**Hall 3 / chair: T. Lee**	**Hall 4 / chair: F.-J. Lapointe**
July 28	**Clustering**	**Principal Component Analysis**	**Statistics (2)**	**Data Analysis in Biology**
15:50-16:10	*Helena Brás Silva, Paula Brito, Joaquim Pinto da Costa* A New Hierarchical Clustering Method based on Graph Theory	*Mariëlle Linting, Jacqueline Meulman, Bart Jan van Os* Nonparametric Inference in Principal Components Analysis by Using Permutation Tests: Two Approaches Compared	*A.M. Shurygin* Theory of stable estimation	*Abdoulaye Baniré Diallo, Vladimir Makarenkov, Mathieu Blanchette, François-Joseph Lapointe* A new efficient method for assessing missing nucleotides in DNA sequences in the framework of a generic evolutionary model
16:10-16:30	*Javier Arroyo, Carlos Maté, Antonio Munoz-San Roque* Hierarchical Clustering for Boxplot Variables	*Rafik Abdesselam* Mixed Principal Component Analysis	*José António Santos, M. Manuela Neves* A Local Maximum Likelihood Estimator for Zero-inflated Count Data	*Vladimir Makarenkov, Alix Boc, Charles F. Delwiche, Alpha Boubacar Diallo, Hervé Philippe* New efficient algorithm for modeling partial and complete gene transfer scenarios
16:30-16:50	*Iven van Mechelen* Simultaneous Clustering Methods Using the Max Operator: The Hierarchical Classes Approach	*Roberto Rocci, Maurizio Vichi* Simultaneous models for clustering and reduction	*Ana Oliveira-Brochado, Francisco Vitorino Martins* Determining the number of components in mixture regressions of normal data	*Edgar Acuña, Jaime Porras* Improving the performance of principal components for classification of gene expression data through feature selection
16:50-17:10	*Marie Chavent, Yves Lechevallier* Empirical comparison of a divisive clustering method with the Ward and the k-means methods	*Bernard Fichet, Jean Gaudart, Bernard Giusiano* Solving oblique bivariate CART	*Galina Andreeva, Jake Ansell, Jonathan Crook* Survival combination scores: a case of a revolving store card	*Taerim Lee* Tree Structured prognostic Model for HCC Using the Gene Expression Data
17:10-17:30	*Bernard Harris* Some applications of combinatorial methods used for cluster verification		*Paolla Berchialla, Silvia Snidero, Alex Stancu, Cecilia Scarinzi, Roberto Corradetti, Dario Gregori* Bayesian models for safety design to prevent foreign body injuries in children	
19:00-	**Conference Dinner**			

Saturday July 29

	Presidential Address	
Saturday	**Hall 1 / chair: Henk Kiers**	
July 29		
9:00-10:00	*David J. Hand*: The length and breadth of classification science: the case study of banking fraud	
10:00-10:20	**Coffee Break**	
	Invited Speakers	
Saturday	**Hall 1 / chair: H.-H. Bock**	**Hall 2 / chair: E. Diday**
July 29		
10:20-11:20	***Boris Mirkin*** Clustering in Different Perspectives: How many clusters are there?	***Paula Brito*** On the Analysis of Symbolic Data
11:20-11:40	**Coffee Break**	

	Contributed Sessions			
Saturday	**Hall 1 / chair: K. Košmelj**	**Hall 2 / chair: S. Korenjak-Černe**	**Hall 3 / chair: J. Meulman**	**Hall 4 / chair: A. Sokolowski**
July 29	**Clustering Temporal Data**	**Two-mode Clustering**	**Multivariate Analysis**	**Application in Social Sciences**
11:40-12:00	*A. Chouakria-Douzal, P. Nagabhushan* Improved Fréchet Distance For Time Series	*Dirk Depril, Iven van Mechelen* One-mode additive clustering for two-way two-mode data: a comparison of algorithms	*Daniel Kosiorowski* About Strein Force in a Capital and Robust Analysis of Planar Shape	*W. Polasek, R. Sellner* How important are ICT services for economic growth?
12:00-12:20	*Bernard Hugueney, Georges Hébrail, Yves Lechevallier* Computing summaries of time series databases with clustering and segmentation	*Tom Wilderjans, Eva Ceulemans, Iven van Mechelen* HIC-model: A global model for coupled binary data	*Georgi I. Nalbantov, Jan C. Bioch, Patrick J.F. Groenen* Binary Classification with Support Hyperplanes	*Felix Rüb, Daniel Werner, Katja Wolf* Classification of regional labour markets for purposes of research and of labour market policy
12:20-12:40	*Vesna Žabkar, Katarina Košmelj* Clustering Time Varying Data of European Advertising Expenditures	*Jan Schepers, Iven van Mechelen* Revealing the nature of interactions in two-way two-mode data by constrained simultaneous clustering models	*Ilse Stuive, Henk A.L. Kiers, Marieke E. Timmerman, Jos M.F. ten Berge* Adjusting incorrect classifications of items in subtests: Oblique Multiple Group method or Confirmatory Common Factor method	*Carlos M.F. Monteiro, Joao Oliveira Soares, Cristina del Campo* Regional Clusters and Socio-economic diversity in the European Union
12:40-13:00	*Gaj Vidmar, Krešimir Matković, Branimir Leskošek, Drago Rudel* Exploratory Analysis of Uterine Electromyographic Data from Pregnant Sheep	*Hans-Friedrich Köhn* Anti-Robinson Structures for Analyzing Three-way Two-mode Data		*Guy Cucumel, Véronique Mallandain, Marie-Marthe Cousineau* Looking for a Typology of Sexual Practices by Adolescents in Three Secondary Schools in Quebec
13:00-14:30	**Lunch**			

	Invited Sessions		
Saturday	**Hall 1 / chair: A. Okada**	**Hall 2 / chair: A. Ferligoj**	**Hall 3 / chair: L. Schmidt-Thieme**
July 29	**Multidimensional scaling of assymmetric relationships data (2) (org. A. Okada)**	**Spatial Classification (org. E. Diday, V. Batagelj)**	**Web mining (org. L. Schmidt-Thieme)**
14:30-14:50	*Kohei Adachi* Nonlinear Principal Component Analysis for Representing Inter-variable Relationships by Trajectories	*Edwin Diday* Spatial Classification	*Christoph Schmitz, Andreas Hotho, Robert Jäschke, Gerd Stumme* Mining Association Rules in Folksonomies
14:50-15:10	*Naohito Chino, Shingo Saburi* A link between the asymmetric MDS and the analysis of contingency table	*Mohamed Cherif Rahal, Edwin Diday* Spatial hierarchical and pyramidal clustering software	*Miha Grčar, Blaž Fortuna, Dunja Mladenič, Marko Grobelnik* kNN Versus SVM in the Collaborative Filtering Framework
15:10-15:30	*Akinobu Takeuchi, Hiroshi Yadohisa* Simulation study of asymmetric k-medoids clustering algorithms for dissimilarity data	*Vladimir Batagelj* Combining data and network analysis	*Karen H. L. Tso, Lars Schmidt-Thieme* Empirical Analysis of Attribute-Aware Recommendation Algorithms with Variable Synthetic Data
15:30-15:50	*Giuseppe Bove* Asymmetric Multidimensional Scaling with External Information: Some Possible Approaches		*Andreas Geyer-Schulz, Bettina Hoser* On the Eigensystems of Operational Accounting Systems
15:50-16:10	**Coffee Break**		

Saturday	**Hall 1**
July 29	
16:10-	**Closing Ceremony**

Part II

Abstracts

Mixed Principal Component Analysis

Rafik Abdesselam

CREM UMR CNRS 6154 - University of Caen, Basse-Normandie
Esplanade de la paix, F-14032 Caen, France
rafik.abdesselam@unicaen.fr

Abstract. The processing of mixed data - quantitative and qualitative variables cannot be carry out directly by classical methods of data analysis. In this work, a factorial method which analyze simultaneously quantitative and qualitative data is described. The proposed Mixed Principal Component Analysis is a standardized principal component analysis of both quantitative variables and the transformation of the dummy variables associated to qualitative variables on quantitative variables through orthogonal projections of configurations of statistical units in the individual-space with a relational inner product. An example resulting from real data illustrates the results of this method, which are also compared with those of the Mixed Data Factorial Analysis proposed by Pagès (2004).

References

PAGÈS, J. (2004): Analyse factorielle de données mixtes. *Revue de Statistique Appliquée, LII(4), 93-111*.

SAPORTA, G. (1990): Simultaneous analysis of qualitative and quantitative data. Società italiana di statistica, *Atti della XXXV riunione scientifica*, 63–72.

SCHEKTMAN, Y. and ABDESSELAM, R. (2000): A Geometrical Relational Model for Data Analyses. In: W.Gaul, O.Opitz and M.Schader (Eds.): *Data Analysis*. Springer, Berlin, 359–368.

TENENHAUS, M. (1977): Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée, XXV(2), 39-56*.

YOUNG, F.W. (1981): Quantitative analysis of qualitative data. *Psychometrika, 46(4), 357-388*.

Keywords

PRINCIPAL COMPONENT ANALYSIS, MULTIVARIATE ANALYSIS OF VARIANCE, GENERALIZED CORRELATION RATIO, RELATIONAL INNER PRODUCT

Improving the Performance of Principal Components for Classification of Gene Expression Data Through Feature Selection

Edgar Acuña and Jaime Porras

Department of Mathematics, University of Puerto Rico at Mayaguez,
Mayaguez, PR 00680, USA

Abstract. The gene expression data is characterized by its considerably great amount of features in comparison to the number of observations. The direct use of traditional statistics techniques of supervised classification can give poor results in gene expression data. Therefore, dimension reduction is recommendable prior to the application of a classifier. In this work, we propose a method that combines two types of dimension reduction techniques: feature selection and feature extraction. First, one of the following feature selection procedures: a univariate ranking based on the Kruskal-Wallis statistic test, the Relief, and recursive feature elimination (RFE) is applied on the dataset. After that, principal components are formed with the selected features. Experiments carried out on eight gene expression datasets using three classifiers: logistic regression, k-nn and rpart, gave good results for the proposed method.

Keywords

DIMENSION REDUCTION, PRINCIPAL COMPONENT ANALYSIS, SUPERVISED CLASSIFICATION, FEATURE SELECTION, CLASSIFICATION OF GENE EXPRESSION DATA

Nonlinear Principal Component Analysis for Representing Inter-Variable Relationships by Trajectories

Kohei Adachi

Graduate School of Human Sciences
Osaka University
Japan

Abstract. To capture nonlinear relationships between variables, we consider representing variables by nonlinear trajectories in a low-dimensional configuration. Though it cannot be attained by ordinary linear PCA (principal component analysis), MCA (multiple correspondence analysis) which is regarded as a generalized PCA method gives nonlinear variable -trajectories. That is, treating the values on variables as nominal categories and giving optimal scores to the categories by MCA, we have the trajectories which connect the points of category scores. However, the MCA trajectories are often too zigzag, when variables have many categories. To deal with these difficulties, we propose a method which gives smooth and comprehensible nonlinear variable-trajectories. In this method, an objective function to be minimized is defined by combining the loss function for MCA and a penalty function. Here, the penalty function expresses the loss of the smoothness of variable-trajectories which are defined as natural cubic splines, and the weight of the penalty is chosen by cross-validation. In the presentation, this method is detailed, an example is given, and the relationships to PCA, MCA, and categorical PCA are discussed.

Keywords

PRINCIPAL COMPONENT ANALYSIS, MULTIPLE CORRESPONDENCE ANALYSIS, NONLINEAR VARIABLE-TRAJECTORIES

Double Hierarchical Mixture Model for Microarray Data

Marco Alfó, Francesca Martella, and Maurizio Vichi

Dipartimento di Statistica, Probabilità e Statistiche Applicate,
Universitá degli Studi di Roma La Sapienza

Abstract. A major empirical aim in microarray data analysis is that it could be to cluster both genes and tissue samples. In the last few years, mixture based clustering techniques have become widely used in microarray data analysis both in gene and tissue sample based clustering context.

We introduce an approach based on a hierarchical structure, allowing for both dependence within clusters and simultaneous clustering of genes and tissue samples.

The proposed approach is obtained by extending the multilevel latent class model proposed by Vermunt (2003) and Li (2005) to two-way continuous data. In order to cluster tissues, we introduce a binary row stochastic matrix of tissue cluster membership (as in Vichi, 2000) using a new reparameterization of the mean vectors of 1st level sub-clusters by extending the proposal of Rocci and Vichi (2002).

To discuss the performance of the proposed double hierarchical mixture model, we present the analysis of a benchmark data set.

References

LI, J. (2005): Clustering based on a multi-layer mixture model, *Journal of Computational and Graphical Statistics, 14(3), 547–568.*

ROCCI, R. and VICHI M. (2002): A two-way model for simultaneous reduction and classification, *Atti della XLI Riunione Scientifica*, Universitá di Milano-Bicocca (5-7 june).

VERMUNT, J.K. (2003): Multilevel latent class models, *Sociological Methodology, 33, 213–239.*

VICHI, M. (2000): Double k-means Clustering for simultaneous classification of Objects and Variables, In Borra et al. (eds): *Advances in Classification and Data Analysis, 43–52.*

Keywords

HIERARCHICAL MIXTURE MODEL, DOUBLE CLUSTERING, MICROARRAY DATA

Survival Combination Scores: A Case of a Revolving Store Card

Galina Andreeva, Jake Ansell, and Jonathan Crook

Credit Research Centre, The University of Edinburgh Management School
50 George Sq., Edinburgh EH9 1RS, United Kingdom

Abstract. To achieve classification into defaulting and non-defaulting credit accounts, credit scoring employs a range of techniques. Traditionally logistic regression is used to predict probability of default. Survival analysis offers an advantage of incorporating time-varying elements into analysis and has successfully been applied to model default and profit for fixed-term credit products. Profit estimation for a revolving credit is complicated by necessity to assess the product usage.

The presentation addresses the application of survival analysis to the area of revolving credit, namely, a store card used for buying 'white' durable goods. An empirical investigation shows there is a relationship between present value of net revenue from a store card account and times to default and to second purchase (a measure of usage). It appears that there is a scope for improving profitability if an application for a store card is assessed by using a model which estimates the revenue and includes the survival probability of default and the survival probability of second purchase (a survival combination model) rather than a static default probability predicted by logistic regression.

References

ANDREEVA, G., ANSELL, J.I. and CROOK, J.N. (2005): Modelling the purchase propensity: analysis of a revolving store card. *Journal of Operational Research Society, 56, 1041–1050.*

Keywords

CREDIT SCORING, BANKING, SURVIVAL ANALYSIS, LOGISTIC REGRESSION

Hierarchical Clustering for Boxplot Variables

Javier Arroyo[1], Carlos Maté[2], and Antonio Muñoz-San Roque[2]

[1] Departamento de Sistemas Informáticos, Universidad Complutense de Madrid, Profesor García-Santesmases s/n, 28040 Madrid, Spain
[2] Instituto de Investigación Tecnológica, ETSI, Universidad Pontificia Comillas, Alberto Aguilera 25, 28015 Madrid, Spain

Abstract. Boxplots are well-known exploratory charts used to extract meaningful information from batches of data at a glance. Their strength lies in their ability to summarize data retaining the key information, which also is a desirable property of symbolic variables. In this paper, boxplots are presented as a new kind of symbolic variable. In addition, two different approaches to measure distances between boxplot variables are proposed. The usefulness of these distances is illustrated by means of a hierarchical clustering of boxplot data.

References

BENJAMINI, Y. (1988): Opening the Box of a Boxplot. *American Statistician, 42/4, 257-262*.

BILLARD, L., and DIDAY, E. (2002): From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association, 98/462, 991-999*.

BOCK, H.H. and DIDAY, E. (2000): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information >From Complex Data.* Springer-Verlag, Heidelberg.

FRIGGE, M., HOAGLIN, D. C., and IGLEWICZ, B. (1989): Some Implementations of the Boxplot. *American Statistician, 43/1, 50-54*.

HOAGLIN, D. C., IGLEWICZ, B., and TUKEY, J. W. (1986): Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association, 81/396, 991-999*.

ICHINO, M., and YAGUCHI, H. (1994): Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Transactions on Systems, Man and Cybernetics, 24/1, 698-708*.

NIBLACK, W., BARBER, R., EQUITZ, W., FLICKNER, M.D., GLASMAN, E.H., PETKOVIC, D., YANKER, P., FALOUTSOS, C., TAUBIN, G., and HEIGHTS, Y. (1993): Querying images by content, using color, texture, and shape. *SPIE Conference on Storage and Retrieval for Image and Video Databases, 1908, 173-187*.

TRENKLER, D. (2002): Quantile-Boxplots. *Communications in Statistics: Simulation and Computation, 31/1, 1-12*.

TUKEY, J. W. (1977).*Exploratory Data Analysis.* Addison-Wesley, Reading.

Keywords

BOXPLOT, SYMBOLIC VARIABLE, METRIC, CLUSTERING

Model Selection in Latent Class Metric Conjoint Analysis: Sequential and Simultaneous Approaches

Daniel Baier and André Buder

Chair of Marketing and Innovation Management, Brandenburg University of Technology, Konrad-Wachsmann-Allee 1, 03046 Cottbus, Germany

Abstract. For model selection in Bayesian latent class procedures various approaches exist: Sequential approaches generate MCMC outputs for different class numbers. Then, subsequently, bridge (Meng, Wong (1996)) or reciprocal importance sampling (Gelfand, Dey (1994)) can be used for model likelihood estimation with respect to the different class numbers and for selecting the adequate model.

Alternatively, simultaneous approaches generate MCMC outputs without a prespecified class number. Reversible jump (Green (1995), Richardson, Green (1997)) as well as birth-and-death (Stephen (2000)) methods are such approaches, where the class number is introduced as an additional model parameter. This paper discusses advantages of these approaches by adapting them to model selection in Hierarchical Bayes Latent Class Conjoint Analysis (HB/LCMCA, see, e.g. Baier, Polasek (2003), Tüchler et al. (2004)). Synthetic and real data are used for comparisons.

References

BAIER, D. and POLASEK, W. (2003): Market Simulation Using Bayesian Procedures in Conjoint Analysis. *Studies in Classification, Data Analysis, and Knowledge Organization, 22, 413–421.*

GELFAND, A.E., and DEY, D.K. (1994): Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society*, Ser. B, 501–514.

GREEN, P.J. (1995): Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82, 711Ű-732.

MENG, X.L. and WONG, W.H. (1996): Simulating Ratios of Normalizing Constants via a Simple Identity. *Statistica Sinica*, 6, 831–60.

RICHARDSON, S. and GREEN, P.J. (1997): On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 59, 731-Ű792.

STEPHEN (2000): Bayesian Analysis of Mixture Models with an Unknown Number of Components: An Alternative to Reversible Jump Methods. *The Annals of Statistics*, 28, 1, 40–74.

TÜCHLER, R., FRÜHWIRTH-SCHNATTER, S., and OTTER, T. (2004): Bayesian Analysis of the Heterogeneity Model. *Journal of Business and Economic Statistics*, 22, 1, 2–15.

Keywords

LATENT CLASSES, BAYESIAN PROCEDURES, CONJOINT ANALYSIS

Comparison of Clustering Approaches for Mixed Variables

Andreja Bandelj[1], Mateja Budin[1], Tomi Deutsch[1], Andrej Kastrin[2], Ana Kolar[1], and Polona Kramar[1]

[1] Graduate program on Statistics, University of Ljubljana,
Kongresni trg 10, SI-1000 Ljubljana, Slovenia
[2] Division of Medical Genetics, Department of Obstetrics and Gynecology, University Medical Centre,
Slajmerjeva 3, SI-1000 Ljubljana, Slovenia

Abstract. The main goal of the study is the comparison of the clustering approach based on Gower's (dis)similarity coefficient (Gower, 1971) and some other clustering approaches for analyzing mixed data. For this purpose two mixed variables data sets are used: the first one relates to European Social Survey and the second one to the demographical data of the 200 world countries.

References

GOWER, J.C. (1971): A general coefficient of similarity and some of its properties. *Biometrics, 27, 857–871*.

Keywords

MIXED DATA, GOWER'S SIMILARITY COEFFICIENT, EUROPEAN SOCIAL SURVEY

Partition-Valued Random Variables

David Banks and Natesh Pillai

Institute of Statistics and Decision Sciences, Duke University,
Durham, NC 27708, USA

Abstract. In some situations, one has a sample in which each observation is a partition of a set. Such cases arise in card-sorting experiments in psychology, or when the same individuals are being clustered on separate features, as can happen in data mining when the number of variables is large. This research describes a statistical model for such data, and demonstrates different kinds of analysis for a card-sorting experiment that studies how well naive people can recognize architectural styles.

Keywords

PARTITIONS, CARD-SORTING EXPERIMENTS, CLUSTERING

Phylogenetic Trees, Additive Tree Metrics and Filiation of Manuscripts: The Case of the Benares Glose

Jean-Pierre Barthélemy[1,2], Marc Le Pouliquen[1,3], and Patrice Bertrand[1]

[1] Département Logique des Usages, Sciences Sociales et de l'Information & Tamcic, UMR CNRS 2872, Ecole Nationale Sup. des Télécommunications de Bretagne, Technopôle de Brest-Iroise, CS 83818, Brest (Breizh), France
[2] CAMS, UMR CNRS 8557, Ecole des Hautes Etudes en Sciences Sociales, 54 bd Raspail, 75270 Paris Cedex 06, France
[3] Université de bretagne occidentale, IUP Génie Mécanique et Productique, 6 av. Le Gorgeu - CS 93837, 29238 Brest Cedex 3, France

Abstract. As far as we know the use of phylogenetic trees to account for filiations of manuscripts goes back to Buneman (1971). The design of a *stemma codicum* (cf. Salemans, 2000) is one of the most rigorous approaches. It needs the reconstruction of the history of the text. Despite in our case we can observe some "contamination" phenomenon's (in biology the term is hybridation), a tree model sounds to be a good approximation. The employed method is essentially Addtree (cf. Sattah and Tversky, 1977) and its extensions to the "grouping method" by Luong (1988) together with the NJ method (cf. Saitou and Nei, 1987). The method used has been developed in editing manuscripts in Sanskrit according with the specificities of this language (in particular, the words are not separated).

References

BUNEMAN, P. (1971): Filiations of manuscripts. In: F.H. Hodson, D.G. Kendall and P. Tautu (Eds.): *Mathematics in archaeological and historical sciences*, Edimburg University Press, 387–395.

SALEMANS, B. J. P. (2000): Building stemmas with the computer in a cladistic, neo-Lachmannian, way: the case of fourteen text versions of Lanseloet van Denemerken. *Ph. D. thesis, Nijmegen: University Press.*

SATTAH, S. and TVERSKY, A. (1977): Additive similarity trees. *Psychometrika, 42, 319–345.*

LUONG, X. (1988): Méthodes d'analyse arborée. Algorithmes. Applications. *Ph. D. thesis, Thèse de doctorat, University of Paris V.*

SAITOU, N. and NEI, M. (1987): The neighbor-joining method: a new method for reconstruction of phylogenetic trees. *Mol. Biol., 4, 406–425.*

Keywords

PHYLOGENETIC TREES, TREE METRICS, FILIATION OF MANUSCRIPTS, SANSKRIT

Comparison of Clustering Approaches for Mixed Data: A Simulation Study

Petra Bastjančič, Nejc Bergant, Renata Blatnik, Miran Juretič, Mojca Omerzu, and Jure Željko

Graduate program on Statistics, University of Ljubljana,
Kongresni trg 10, SI-1000 Ljubljana, Slovenia

Abstract. One of the most successful clustering approaches for mixed-type variables is to use Gower's similarity coefficient among units (Gower, 1971). The effect of several different data transformations is studied on simulated interval-type variables as well as on different kinds of mixed-type variables. Additionally, several different types of simulated data structures are used to compare the effects of data transformation.

References

GOWER, J.C. (1971): A general coefficient of similarity and some of its properties. *Biometrics, 27, 857–871*.

Keywords

MIXED DATA, GOWER'S SIMILARITY COEFFICIENT , SIMULATED DATA, DATA TRANFORMATION

Analyzing the Structure of U.S. Patents Network

Vladimir Batagelj[1], Nataša Kejžar[2], Simona Korenjak-Černe[3], and Matjaž Zaveršnik[1]

[1] Department of Mathematics, FMF, University of Ljubljana, Jadranska 19, SI-1000 Ljubljana, Slovenia
[2] Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia
[3] Faculty of Economics, EF, University of Ljubljana Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia

Abstract. The U.S. patents network is a network of almost 3.8 millions patents (network vertices) from the year 1963 to 1999 (Hall et al., 2001) and more than 16.5 millions citations (network arcs). It is an example of a very large citation network.

We analyzed the U.S. patents network with the tools of network analysis in order to get insight into the structure of the network as an initial step to the study of innovations and technical changes based on patents citation network data.

In our approach the SPC (Search Path Count) weights, proposed by Hummon and Doreian (1989), for vertices and arcs are calculated first. Based on these weights vertex and line islands (Batagelj and Zaveršnik, 2004) are determined to identify the main themes of U.S. patents network. All analyses were done with `Pajek` – a program for analysis and visualization of large networks. As a result of the analysis the obtained main U.S. patents topics are presented.

References

BATAGELJ, V. and MRVAR, A. (2003): Pajek – Analysis and Visualization of Large Networks. In: Jünger, M., Mutzel, P., (Eds.): *Graph Drawing Software*. Springer, Berlin, 77-103.

BATAGELJ, V. and ZAVERŠNIK, M.: Islands – identifying themes in large networks. Presented at Sunbelt XXIV Conference, Portorož, May 2004.

HUMMON, N.P. and DOREIAN, P. (1989): Connectivity in a Citation Network: The Development of DNA Theory. *Social Networks, 11, 39-63*.

Keywords

PATENTS, CLUSTERING, NETWORK ANALYSIS, LARGE NETWORK, CITATION NETWORK, ACYCLIC NETWORK, SEARCH PATH COUNT WEIGHTS, ISLAND METHOD, VISUALIZATION

Three Dimensional Blockmodeling

Vladimir Batagelj[1], Patrick Doreian[2], and Anuška Ferligoj[3]

[1] Department of Mathematics, FMF, University of Ljubljana,
Jadranska 19, SI-1000 Ljubljana, Slovenia
[2] Department of Sociology, University of Pittsburgh,
2406 WWPH, Pittsburgh, PA 15260, USA
[3] Faculty of Social Sciences, University of Ljubljana,
Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia

Abstract. Currently, there are two broad approaches to blockmodeling: conventional blockmodeling and generalized blockmodeling. This paper considers both approaches in an effort to apply blockmodeling to three dimensional network structures viewed as three-mode network data. Such data arise naturally in many situations and include the following: multiple distinct relations, indicator relations for some underlying relation, full three-mode networks (where the modes are all distinct), and temporal networks. The approaches to blockmodeling such structures that we propose and develop are fourfold: indirect methods, coupled direct and indirect methods, graph theoretical methods, and full generalized blockmodeling. While temporal three dimensional networks have the same logical structure as the other three dimensional network structures, they cannot be treated in exactly the same fashion because the inclusion of time introduces major restrictions. A formalization of the three dimensional blockmodeling problem is presented together with a formal statement of the methods for solving this problem. These methods are applied to a variety of real three-mode network data sets.

Keywords

NETWORK ANALYSIS, GENERALIZED BLOCKMODELING, THREE-MODE NETWORK DATA, GRAPH, RELATION

Combining Data and Network Analysis

Vladimir Batagelj

University of Ljubljana, FMF, Dept. of mathematics
Jadranska 19, 1000 Ljubljana, Slovenia

Abstract. Let $\mathbf{U}$ be a *set of multivariate units* and d a *dissimilarity* on it. They determine two types of *multivariate networks*:

The k *nearest neighbors graph* $\mathbf{G}_N(k) = (\mathbf{U}, A)$

$$a(\mathrm{X}, \mathrm{Y}) \in A \Leftrightarrow \mathrm{Y} \text{ is among the } k \text{ closest neighbors of } \mathrm{X}$$

By setting for $a(\mathrm{X}, \mathrm{Y}) \in A$ its value to $w(a) = d(\mathrm{X}, \mathrm{Y})$ we obtain a network.
In the case of equidistant pairs of units we have to decide – or to include them all in the graph, or specify an additional selection rule.

The *fixed-radius neighbors graph* $\mathbf{G}_B(r) = (\mathbf{U}, E)$

$$e(\mathrm{X} : \mathrm{Y}) \in E \Leftrightarrow d(\mathrm{X}, \mathrm{Y}) \leq r, \qquad w(e) = d(\mathrm{X}, \mathrm{Y})$$

The multivariate networks provide a link between data analysis and network analysis. We will illustrate this with analysis of some classical data sets such as Fisher's Iris data. It seems that the r-neighbors network makes sense only when the 'density' of groups doesn't change.

There is no known subquadratic algorithm for determining multivariate networks. The basic paper on the subject is Fukunaga and Narendra (1975). Therefore the approach can't be applied on very large data sets.

In the case of (sparse) relational constraints we can efficiently determine the network by computing the dissimilarities only between linked units. This approach will be illustrated with an analysis of large network of neighboring territorial units.

References

FUKUNAGA, K., NARENDRA, P.M. (1975), A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, **C-24**, 750-753.

LI YANG (2006): Building k-Connected Neighborhood Graphs for Isometric Data Embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 5.

Keywords

NETWORKS, DISSIMILARITIES, NEAREST NEIGHBORS, CLUSTERING, RELATIONAL CONSTRAINTS, STATISTICAL PROGRAM R

Revealing Hidden Relevance in Complex Laddering Networks

Zenel Batagelj[1], Vladimir Batagelj[2], Bojan Korenini, and Vanja Govednik[1]

[1] CATI d.o.o.,
Tržaška 2 , SI-1000 Ljubljana, Slovenia
[2] Department of Mathematics, FMF, University of Ljubljana,
Jadranska 19, SI-1000 Ljubljana, Slovenia

Abstract. In this paper an alternative approach to quantification of laddering, which is one of the methods for identifying a productŠs benefits and values, will be presented. It tries to address some criticisms of existing approaches. Identifying benefits and values of products have gained increasing importance since they became - more than product intrinsic aspects - the basis for communication/positioning and the justification of distinction from other products within the same category. The method proposed leads to ladders ranked according importance and relevance, directly useful for the end clients. It can reveal new possibilities for positioning and values to trigger consumers. The method has the potential to redefine the clientsŠ creative process and actively participate in their communication strategy.

References

HOFSTEDE, F., AUDENAERT, A., STEENKAMP, J.-B. E. M., WEDEL M. (1998): An Investigation into the Association Pattern Technique as a Quantitative Approach to Measuring Means-End Chains, *International Journal of Research in Marketing, 15, 1, 37–50*.

KORENINI, B., BATAGELJ, Z. (2001): Segmenting the market according to consumer's benefits and value orientations - the application of structurally determined laddering method; *Paper presented at International Sunbelt XXI Social Network Conference, Budapest,Hungary*.

REYNOLDS, T. J., GUTMAN, J. (1988): Laddering Theory, Method, Analysis, and Interpretation, *Journal of Advertising Research, February/March, 11–31*.

Keywords

CHARACTERISTICS, BENEFITS, VALUES, LADDERING, ASSOCIATION PATTERN TECHNIQUE, STRUCTURALLY DETERMINED LADDERING

Spectral Clustering and Multidimensional Scaling: A Unified View

François Bavaud

Section d'Informatique et de Méthodes Mathématiques
Faculté des Lettres, Université de Lausanne
CH-1015 Lausanne, Switzerland

Abstract. Scalar products between features define similarities between objects, and reversible Markov chains define weighted graphs describing a stationary flow. It is natural to expect flows and similarities to be related: somehow, the exchange of flows between objects should enhance their similarity, and transitions should preferentially occur between similar states.

This paper formalizes the above intuition by demonstrating in a general framework that the symmetric matrices K and F possess an identical eigenstructure, where the kernel K is a measure of similarity between objects, and the symmetrized transition F is a measure of flows. Diagonalizing K yields principal components analysis as well as mutidimensional scaling, while diagonalizing F yields spectral clustering. Precisely, we demonstrate that the eigenvectors of K and F coincide and that their eigenvalues are related in a linear or non-linear way.

The general "$K - F$ connection" described here hence formalizes a theme whose various instances have already been encountered and addressed in the classical setup (such as correspondence analysis) or in the kernel and machine learning setup, at least implicitly. The relative generality of the present approach (weighted objects, weighted variables, weighted graphs) might provide some guidance for defining the appropriate objects (kernels, scalar products, similarities or affinities, etc.). Also, the same formalism permits to characterize a broad family of *separable auto-covariances*, relevant in spatial statistics.

Keywords

CORRESPONDENCE ANALYSIS, CHEEGER'S INEQUALITY, EUCLIDEAN DISSIMILARITIES, KERNEL METHODS, MARKOV CHAIN, MULTIDIMENSIONAL SCALING, NORMALIZED MINIMAL CUT, PRINCIPAL COMPONENTS ANALYSIS, SPECTRAL CLUSTERING, SPATIAL AUTOCORRELATION, WEIGHTED GRAPHS

Evaluation of Allocation Rules under Some Cost Constraints

Farid Beninel[1] and Michel Grun Rehomme[2]

[1] Université de POITIERS, UMR CNRS 6086
IUT- STID, 8 rue Archimède, 79000 Niort, FRANCE
[2] Université PARIS2, ERMES, UMR CNRS 7017
92 rue d'Assas, 75006 Paris, FRANCE

Abstract. Usually, assignment rules are built on the basis of an implicit assumption that the different cases of error have a same cost.
The difference between costs is more often considered when building the rule than in its tests and validation.
Here, we focus on the situation where given an assigned stratified sample and the errors cost values. In such a context, we study the behavior of a statistic-index of the rule quality. This index generalize *the corrected classified rate*. Our aim is to associate, to the observed value of such an index, a level of significance taking into account of the misallocations cost.
From a mathematical point of view, we deal with a *non linear programming* problem. In the case of three classes, an analytical solution is given and for the more general case, we propose an *ad hoc* optimization program.

References

ADAMS,N.M., HAND, D.J. (1999): Comparing classifiers when the misallocation costs are uncertain. *Pattergn recognition, 32, 1139-1147*

BREIMAN, L., FRIEDMAN, J., OHLSEN, R., STONE,C. (1984): *Classification and regression trees*. Wadsmorth, Belmont.

GIBBONS, J.D., PRATT, J.W. (1975): p-value: interpretation and methodology. *JASA, 29/1, 20-25*

GOVAERT, G. (2003): *Analyse des données*. Lavoisier serie "traitement du signal et de l'image", Paris, pp.362.

SEBBAN, M., RABASEDA,S., BOUSSAID,O.(1996): Contribution of related geometrical graph in pattern and recognition. In: E. Diday, Y. Lechevallier and O. Optiz (Eds.): *Ordinal and symbolic data Analysis*. Springer, Berlin, 167–178.

Keywords

COST ERRORS, MISALLOCATIONS, NON LINEAR PROGRAMMING, QUALITY INDEX, STATISTICAL LEARNING

Bayesian Models for Safety Design to Prevent Foreign Body Injuries in Children

Paola Berchialla[1], Silvia Snidero[2], Alex Stancu[2], Cecilia Scarinzi[2], Roberto Corradetti[2], and Dario Gregori[1]

[1] Department of Public Health and Microbiology, University of Torino, Via Santena 26, 10126 Torino, Italy
[2] Department of Statistics and Applied Mathematics, University of Torino, Piazza Arbarello 8, 10122 Torino, Italy

Abstract. The entry of a small item into the upper aero-digestive ways is one of the leading causes of injuries in children up to 14 years old. The European Survey on Foreign Bodies Injuries Study collected data from 19 European countries in the years 2000 to 2002 according to ICD9-CM 931 to 935. The goal of this paper is to show how the Bayesian models along with Markov chain Monte Carlo techniques can be used to formulate a model for use in a quantitative risk assessment aimed at identifying the critical factors that does not meet the necessary safety expectations Due to the skewness of the data, a log transformation was performed on continuous variables and a log normal model was compared to a gamma model. Finally, model selection was carried out based on MCMC simulations from posterior model parameter distributions. Inference, in the light of evidence, can be made on all domain variables making it possible to sample from the distributions of variables of interest such as volume or shape of objects which caused injuries. Results show how the knowledge of such distribution can be helpful in implementing a safety design of the products.

References

RIMELL, F. L., THOME Jr, A., STOOL, S., REILLY, J. S., RIDER, G., STOOL, D. and WILSON, C. L. (1996): Characteristics of objects that cause choking in children. *Journal of American Medical Association, 274(22),1763–1766*.

STOOL, D., RIDER, G. and WELLING, J. R. (1998): Human factors project: development of computer models of anatomy as an aid to risk management. *International Journal of Pediatrics and Otorhinolaryngology, 43(3), 217–227*.

RIDER, G., MILKOVICH, S., STOOL, D., WISEMAN, T., DORAN, C. and CHEN, X. (2000): Quantitative risk analysis. *Injury Control and Safety Promotion, 7(2),115–133*.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N., LUNN, D. (2003): WinBUGS Version 1.4 User Manual. MRC Biostatistic Unit, Cambridge, UK.

Keywords

QUANTITATIVE RISK ASSESSMENT, BAYESIAN MODELS, FOREIGN BODY INJURY

Paired-Hierarchies and Associated Dissimilarities

Patrice Bertrand

ENST Bretagne, LUSSI, 2 rue de la Châtaigneraie, 35576 Cesson Sévigné, France

Abstract. Given any collection $\mathcal{C}$ of nonempty subsets of a ground set S, we consider the relation $\sim$ defined on $\mathcal{C}$ by $A \sim B$ if either $A = B$ or $A \cap B \notin \{\emptyset, A, B\}$. Assuming that $\mathcal{C}$ contains the nontrivial subsets of S, it can be noticed that $\mathcal{C}$ is a hierarchy if and only if the relation $\sim$ is an equivalence relation whose classes are singletons. Our aim is to investigate the more general case where each equivalence class is either a singleton or a pair of subsets. We name such collections *paired-hierarchies*. In Bertrand (2002), it was proved that the paired-hierarchies (also called 2-3 hierarchies) define a pyramidal parsimonious clustering model. It results from the bijection introduced by Diday (1984) and by Fichet (1987), that the (weakly) indexed paired-hierarchies induce dissimilarities that are robinsonian. In this communication, we will present some structural properties of the paired-hierarchies together with their induced dissimilarities. Finally, we will present and illustrate the so-called APHC clustering algorithm that directly extends the well-known Agglomerative Hierarchical Clustering (AHC) algorithm in order to generate (weakly) indexed paired-hierarchies (cf. Bertrand (2002) and Chelcea *et al.* (2004)).

References

BERTRAND, P. (2002): Systems of sets such that each set properly intersects at most one other set - Application to pyramidal clustering. *Cahier du Ceremade*, University Paris-Dauphine, France, research report 2002-2 (final version to appear in *Discr. Appl. Maths*).

CHELCEA, S., BERTRAND, P. and TROUSSE, B. (2004): Un Nouvel Algorithme de Classification Ascendante 2-3 Hiérarchique. In: *Reconnaissance des Formes et Intelligence Artificielle (RFIA 2004)*, vol. 3, 1471 – 1480, Toulouse, France.

DIDAY, E. (1986): Orders and overlapping clusters in pyramids. In: J. De Leeuw et al. (Eds.): *Multidimensional Data Analysis*. DSWO Press, Leiden, 201 – 234.

FICHET, B. (1987): Data Analysis: geometric and algebraic structures. In: Y.A. Prohorov and V.V. Sazonov (Eds.): *Proceedings of the 1st World Congress of the BERNOULLI SOCIETY, Tachkent, 1986*, V.N.U. Science Press, vol. 2, 123 – 132.

Keywords

PYRAMIDAL PARSIMONIOUS CLUSTERING, HIERARCHIES, ROBINSONIAN DISSIMILARITIES

Simultaneous Model-Based Clustering of Data Arising from Different Populations

Christophe Biernacki

Laboratory of Mathematics, UMRS CNRS 8524, University Lille 1,
F-59655 Villeneuve d'Ascq, France

Abstract. It is common to perform clustering methods *independently* on different data sets while (i) all individuals are described by the same variables and (ii) partitions with identical meaning are expected in all data sets (males/females in biology for instance). In this situation, exhibiting a formal relation between these related data sets may allow to apply a clustering process *simultaneously* on all of them in order to improve estimated partitions.

In the standard multivariate Gaussian model-based clustering (see for instance Banfield and Raftery 1993), a conditional linear mapping between each couple of populations is established under few assumptions and several *interpopulation* models of constraints allow to control parsimony of this linear relationship (see Biernacki *et al.* 2002 in a generalized discriminant analysis context). They are also combined with some classical *intrapopulation* models of constraints between components (for instance homoscedasticity or heteroscedasticity) and associated parameters of this combination are estimated by using the EM algorithm (Dempster *et al.* 1977). Choosing between independent and simultaneous clustering and also between all models of constraints can be achieved with any standard information criterion.

A biological illustration is provided through three populations which differ morphologically over their geographical range (Thibault *et al.* 1997). It appears that, in this example, simultaneous clustering for estimating sex of birds highly overperformed standard independent clustering.

References

BANFIELD, J.D. and RAFTERY, A.E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics, 49, 803–821*.

BIERNACKI, C., BENINEL, F. and BRETAGNOLLE, V. (2002). A Generalized Discriminant Rule when Training Population and Test Population Differ on their Descriptive Parameters. *Biometrics, 58/2, 387–397*.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 34, 1–38*.

THIBAULT, J.C., BRETAGNOLLE, V. and RABOUAM, C. (1997). Cory's Shearwater Calonectris Diomedea. *Birds of Western Paleartic Update, 1, 75–98*.

Keywords

GAUSSIAN MIXTURES, EM ALGORITHM, DISSIMILAR POPULATIONS, INTERPOPULATION MODELS, BIOLOGY

Dependence in Interval-Valued Observations

Lynne Billard

University of Georgia, USA

Abstract. The problem of obtaining the covariance function between two interval-valued random variables (X,Y) is addressed. Previous attempts to derive this function include its incorporation as part of the regression coefficient in a linear regression model obtained by Billard and Diday (2000), and the extension of the methodology used by Bertrand and Goupil (2000) in their derivation of the empirical sample variance, both for observations from an interval-valued random variable. Those derivations are limited because they only apply to special cases. In the current work, a general formulation for the covariance function is presented, one which, e.g., captures the internal variation of observations with different interval widths. The new formulation is easily extended to modal interval-valued (i.e., histogram-valued) observations.

A Tree-Based Similarity for Evaluating Concept Proximities in an Ontology

Emmanuel Blanchard, Pascale Kuntz, Mounira Harzallah, and Henri Briand

Laboratoire d'informatique de Nantes Atlantique
Site École polytechnique de l'université de Nantes
rue Christian Pauc
BP 50609 - 44306 Nantes Cedex 3
emmanuel.blanchard@univ-nantes.fr

Abstract. The problem of evaluating semantic similarity in a network structure knows a noticeable renewal of interest linked to the importance of the ontologies in the semantic Web. Different semantic measures have been proposed in the literature to evaluate the strength of the semantic link between two concepts or two groups of concepts within either two different ontologies or the same ontology. This paper presents a theoretical study synthesis of some semantic measures based on an ontology restricted to subsumption links. We outline some limitations of these measures and introduce a new one: the Proportion of Shared Specificity. This measure which does not depend on an external corpus, takes into account the density of links in the graph between two concepts. A numerical comparison of the different measures has been made on different large size samples from WordNet.

References

FELLBAUM, C., editor (1998): *WordNet: An electronic lexical database*. MIT Press.

GANESAN, P., GARCIA-MOLINA, H., AND WIDOM, J. (2003): Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst.*, 21(1):64–93.

GRUBER, T. R. (1993): A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

LIN, D. (1998): An information-theoretic definition of similarity. In *Proc. of the 15th Int. Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann.

RADA, R., MILI, H., BICKNELL, E., AND BLETTNER, M. (1989): Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

RESNIK, P. (1995): Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th Int. Joint Conf. on Artificial Intelligence*, volume 1, pages 448–453.

Keywords

DISSIMILARITY, HIERARCHY, ONTOLOGY

Class Prototypes for Complex (Interval) Data

Hans-Hermann Bock

Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany

Abstract. A typical form of complex data is provided by interval data where each object k is described by a p-dimensional interval (hypercube) $x_k = [a_k, b_k] \in \mathcal{H}^p$ with $a_k, b_k \in R^p$ and $\mathcal{H}^p$ the set of all hypercubes in R^p. In this paper we consider the problem of defining, for a given set of n hypercubes $x_1, ..., x_n \in \mathcal{H}^p$ (= data), a suitable 'prototype' or 'center' z (in an appropriate family $\mathcal{Z}$ of candidates, as well as some extensions of this problem. Obviously, the optimal z will depend on the candidates' set $\mathcal{Z}$ ($= R^p$, $\mathcal{H}^p$,...).

Since in R^p the classical mean value $\overline{x} := (\Sigma_k x_i)/n$ is the solution of the optimality problem $g(z) := \Sigma_k ||x_k - z||^2 \rightarrow \min_{z \in R^p}$, suitable approaches start with a similar optimality problem $g(z) := \Sigma_k d(x_k, z) \rightarrow \min_{z \in \mathcal{Z}}$ where $d(x, z)$ is a suitable dissimilarity measure. Our paper presents a survey on solutions ('centrocubes') for various choices of d and $\mathcal{Z}$.

The problem can be generalized in various ways: considering a random vector (hypercube) X with a distribution P^X on R^p (instead of empirical data only), or embedding the problem into the framework of random set theory. In particular, we present a paradox with an unexpected centrocubes in the case of the multidimensional normal distribution.

References

BOCK, H.-H. (2002): Clustering Methods and Kohonen Maps for Symbolic Data. *J. Japan. Soc. Comput. Statistics 15, 1-13.*

BOCK, H.-H. (2005): Optimization in symbolic data analysis: dissimilarities, class centers, and clustering. In: D. Baier, R. Decker, L. Schmidt-Thieme (eds.): *Data Analysis and Decision Support.* Springer, Heidelberg, 2005, 3-10.

BOCK, H.-H. and DIDAY, E. (2000): *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data.* Studies in Classification, Data Analysis, and Knowledge Organization. Springer Verlag, Heidelberg-Berlin.

BOCK, H.-H., and PÄRNA, K. (2005): *Fitting rectangles to multivariate distributions.* Working paper, RWTH Aachen University.

MOLCHANOV, I. (1997): Statistical Problems for Random sets. In: J. Goutsias (ed.): *Random Sets: Theory and Applications. Springer, Heidelberg, 27-45.*

NORDHOFF, O. (2003): *Erwartungswerte zufälliger Quader.* Diploma thesis, Institute of Statistics, RWTH Aachen University.

Keywords

CLASS PROTOTYPES, CENTROCUBES, INTERVAL DATA, SYMBOLIC DATA

Crisp Partitions Induced by a Fuzzy Set

Slavka Bodjanova

Department of Mathematics, Texas A&M University-Kingsville,
MSC 172, Kingsville, TX 78363, U.S.A.

Abstract. Relationship between fuzzy sets and crisp partitions defined on the same finite set of objects X is studied. Granular structure of a fuzzy set is described by rough fuzzy sets and the quality of approximation of a fuzzy set by a crisp partition is evaluated. Special attention is given to partitions called scales, whose clusters can be labeled by terms from an ordered linguistic scale. Classification of membership grades of a fuzzy set into linguistic categories is discussed. Measure of rough dissimilarity between clusters $C_r,\ C_s$ from a crisp partition of X with respect to a fuzzy set A defined on X is introduced. This measure evaluates how the roughness of approximation of membership grades $A(x), x \in C_r \cup C_s$ increases when we approximate $A(x)$ by the coarser cluster $C_r \cup C_s$ instead of by two separate clusters C_r and C_s. Rough dissimilarity will be used in the search for a reasonable 2-granule structure of A and consequently for defuzzification of A. If the quality of approximation of A by 2-cluster partition is low, we will search for a reasonable k-granule structure, $k > 2$. Finally, we will use rough dissimilarity to create a fuzzy proximity relation on the set of all elementary crisp clusters induced by A. Proximity relation will lead to the construction of a hierarchical crisp approximation of fuzzy set A.

References

DUBOIS, D. and PRADE, H. (1990): Rough fuzzy sets and fuzzy rough sets. *International Journal of General Systems, 17, 191–229.*

KLIR, G.J. and YUAN, B. (1995): *Fuzzy sets and fuzzy logic: Theory and applications.* Prentice Hall, Upper Saddle River.

PAWLAK, Z. (1982): Rough sets. *International Journal of Computer and Information Sciences, 11, 341–356.*

BILGIC, T. and TURKSEN, I.B. (2000): Measurement of membership functions: Theoretical and empirical work. In: D. Dubois, H.Prade (Eds.): *Fundamentals of fuzzy sets.* Kluwer, Dordrecht.

VIERTL, R. (1996): *Statistical methods for non-precise data.* CRC Press, Boca Raton.

Keywords

FUZZY SETS, ROUGH SETS, ROUGH FUZZY SETS, PARTITIONS, MEASURE OF DISSIMILARITY, PROXIMITY RELATION

Asymmetric Multidimensional Scaling with External Information: Some Possible Approaches

Giuseppe Bove

Dipartimento di Scienze dell'Educazione,
Università degli Studi *Roma Tre*, Italy
bove@uniroma3.it

Abstract. Asymmetric relationships contained in square data matrices like proximities (e.g. similarity ratings), preferences (e.g. socio-matrices), flow data (e.g. import-export, brand switching), can be represented in low-dimensional spaces by scalar product or Euclidean distance models (MDS models), opportunely modified by increasing the number of parameters (see e.g. Zielman and Heiser, 1996). In many applications additional information (external information) on the objects is available that could be conveniently incorporated in the data analysis. For instance, this allows to analyze the contribution of variables suggested from theoretical knowledge to the explanation of the relationships in the data. To this aim many methods were proposed in the context of symmetric MDS (see e.g. Borg and Groenen 1997, chapter 10), while a lack of proposals seems to characterize asymmetric MDS. In this communication some possible approaches to asymmetric multidimensional scaling with external information to analyze graphically asymmetric proximity matrices are discussed. In particular, a method based on the unique decomposition of the data matrix in its symmetric and skew-symmetric components, recently proposed by Bove and Rocci (2004), and a proposal to incorporate external information in biplot method are considered. The presented methods allow joint or separate analyzes of symmetry and skew-symmetry.

References

BORG, I. and GROENEN P. (1997): *Modern multidimensional scaling. Theory and applications*. Springer, Berlin.

BOVE, G. and ROCCI R. (2004): A method of asymmetric multidimensional scaling with external information. In: *Proceedings of the XLII Scientific Meeting of the Italian Statistical Society*. CLEUP, Padova, 631-634.

ZIELMAN, B. and HEISER W. J. (1996): Models for asymmetric proximities. *British Journal of Mathematical and Statistical Psychology, 49, 127–146.*

Keywords

VISUALIZATION, MULTIDIMENSIONAL SCALING, ASYMMETRY

A Novel Mixture-Model Cluster Analysis with Genetic EM Algorithm and Information Complexity as the Fitness Function

James E. Wicker[1], **Hamparsum Bozdogan**[2], and Halima Bensmail[3]

[1] Department of Physics and Astronomy,
University of Tennessee,
Knoxville, Tennessee 37996-0532 U.S.A.
[2] Department of Statistics,Operations, and Management Science
University of Tennessee,
Knoxville, Tennessee 37996-0532 U.S.A.
[3] Department of Statistics,Operations, and Management Science
University of Tennessee,
Knoxville, Tennessee 37996-0532 U.S.A.

Abstract. Methods of clustering based on Genetic Algorithms (GAŠs) offer new and novel approaches to analyzing complex multivariate data to determine the patterns and the number of clusters in the data. The traditional methods based on the iterative EM algorithm can have poor accuracy in parameter estimation, especially when the clusters overlap and show complex covariance structures. To improve this performance, in this paper, for the first time, we introduce a new and novel Genetic EM (GEM) Algorithm that does not show strong dependence on initial conditions and utilizes the global search properties of genetic algorithms (GAs) to accurately estimate model parameters describing the multivariate clusters. Our method is flexible to use Genetic K-Means (GKM) or Genetic Regularized Mahalanobis (GARM) distances to compute the initial cluster parameters, with little difference in the final results. This innovation allows our algorithm to find optimal parameter estimates of complex hyperellisoidal clusters. We develop and score the information complexity (ICOMP) criterion of Bozdogan (1994a,b, 2004) as our fitness function to choose the number of clusters present in the data sets and compare our results with other model selection criteria.

We develop the GEM approach to handle both normal (Gaussian) and non-normal (non-Gaussian) large dimensional heterogeneous data sets using adaptive multivariate mixtures of kernels of Bensmail and Bozdogan (2006a,b) to achieve flexibility in currently practiced mixture-model clustering techniques and to obtain both accurate classification error rates and accurate cluster parameter estimates.

We will highlight and demonstrate our approach on simulated data sets with different configurations and on complex real data sets collected from astronomical observations. We believe that these new methods can be implemented in data mining applications and in complex pattern recognition problems in many cross-disciplinary fields.

References

BOZDOGAN, H. (1994a): Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity. In *Multivariate Statistical Modeling, Vol. 2, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, H. Bozdogan (ed.), Kluwer Academic Publishers, Dordrecht, the Netherlands, 1994, pp. 69-113.

BOZDOGAN, H. (1994b): Choosing the Number of Clusters, Subset Selection of Variables, and Outlier Detection in the Standard Mixture-Model Cluster Analysis. Invited paper in *New Approaches in Classification and Data Analysis*, E. Diday et al. (Eds.), Springer-Verlag, New York, 1994, pp. 169-177.

BOZDOGAN, H. (2004): Intelligent Data Mining with Information Complexity, Statistical Modeling, and Genetic Algorithms: A Three-Way Hybrid. In *Statistical Data Mining & Knowledge Discovery*, H. Bozdogan (Ed.), Chapman & Hall/CRC, 2004, 15-56.

BENSMAIL, H. and BOZDOGAN, H. (2006a): Adaptive Multivariate Kernel Mixture-Model Cluster Analysis for Mixed Data. Working paper.

BENSMAIL, H. and BOZDOGAN, H. (2006b): Bayesian Cluster Analysis for Gaussian and Non-Gaussian Data With Missing Values. Working paper.

Keywords

MIXTURE-MODEL CLUSTER ANALYSIS, GENETIC EM ALGORITHM, INFORMATION COMPLEXITY

The Data Sets from Aitchison's Book in the "Compositions" Package

Matevž Bren[1,2]

[1] Faculty of Organisational Sciences, University of Maribor, Kidričeva 55[a], SI-4000 Kranj, Slovenia
[2] Institute of Mathemetics, Phisics and Mechanics, Jadranska 19, SI-1000 Ljubljana, Slovenia

Abstract. Compositions (compounds, mixtures, alloy ...) can be represented with vectors of the portions of individual components. The portions are nonnegative and they have constant sum like 100% or 1, in case of full compositions. In case of subcompositions where meaningless parts have been removed, the sum of parts is meaningless. In accordance with this nature of the data a suitable sample space and adequate methods should be implemented for the analysis.

In June 2005 a "compositions" package, authors K. Gerald van den Boogaart and Raimon Tolosana Delgado was published and is now available at

`http://cran.r-project.org/src/contrib/Descriptions/compositions.html`

The package supports four different multivariate scales represented by four classes. The classes differ by the assumption whether or not the total amount is meaningful for the problem and whether the geometry of the differences is a relative (log-scale) distance or a absolute (Euclidean) distance. For the analysis different graphical presentations, descriptive statistics and multivariate methods are implemented.

In the package we will include the data sets from Aitchison's book the groundwork on compositional data analysis for twenty years. There are forty data sets that serve as examples in the book and wil be available with the "compositions" package under the GNU Public Library Licence Version 2.

With these data sets also examples of data analysis and graphical presentations will be provided.

References

AITCHISON, J. (1986): *The Statistical Analysis of Compositional Data*. Chapman & Hall, New York.

Van den BOOGAART, K. G., TOLOSANA R. (2005): The compositions Package. `http://cran.r-project.org/src/contrib/Descriptions/compositions.html`

BREN, M. and BATAGELJ, V. (2005): Compositional data analysis with R. In: G. Mateu-Figueras, C. Barceló-Vidal (Ed.): *CODAWORK'05*. Girona: University of Girona.

Keywords

COMPOSITIONAL DATA, COMPOSITIONS PACKAGE, R LANGUAGE AND ENVIRONMENT, GRAPHICAL PRESENTATIONS

On the Analysis of Symbolic Data

Paula Brito

Faculdade de Economia/NIAAD-LIACC, Universidade do Porto
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal

Abstract. Symbolic data extend the classical tabular model, where each individual, represented by a row, takes exactly one value for each variable represented by a column, by allowing multiple, possibly weighted, values for each variable. New variable types - interval, categorical multi-valued and modal variables - have been introduced, which allow representing variability and/or uncertainty inherent to the data.

But are we still in the same framework when we allow for the variables to take multiple values? Are the definitions of basic notions still so straightforward? What properties remain valid?

In this talk we will discuss some issues that arise when trying to apply classical data analysis techniques to symbolic data. The central question of the evaluation of dispersion, and the consequences of different possible choices in the design of multivariate methods, will be addressed.

Dispersion is a key issue in clustering, since the result of any clustering method depends heavily on the scales used for the variables; natural clustering structures can sometimes only be detected after an appropriate rescaling of variables. The standardization problem has been addressed by De Carvalho, Brito & Bock, and three standardization techniques for interval-type variables have been proposed.

Furthermore, many exploratory multivariate methodologies rely heavily on the notion of linear combination and on the properties of dispersion measures under linear transformations. This problem has been addressed in a recent work of Duarte Silva & Brito in the context of linear discriminant analysis of interval data.

Keywords

SYMBOLIC DATA ANALYSIS, INTERVAL DATA, STANDARDIZATION, CLUSTERING, DISCRIMINANT ANALYSIS

Assessing Congruence among Ultrametric Distance Matrices

Véronique Campbell, Pierre Legendre, and François-Joseph Lapointe

Département de sciences biologiques, Université de Montréal,
C.P. 6128, Succ. Centre-ville, Montréal, Québec, H3C 3J7, Canada

Abstract. Ultrametric matrices representing different dendrograms often have to be combined in a single multivariate analysis. Before combining these matrices, it is important to determine whether they convey the same information. A test of congruence among distance matrices (CADM) has been developed by Legendre and Lapointe (2004) to test the null hypothesis that the matrices are incongruent with one another. CADM is an extension of the Mantel test to more than two distance matrices. Previous simulations have shown that this test has a correct rate of type I error and good power when applied to independently-generated distance matrices. In this study, we tested the type I error rate and power of CADM with ultrametric matrices. We used different numbers of randomly generated dendrograms (as described by Lapointe and Legendre 1991) to test the type I error rate when H_0 is true by construct. The power of the test was assessed through simulations for partly similar matrices. An application to the classification of single-malt Scotch whiskies will also be presented (Lapointe and Legendre 1994).

References

LAPOINTE, F.-J. and LEGENDRE, P. (1991): The generation of random ultrametric matrices representing dendrograms. *Journal of Classification, 8, 177–200*.

LAPOINTE, F.-J. and LEGENDRE, P. (1994): A classification of pure malt Scotch whiskies. *Applied Statistics, 43, 237–257*.

LEGENDRE, P. and LAPOINTE, F.-J. (2004): Assessing congruence among distance matrices: single-malt Scotch whiskies revisited. *Australian and New Zealand Journal of Statistics, 46, 615–629*.

Keywords

CONGRUENCE, POWER, TYPE I ERROR, ULTRAMETRIC DISTANCE MATRICES

Empirical Comparison of a Divisive Clustering Method with the Ward and the K-Means Methods

Marie Chavent[1] and Yves Lechevallier[2]

[1] Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,
Université Bordeaux1, 351 Cours de la libération, 33405 Talence Cedex, France
[2] Institut National de Recherche en Informatique et en Automatique,
Domaine de Voluceau-Rocquencourt B.P.105, 78153 Le Chesnay Cedex, France

Abstract. DIVCLUS-T is a descendant hierarchical clustering method based on the same monothetic approach than classification and regression trees but from an unsupervised point of view. The aim is not to predict a continuous variable (regression) or a categorical variable (classification) but to construct a hierarchy of partitions. The dendrogram of the hierarchy is easy to interpret and can be read as decision tree. An example of this new type of dendrogram is given on a small categorical dataset. DIVCLUS-T is then compared empirically with two polythetic clustering methods: the Ward ascendant hierarchical clustering method and the k-means partitional method. The three algorithms are applied and compared on six databases of the UCI Machine Learning repository.

References

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984): *Classification and regression Trees*, C.A:Wadsworth.

CHAVENT, M. (1998): A monothetic clustering method. *Pattern Recognition Letters, 19, 989–996*.

HETTICH, S., BLAKE, C.L. and MERZ, C.J. (1998): *UCI Repository of machine learning databases, http://www.ics.uci.edu/ mlearn/MLRepository.html*. Irvine, CA: University of California, Department of Information and Computer Science.

Keywords

DESCENDANT HIERARCHICAL CLUSTERING, MONOTHETIC CLUSTER, DECISION DENDROGRAM, INERTIA CRITERION

Iterated Boosting for Outlier Detection

Nathalie Cheze[1,2] and Jean-Michel Poggi[1,3]

[1] Laboratoire de Mathématique – U.M.R. C 8628, "Probabilités, Statistique et Modélisation", Université Paris-Sud, Bât. 425, 91405 Orsay cedex, France
[2] Université Paris 10-Nanterre, Modal'X, France
[3] Université Paris 5, France

Abstract. A procedure for detecting outliers in regression problems based on information provided by boosting trees is proposed. Boosting is meant for dealing with observations that are hard to predict, by giving them extra weights. In the present paper, such observations are considered to be possible outliers, and a procedure is proposed that uses the boosting results to diagnose which observations could be outliers. The key idea is to select the most frequently resampled observation along the boosting iterations and reiterate boosting after removing it. The selection criterion is based on Tchebychev's inequality applied to the maximum over the boosting iterations of the average number of appearances in bootstrap samples. So the procedure is noise distribution free. A lot of well-known bench data sets are considered and a comparative study against two classical competitors allows to show the value of the method.

References

CHEZE, N. and POGGI, J-M. (2005): Outlier Detection by Boosting Regression Trees. *Preprint 2005-17, Orsay* . www.math.u-psud.fr/ biblio/ppo/2005/

DRUCKER, H. (1997): Improving Regressors using Boosting Techniques. In: Proc. of the 14th Int. Conf. on Machine Learning. Morgan Kaufmann, 107–115.

GEY, S. and POGGI, J-M. (2006): Boosting and Instability for Regression Trees. *Computational Statistics & Data Analysis, 50, 2, 533-550*.

ROUSSEEUW, P.J. and LEROY, A. (1987): *Robust regression and outlier detection*. Wiley.

VERBOVEN, S. and HUBERT, M. (2005): LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75, 127-136.

Keywords

BOOSTING, OUTLIERS, REGRESSION, TREES

A Link between the Asymmetric MDS and the Analysis of Contingency Table

Naohito Chino and Shingo Saburi

Department of Psychology, Aichi Gakuin University,
Nisshin-city, Aichi, Japan

Abstract. The examination of some types of symmetry contained in an NĄ NĄ M contingency table provides an important link between two different areas of statistical research and application. These are the asymmetric multidimensional scaling (MDS) in psychometrics and the other the analysis of contingency table in statistics. Zielman & Heiser (1996) and De Rooij & Heiser (2003, 2005) have made an inroad regarding this problem with the usual two-way contingency table. A maximum likelihood asymmetric MDS method proposed recently by Chino and Saburi (Chino, 1992; Saburi & Chino, 2004, 2005, 2006) assumes that similarity or dissimilarity judgments of subjects are done by the method of successive categories (Torgerson, 1958). As a by-product we have an $N \times N \times M$ contingency table, where N is the number of, say, members of a small group, and M is the number of successive categories. One approach to such a three-way contingency table may be to use a two-step procedure, in which we may examine some types of asymmetry contained in the table first by combining some traditional tests of symmetry and related tests, and then apply a congruous asymmetric MDS model with the uncovered feature of asymmetry to the data. We shall discuss advantages and shortfalls of this approach and suggest further approaches to this kind of data.

References

CHINO, N. (1992): Metric and nonmetric Hermitian canonical models for asymmetric MDS. *Proceedings of the 20th Annual Meeting of the Behaviormetric Society of Japan*, 246-249.

SABURI, S. and CHINO, N. (2005): A maximum likelihood method for asymmetric MDS (2). *Proceedings of the 33rd Annual Meeting of the Behaviormetric Society of Japan*, 404-407.

De ROOIJ, M. and HEISER, W. J. (2005): Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika, 70, 99-122*.

TORGERSON, W. S. (1958): *Theory and methods of scaling*. Wiley, New York.

ZIELMAN, B. and HEISER, W. J. (1996): Models for asymmetric proximities. *British Journal of Mathematical and Statistical Psychology, 49, 127-146*.

Keywords

ASYMMETRIC MDS, TESTS FOR SYMMETRY, CONTINGENCY TABLES

Improved Fréchet Distance for Time Series

A. Chouakria-Douzal[1] and P. Nagabhushan[2]

[1] TIMC-IMAG, Université Joseph Fourier Grenoble 1,
F-38706 LA TRONCHE Cedex, France
Ahlame.Douzal@imag.fr
[2] Dept. of Studies in Computer Science, University of Mysore
Manasagangothri, Mysore, Karnataka- 570 006, India
pn@amrita.edu

Abstract. This paper focuses on the Fréchet distance introduced by Maurice Fréchet in 1906 to account for the proximity between curves (Fréchet (1906)). The major limitation of this proximity measure is that it is based on the closeness of the values independently of the local trends. To alleviate this set back, we propose a dissimilarity index extending the above estimates to include the information of dependency between local trends. A synthetic dataset is generated to reproduce and show the limited conditions for the Fréchet distance. The proposed dissimilarity index is then compared with the Fréchet estimate and results illustrating its efficiency are reported.

Keywords

TIME SERIES, FRÉCHET DISTANCE, LOCAL CORRELATION

Looking for a Typology of Sexual Practices by Adolescents in Three Secondary Schools in Quebec

Guy Cucumel[1], Véronique Mallandain[2], and Marie-Marthe Cousineau[2]

[1] École des sciences de la gestion, Université du Québec à Montréal,
C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8
(e-mail: cucumel.guy@uqam.ca)
[2] École de criminologie, Université de Montréal,
C.P. 6128, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3J7
(e-mail: veroniquemallandain@hotmail.com)
(e-mail: cousinem@cicc.umontreal.ca)

Abstract. There is a great deal of misunderstanding in Quebec, as well as in the rest of Canada, about the sexual practices of young people. Young people are often criticised for a perceived lack of control in their sexual lives, particularly for becoming sexually active at an earlier age, for the number of sexual partners and their participation in 'unconventional' sexual acts. Involvement in such sexual behaviours is thought to contribute to an increase in juvenile prostitution. We argue that this image of the sexual behaviour of young people applies only to a small core, and that the sexual practices of the young are much more conventional than is generally thought.

We present a taxonomy of sexual practices that depicts the characteristic sexual behaviours of adolescents. The data used to construct this taxonomy are drawn from a self-completion questionnaire exploring various dimensions of young peoples' lives, from a sample of more than 1600 students from three secondary schools. The taxonomy is derived from these data through a combination of factor and cluster analysis (Lebart, Morineau and Piron, 2000) and provides a portrait of the sexual practices of contemporary Quebecois adolescents.

References

LEBART, L., MORINEAU A. and PIRON M. (2000): *Statistique exploratoire multidimensionnelle*. Dunod, Paris.

Keywords

CLUSTER ANALYSIS, SEXUAL PRACTICES, ADOLESCENTS

Comparison of Distance Indices Between Partitions

L. Denœud[1 2] and A. Guénoche[3]

[1] École nationale supérieure des télécommunications, 46, rue Barrault, 75634 Paris cedex 13 (e-mail: denoeud@infres.enst.fr)
[2] CERMSEM CNRS-UMR 8095, MSE, Université Paris 1 Panthéon-Sorbonne, 106-112, boulevard de l'Hôpital, 75647 Paris cedex 13
[3] Institut de Mathématiques de Luminy, 163, avenue de Luminy, 13009 Marseille (e-mail: guenoche@iml.univ-mrs.fr)

Abstract. In this paper, we compare five classical distance indices on $\mathcal{P}_n$, the set of partitions on n elements. First, we recall the definition of the *transfer distance* between partitions and an algorithm to evaluate it. Then, we build sets $\mathcal{P}_k(P)$ of partitions at k transfers from an initial partition P. Finally, we compare the distributions of the five index values between P and the elements of $\mathcal{P}_k(P)$.

We conclude, according to the studied partitions, that the Jaccard and the Johnson indices are the most accurate to compare close partitions. The corrected Rand index comes after. We also illustrate that these classical indices are correlated with the small values of the transfer distance, but only when n is large enough. For small n, the transfer distance is a much appropriate measure of the closeness of partitions.

Keywords

PARTITIONS, DISTANCE INDICES, TRANSFER DISTANCE, ENUMERATION OF CLOSE PARTITIONS

One-Mode Additive Clustering for Two-Way Two-Mode Data: A Comparison of Algorithms

Dirk Depril and Iven Van Mechelen

Departement of Psychology, Katholieke Universiteit Leuven
Tiensestraat 102, 3000 Leuven, Belgium

Abstract. Shepard & Arabie (1979) proposed the ADCLUS model as an additive model for two-way one-mode object by object similarity data based on overlapping object clusters. In case of two-way two-mode object by variable data, one may also wish to construct an additive model based on an overlapping clustering of the objects. One possible solution to this problem could be to first derive similarities between all pairs of objects and then to subsequently fit the ADCLUS model to them. Yet, there is an infinite number of ways to derive one-mode similarities from two-mode data and as such, one may wish to go for a direct additive clustering of two-mode data. For this purpose, Mirkin (1990) proposed an additive clustering model for two-mode data that implies overlapping object clusters. To fit this model to a given data set (with a least squares loss function), Mirkin further proposed a sequential fitting (SEFIT) algorithm. Unfortunately, as shown in this talk, Mirkin's algorithm has some problems. As a way out we propose three novel algorithms: two of an alternating least squares type and a simulating annealing approach. Simulation results will be presented on the absolute and comparative performance of the different algorithms.

References

DEPRIL, D. & VAN MECHELEN, I. (2005): One mode additive clustering of multiway data. In: J. Janssen & (Eds.): *Applied stochastic models and data analysis.* Brest, France, 724–729.

MIRKIN, B. G. (1990): A sequential fitting procedure for linear data analysis models. *Journal of Classification, 7, 167–195*.

SHEPARD, R. & ARABIE, P. (1979): Additive clustering: Representations of similarities as combinations of discrete overlapping properties. *Psychological Review, 86, 87–123*.

Keywords

ADDITIVE CLUSTERING, TWO-MODE DATA, COMBINATORIAL DATA ANALYSIS, SIMULATION STUDY

A Dynamic Clustering Method for Mixed Feature-Type Symbolic Data

Renata M.C.R. de Souza, Francisco de A.T. de Carvalho, and Daniel Ferrari Pizzato

Centro de Informatica - CIn/UFPE, Av. Prof. Luiz Freire, s/n, Cidade Universitaria, CEP 50740-540, Recife-PE, Brasil, {rmcrs,fatc,dfp}@cin.ufpe.br

Abstract. A dynamic clustering method for mixed feature-type symbolic data is presented. The proposed method needs a previous pre-processing step to transform Boolean symbolic data into modal symbolic data. The presented dynamic clustering method has then as input a set of vectors of modal symbolic data and furnishes a partition and a prototype to each class by optimizing an adequacy criterion based on a suitable squared Euclidean distance. To show the usefulness of this method, an application with a real interval symbolic is considered. The car data set consists of a set of 33 car models described by 8 interval and 3 nominal variables. Our aim is to compare the approach presented in (Chavent et al (2003)), which transforms interval symbolic data on modal symbolic data represented by non-cumulative weight distributions, with the approach presented in this paper, which transforms interval symbolic data on modal symbolic data represented by cumulative weight distributions. The accuracy of the results furnished by the clustering method introduced in this paper was assessed by the adjusted Rand index. These results clearly show that the accuracy of the clustering method using cumulative weight vectors of interval data is superior to that which uses non-cumulative weight vectors.

References

BOCK, H. H. and DIDAY, E. (2000): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Springer, Heidelberg.

CHAVENT, M., De CARVALHO, F. A. T., LECHEVALLIER, Y. and VERDE, R. (2003). Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle. *Revue de Statistique Appliquée, v. LI, n. 4, p. 5–29*.

SOUZA, R. M. C. R. and De CARVALHO, F. A. T. (2004): Clustering of interval data based on city-block distances. *Pattern Recognition Letters, **25** (3), 353-365*.

Keywords

SYMBOLIC DATA ANALYSIS, DYNAMICAL CLUSTERING, INTERVAL DATA, MODAL DATA, EUCLIDEAN DISTANCE

Biclustering of Microarray Data in a Bayesian Framework

Thomas Dhollander, Qizheng Sheng, and Yves Moreau

Department of Electrical Engineering (ESAT), BioI-SISTA-SCD, Katholieke Universiteit Leuven.
Kasteelpark Arenberg 10, 3001 Heverlee-Leuven, Belgium

Abstract. Biclustering is the simultaneous search for a set of samples and a set of attributes for which these samples 'belong together'. In bioinformatics, this problem is important for the analysis of microarray datasets, where gene expression patterns (samples) are measured across a set of biological conditions (attributes). Indeed, genes are tightly coexpressed (correlated expression patterns) only in the subset of experimental conditions where common parts of their regulatory programs are active. Other - possibly overlapping - parts of those programs are triggered under other experimental conditions, leading to the notion of overlapping transcriptional modules. Classical clustering mostly fails to detect such structure in noisy data.

A Bayesian probabilistic framework shows promise for bicluster pattern discovery, directed by a soft query in the form of prior probabilities. It allows biologists to tackle specific biological problems by querying their data with a seed of genes, which they believe to have a common function. We apply Gibbs sampling and EM-related approaches to identify biclusters in the neighborhood of these seeds.

Additionally, we extend this probabilistic framework to a multiple-module setting, using functional ontology tree hierarchies and corresponding known gene annotations as prior information on the hierarchy of the modules.

References

SHENG Q., DE MOOR B. and MOREAU Y. (2003): Biclustering microarray data by Gibbs sampling. *Bioinformatics, 19(Suppl. 2), II196–II205*.

Keywords

GENE EXPRESSION DATA, BICLUSTERING, GIBBS-SAMPLING, FUNCTIONAL ANNOTATION, BIOINFORMATICS

A New Efficient Method for Assessing Missing Nucleotides in DNA Sequences in the Framework of a Generic Evolutionary Model

Abdoulaye Baniré Diallo[1], Vladimir Makarenkov[2], Mathieu Blanchette[1], and François-Joseph Lapointe[3]

[1] McGill Centre for Bioinformatics and School of Computer Science, McGill University 3775 University Street, Montreal, Quebec, H3A 2A7, Canada
[2] Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), H3C 3P8, Canada
[3] Département de sciences biologiques, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal (Québec), H3C 3J7, Canada

Abstract. The problem of phylogenetic inference from datasets including incomplete characters is among the most relevant issues in systematic biology. In this paper, we propose a new probabilistic method for estimating unknown nucleotides before computing evolutionary distances. It is developed in the framework of the Tamura-Nei evolutionary model (Tamura and Nei (1993)). The proposed strategy is compared, through simulations, to existing methods "Ignoring Missing Sites" (IMS) and "Proportional Distribution of Missing and Ambiguous Bases" (PDMAB) included in the PAUP package (Swofford (2001)).

References

DIALLO, Ab. B., DIALLO, Al. B. and MAKARENKOV, V. (2005): Une nouvelle méthode efficace pour l'estimation des données manquantes en vue de l'inférence phylogénétique. In: *Proceeding of the 12th meeting of Société Francophone de Classification*.Montréal, Canada, *121–125*.

HUELSENBECK, J. P. (1991): When are fossils better than existent taxa in phylogenetic analysis? *Systematic Zoology, 40, 458–469*.

SWOFFORD, D. L. (2001): PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. *Sinauer Associates*, Sunderland, Massachusetts.

TAMURA, N. and NEI, M. (1993): Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution, 10/3, 512–526*.

WIENS, J. J. (1998): Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology, 52, 528–538*.

WIENS, J. J. (2003): Does adding characters with missing data increase or decrease phylogenetic accuracy. *Systematic Biology, 47, 625–640*.

Keywords

MISSING DATA, PHYLOGENETIC INFERENCE, PHYLOGENETIC TREE, EVOLUTIONARY MODEL, DNA SEQUENCES

Model Selection for the Binary Latent Class Model A Monte Carlo Simulation

José G. Dias

Department of Quantitative Methods – UNIDE,
Higher Institute of Social Sciences and Business Studies – ISCTE,
Av. das Forças Armadas, Lisboa 1649–026, Portugal
jose.dias@iscte.pt

Abstract. This paper addresses model selection using information criteria for binary latent class (LC) models. A Monte Carlo study sets an experimental design to compare the performance of different information criteria for this model, some compared for the first time. Furthermore, the level of separation of latent classes is controlled using a new procedure (Dias, 2004). The results show that the Akaike information criterion with 3 as penalizing factor (AIC3: Bozdogan, 1993) has a balanced performance for binary LC models.

References

BOZDOGAN, H. (1993): Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix. In: O. Opitz, B. Lausen, and R. Klar (Eds.): *Information and Classification, Concepts, Methods and Applications*. Springer, Berlin, 40–54.

DIAS, J.G. (2004): Controlling the Level of Separation of Components in Monte Carlo Studies of Latent Class Models. In: D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul (Eds.): *Classification, Clustering, and Data Mining Applications*. Springer, Berlin, 77–84.

Keywords

MIXTURE MODELS, LATENT CLASS MODELS, BINARY DATA, MODEL SELECTION, LEVEL OF SEPARATION OF COMPONENTS, INFORMATION CRITERIA, MONTE CARLO STUDIES, LAPLACE-EMPIRICAL CRITERION

Spatial Classification

Edwin Diday

LISE-CEREMADE,Université Paris IX Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris $16^{iéme}$. France.
diday@ceremade.dauphine.fr

Abstract. The aim of a spatial classification is to position the units on a spatial network and to give simultaneously a set of structured classes of these units *"compatible" with the network*. We introduce the basic needed definitions: compatibility between a classification structure and a tessellation, (m, k)-networks as a case of tessellation, convex, maximal and connected subsets in such networks, spatial pyramids and spatial hierarchies. Like Robinsonian dissimilarities induced by indexed pyramids generalize ultrametrics induced by indexed hierarchies we show that a new kind of dissimilarities called "Yadidean" induced by Spatial Pyramids generalize Robinsonian dissimilarities. We focus on spatial pyramids where each class is a convex for a grid, and we show that there are several one-to-one correspondences with different kinds of Yadidean dissimilarities. These new results produce also, as a $special$ case, several one to one correspondences between spatial hierarchies (resp. standard indexed pyramids) and Yadidean ultrametrics (resp. Robinsonian) dissimilarities. Qualities of spatial pyramids and their supremum under a given dissimilarity are considered. We give a $constructive$ algorithm for convex spatial pyramids illustrated by an example. We show finally on a simple example that Spatial pyramids on symbolic data can produce a geometrical representation of conceptual lattices of "symbolic objects".

Keywords

PYRAMIDAL CLUSTERING, SPATIAL CLASSIFICATION, SYMBOLIC DATA ANALYSIS, CONCEPTUAL LATTICES, KOHONEN MAPPING

Mining High-Throughput Biological Data: Methods, Algorithms, and Applications

Eytan Domany

Department of Physics of Complex Systems, Weizmann Institute of Science,
Rehovot 76100 Israel

Abstract. DNA chips are novel experimental tools that have revolutionized research in molecular biology and generated considerable excitement. A single chip allows simultaneous measurement of the level at which thousands of genes are expressed. A typical experiment uses a few tens of such chips, each devoted to one sample - such as material extracted from a tumor. Hence the results of such an experiment consist of a table, of several thousand rows (one for each gene) and 50 - 100 columns (one for each sample). Extracting relevant information from such a large, complex and noisy data set requires development of novel methods of analysis.

I will briefly demonstrate how we combine standard statistical analysis with novel unsupervised methods (clustering: Blatt et al., 1996; bi-clustering: Getz et al., 2000; and sorting: Tsafrir et al., 2005) to mine expression data obtained from leukemia (Rozovskaia et al., 2003), cervical (Rosty et al., 2005) and colon cancer (Tsafrir et al., 2005, 2006) patients.

References

BLATT, M., WISEMAN, S. and DOMANY, E. (1996): Superparamagnetic clustering of data. *Physical Review Letters, 76, 3251–3254*.

GETZ, G., LEVINE, E. and DOMANY, E. (2000): Coupled two-way clustering analysis of gene microarray data. *PNAS, 97, 12079–12084*.

ROSTY, C. et al. (2005): Identification of a proliferation gene cluster associated with HPV E6/E7 expression level and viral DNA load in invasive cervical carcinoma. *Oncogene 24, 7094-7104*.

ROZOVSKAIA, T. et al. (2003): Expression profiles of acute lymphoblastic and myeloblastic leukemias with ALL-1 rearrangements. *PNAS 100, 7853–7858*.

TSAFRIR, D., TSAFRIR, I., EIN-DOR, L., ZUK, O. and DOMANY, E. (2005): Sorting points into neighborhoods (SPIN): Data analysis and visualization by ordering distance matrices. *Bioinformatics, 21, 2301–2308*.

TSAFRIR, D. et al (2006): Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Research, 66, 2129–2137*.

Keywords

SUPERPARAMAGNETIC CLUSTERING (SPC), BICLUSTERING (CTWC), SORTING POINTS INTO NEIGHBORHOODS (SPIN), DNA MICROARRAY DATA, CANCER

Some Open Problem Sets for Generalized Blockmodeling

Patrick Doreian

Department of Sociology, University of Pittsburgh,
2406 WWPH, Pittsburgh, PA 15260, USA

Abstract. This paper provides an introduction to the blockmodeling problem of how to cluster networks, based solely on the structural information contained in the relational ties, and a brief overview of generalized blockmodeling as an approach for solving this problem. Following a formal statement of the core of generalized blockmodeling, a listing of the advantages of adopting this approach to partitioning networks is provided. These advantages, together with some of the disadvantages of this approach, in its current state, form the basis for proposing some open problem sets for generalized blockmodeling. Providing solutions to these problem sets will transform generalized blockmodeling into an even more powerful approach for clustering networks of relations.

Keywords

BLOCKMODELING, GENERALIZED BLOCKMODELING, CLUSTER ANALYSIS, OPTIMIZATION, SOCIAL NETWORKS, PARTITIONS

Reframing Author Cocitation Analysis

David Dubin

Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign, USA

Abstract. Author cocitation analysis (ACA) employs a variety of clustering and scaling techniques for exploring patterns in bibliometric data. In a 2003 JASIST article, Per Ahlgren, Bo Jarneving, and Ronald Rousseau criticized the use of Pearson's r as a similarity measure in ACA, citing properties that they believe make r undesirable. Howard White's response to Ahlgren et al, and the subsequent exchange in the JASIST letters column suggest that behind this methodological controversy there lurk contrasting intuitions about both the reality ACA is supposed to reveal and its relationship to the bibliographic evidence on which ACA is based. A reframing of the ACA data analysis methodology puts some of these controversies in a new light.

References

AHLGREN, P. and JARNEVING, B. and ROUSSEAU, R. (2003): Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54/6, 550–560*.

AHLGREN, P. and JARNEVING, B. and ROUSSEAU, R. (2004): Author Cocitation Analysis and Pearson's r. *Journal of the American Society for Information Science and Technology, 55/9, 843*.

MCCAIN, K. W. (1990): Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41/6, 433–443*.

WHITE, H. D. (2003): Author Cocitation Analysis and Pearson's r. *Journal of the American Society for Information Science and Technology, 54/13, 1250–1259*.

WHITE, H. D. (2003): Replies and a Correction. *Journal of the American Society for Information Science and Technology, 55/9, 843–844*.

Keywords

CITATION ANALYSIS, BIBLIOMETRICS, SIMILARITY, INFORMATION SCIENCE

Model Based Two-Mode Partitioning with Dependence Between Row and Column Clusters

Alessio Farcomeni and Maurizio Vichi

Dipartimento di Statistica, Probabilita' e Statistica Applicata
University of Rome "La Sapienza"

Abstract. In this paper a simultaneous clustering of rows and columns of a two-way two mode data matrix is presented. The observed data matrix is supposed divided into sub-matrices such that their union gives back the entire data set. However, blocks are not necessarily defined by the cross-product of row and column clusters, as it is usually done in the literature on two-mode clustering. This allows for each class of a partition of one mode, to obtain a specific partition for the other mode. The proposed clustering model is estimated via maximum likelihood approach by using an EM type algorithm.

In this setting dependence within blocks is taken into account by modeling. Extensive simulations of the proposed algorithm are given. In particular, criteria to select the number of row and column clusters are explored. Furthermore, in order to improve the chance of global maximum convergence and to reduce the number of iterations of the EM algorithm several starting points are compared. Finally, an application in microarray data analysis illustrates the proposed methodology.

References

ROCCI R. and VICHI, M. (2005): Two-mode multi-partitioning. *Computational Statistics & Data Analysis, in revision*.

Van MECHELEN, I., BOCK, H.-H., and De BOECK, P. (2004): Two mode clustering methods: a structured overview. *Statistical Methods in Medical Research, 13, 363–394*.

Keywords

TWO-MODE CLUSTERING, EM, MAXIMUM LIKELIHOOD CLUSTERING

Solving Oblique Bivariate CART

Bernard Fichet, Jean Gaudart, and Bernard Giusiano

Equipe Biomathématiques
Laboratoire d'Informatique Fondamentale
Universités de Marseille
27, Bd Jean Moulin F-13005 Marseille
Name.Surname@medecine.univ-mrs.fr

Abstract. In the usual framework of regression analysis with a sample drawn from d (continuous) predictor variables and a (continuous) response variable, CART procedure is twofold. First (the classification), a binary hierarchy is built on units, via a dividing procedure. For a given cluster, the dichotomy is chosen to maximise the between variance of the response variable. Second (the regression), the admissible dichotomies of a cluster, are those splitting any predictor variable from any threshold. Geometrically, in the d-dimensional predictor space, every dichotomy is associated with a hyperplane orthogonal to one axis. A hierarchy built by this way, is referred to as an axis-parallel decision tree.

Oblique decision trees extend this procedure by considering any hyperplane to split a cluster. Although attractive, such an extension has been proved to be NP-hard. Since then, a great deal of attention has been paid to stochastic algorithms.

However, solving oblique CART in the bivariate case (d=2), is quite easy. We here propose an efficient algorithm. We have to treat a priori infinitely many directions of the predictor plane P. In fact, using continuity arguments, it may be proved that P is divided in a quadratic number of angular sectors, in such a way that only one direction by sector has to be treated. Moreover, the results on one sector are easily deduced from those of a consecutive one. Our algorithm is based upon those properties and explores all sectors in the anticlockwise sense.

Keywords

HIERARCHY, REGRESSION, OBLIQUE DECISION TREE, ALGORITHM

Comparison of Two Methods for Detecting and Correcting Systematic Error in High-Throughput Screening Data

Andrei Gagarin[1], Dmytro Kevorkov[1], Vladimir Makarenkov[2], and Pablo Zentilli[2]

[1] Laboratoire LaCIM, Université du Québec à Montréal,
C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8
[2] Département d'Informatique, Université du Québec à Montréal,
C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8

Abstract. High-throughput screening (HTS) is an efficient technological tool for drug discovery in the modern pharmaceutical industry. It consists of testing thousands of chemical compounds per day to select active ones. This process has many drawbacks that may result in missing a potential drug candidate or in selecting inactive compounds. We describe and compare two statistical methods for correcting systematic errors that may occur during HTS experiments. Namely, the collected HTS measurements and the hit selection procedure are corrected.

References

BRIDEAU, C., GUNTER, B., PIKOUNIS, W. and LIAW, A. (2003): Improved statistical methods for hit selection in high-throughput screening. *Journal of Biomolecular Screening, 8, 634-647*.

ELOWE, N.H., BLANCHARD, J.E., CECHETTO, J.D. and BROWN, E.D. (2005): Experimental screening of dihydrofolate reductase yields a "test set" of 50,000 small molecules for a computational data-mining and docking competition. *Journal of Biomolecular Screening, 10, 653-657*.

HEUER, C., HAENEL, T., PRAUSE, B. (2003): A novel approach for quality control and correction of HTS data based on artificial intelligence. *The Pharmaceutical Discovery & Development Report*. PharmaVentures Ltd. [Online].

KEVORKOV, D. and MAKARENKOV, V. (2005): Statistical analysis of systematic errors in HTS. *Journal of Biomolecular Screening, 10, 557-567*.

MAKARENKOV, V., KEVORKOV, D., GAGARIN, A., ZENTILLI, P., MALO, N. and NADON, R. (2006): An efficient method for the detection and elimination of systematic error in high-throughput screening. Submitted.

ZHANG, J.H., CHUNG, T.D.Y. and OLDENBURG, K.R. (1999): A Simple Statistic Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *Journal of Biomolecular Screening, 4, 67-73*.

Keywords

HIGH-THROUGHPUT SCREENING, SYSTEMATIC ERROR, STATISTICAL ANALYSIS, DATA NORMALIZATION, HIT SELECTION

Clustering Data with Multiple Cluster Structures: A Model-Based Perspective

Giuliano Galimberti and Gabriele Soffritti

Dipartimento di Scienze Statistiche, Università di Bologna,
via delle Belle Arti 41, 40126 Bologna, Italy

Abstract. Nowadays technology advances have made the data collection easier and faster, resulting in larger, more complex data sets with many units and variables. In this situation the choice of the variables that are used to describe units becomes a crucial step in cluster analysis: in fact it is possible that different subsets of variables define different partitions of the same set of units, that is, different cluster structures.

The problem of identifying the cluster structures hidden in a data matrix has been considered only recently in the statistical literature. The proposed solutions are based on very different approaches and in some cases have been conceived for the analysis of specific types of data. Starting from a model-based perspective, in this work some recent methodological developments on this issue are presented.

References

BELITSKAYA-LEVY, I. (2006): A generalized clustering problem with application to DNA microarrays. *Statistical Applications in Genetics and Molecular Biology, 5 (1), Article 2*.

GALIMBERTI, G. and SOFFRITTI, G. (2006): Identifying multiple cluster structures through latent class models. In: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (Eds.): *From Data and Information Analysis to Knowledge Engineering*. Springer-Verlag, Berlin-Heidelberg, 174–181.

GALIMBERTI, G. and SOFFRITTI, G. (2006): How many cluster structures? Answers via a model-based procedure. Submitted.

HASTIE, T., TIBSHIRANI, R., EISEN, M.B., ALIZADEH, A. *et al.* (2000): Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology, 1, 1–21*.

Keywords

CLUSTER ANALYSIS, MIXTURE MODELS, MODEL SELECTION, CLUSTER STRUCTURE

Latent Class Analysis for Financial Data

Attilio Gardini[1], Michele Costa[1], and Stefano Iezzi[2]

[1] Department of Statistical Sciences, University of Bologna,
Via Belle Arti 41, 40126 Bologna, Italy
[2] Economic Research Department, Bank of Italy,
Via Nazionale 91, 00184 Roma, Italy

Abstract. This paper deals with optimal international portfolio choice by developing a latent class approach based on the distinction between international and non-international investors. For thirty years financial literature has been looking for convincing explanations of systematic low international investments: share of foreign assets held by domestic investors is greatly lower than expected by risk-return efficient portfolios: this puzzle is usually called equity home bias.

On the basis of micro data, we analyze the effects of many social, demographic, economic and financial characteristics on the probability to be an international investor. We specify a latent class model which allows to test the existence of two groups of investors: the sub-group of investors who are completely precluded from investment in foreign assets, and the sub-group of investors who are not prevented from investing in foreign assets.

Our results shows how traditional measures of equity home bias are upward biased because they do not allow for the existence of international investment rationing operators. On the contrary, by resorting to latent class analysis it is possible to detect the unobservable distinction between international investors and investors who are precluded from operating into international financial markets and, therefore, to obtain an unbiased measure of equity home bias.

References

BARTHOLOMEW, D.J. and LEUNG, S.O. (2002): A Goodness of Fit for Sparse 2p Contingency Tables. *British Journal of Mathematical and Staistical Psychology, 55, 1–15*.

DEMPSTER, A.P. and LAIRD, N.M. and RUBIN D.B. (1977): Maximum Likelihood from Incomplete Data via EM Algorithm. *Journal of the Royal Statistical Society, Series B, 39:1, 1–38*.

LEWIS, K.K. (1999): Trying to Explain Home Bias in Equities and Consumption. *Journal of Economic Literature, 37, 571–608*.

Keywords

LATENT CLASS ANALYSIS, FINANCIAL DATA, EQUITY HOME BIAS

Sub-Species of *Homopus Areolatus*? Biplots and Small Class Inference with Analysis of Distance

Sugnet Gardner and Niël J le Roux

Department of Statistics and Actuarial Science, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa

Abstract. A canonical variance analysis (CVA) biplot can visually portray a one-way MANOVA. Both techniques are subject to the assumption of equal class covariance matrices. In the application considered, very small sample sizes resulted in some singular class covariance matrix estimates and furthermore it seemed unlikely that the assumption of homogeneity of covariance matrices would hold. Analysis of distance (AOD) is employed as nonparametric inference tool. In particular, AOD biplots are introduced for a visual display of samples and variables, analogous to the CVA biplot.

Keywords

ANALYSIS OF DISTANCE, BIPLOT, CANONICAL VARIATE ANALYSIS, MULTIDIMENSIONAL SCALING, MULTIDIMENSIONAL SCATTERPLOT, MANOVA, PERMUTATION TEST, PRINCIPAL COMPONENT ANALYSIS, PRINCIPAL COORDINATE ANALYSIS, STATISTICAL GRAPHICS

Combining Classifiers of Different Types

Eugeniusz Gatnar

Institute of Statistics, Katowice University of Economics,
ul. Bogucicka 14, 40-226 Katowice, Poland

Abstract. Model fusion has proved to be a very successful strategy to obtain accurate prediction models. The key issue, however, as Tumer and Ghosh (1996) have shown, is the diversity of the component classifiers because classification error of an ensemble depends on the correlation between its members. Simply, the more different the combined models are, the more accurate the ensemble is.

Classifier fusion consists of two steps: first, a set of independent models is formed, and then their outputs are aggregated into an ensemble. Existing methods, e.g. Bagging, Boosting or RandomForest and their variants differ in the way the component models are built, and the way their outputs are combined.

The majority of existing ensemble methods, e.g. RandomForest developed by Breiman (2001), combine the same type of models (trees) in different feature spaces. In order to promote the diversity of the ensemble members, we propose to apply classifiers of different types, because they can partition the same feature space in very different ways (e.g. trees and SVM).

In our experiments we have formed a number of different classifiers, i.e. linear and quadratic classifiers, trees, neural networks, SVM models and nearest neighbors, with various sets of their parameters, and then combined them using majority voting. The obtained results showed that ensembles built with the proposed method outperformed those built for different feature spaces.

References

BREIMAN, L. (2001): Random Forests. *Machine Learning 45, 5–32*.

GATNAR, E. (2002): Tree-based models in statistics: three decades of research. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Analysis*. Springer, Berlin, 399–408.

GATNAR, E. (2005a): Dimensionality of Random Subspaces. In: C. Weihs and W. Gaul (Eds.): *Classification - The Ubiquitous Challenge*. Springer, Heidelberg, 129–136.

GATNAR, E. (2005b): A Diversity Measure for Tree-Based Classifier Ensembles. In: D. Baier, R. Decker, and L. Schmidt-Thieme (Eds.): *Data Analysis and Decision Support*. Springer, Heidelberg, 30–38.

TUMER, K. and GHOSH, J. (1996): Analysis of decision boundaries in linearly combined neural classifiers, *Pattern Recognition, 29, 341–348*.

Keywords

CLASSIFICATION, ENSEMBLE METHODS, CLASSIFIER FUSION

On the Eigensystems of Operational Accounting Systems

Andreas Geyer-Schulz and Bettina Hoser

Institut für Informationswirtschaft und -management,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Abstract. In this contribution we show how the Eigensystem of an operational accounting system can be derived, analyzed, and interpreted. For this purpose transactions in the accounting system are represented as the flow on an asymmetric directed graph which is transformed into a complex Hermitian adjacency matrix whose Eigensystem is computed and analyzed. We discuss how shifts in structure and interesting behavioral patterns can be identified as well as the informational implications of accounting conventions on the Eigensystem. The method is suitable for accounting systems which are as part of AAA-systems embedded in technical infrastructure as e.g. telecommunication or energy networks.

References

HOSER, B. and GEYER-SCHULZ, A. (2005): Eigenspectralanalysis of Hermitian Adjacency Matrices for the Analysis of Group Substructures. *Journal of Mathematical Sociology, 29(4), 265 - 294*.

FRANKE, M., GEYER-SCHULZ, A., HOSER, B. (2006): On the Analysis of Asymmetric Directed Communication Structures in Electronic Election Markets, 37 - 59. In: BILLARI et al. (2006): Agent-Based Computational Modelling: Applications in Demography, Social, Economic and Environmental Sciences, Physica, Heidelberg.

Keywords

FINANCIAL ACCOUNTING, EIGENSYSTEMS ANALYSIS

Revised Boxplot Based Discretization as the Kernel of Automatic Interpretation of Classes Using Numerical Variables

Karina Gibert[1] and Alejandra Pérez-Bonilla[2]

[1] Department Statistics and Operations Research.
Technical University of Catalonia.
Campus Nord, Edif. C5, C\ Jordi Girona, 1-3, 08034 Barcelona, SPAIN. karina.gibert@upc.edu
[2] Department Statistics and Operations Research.
Technical University of Catalonia.
alejandra.perez@upc.edu –Intern CONICYT, Chile.
symbolic data analysis

Abstract. In this paper the impact of improving *Boxplot based discretization (BbD)* on the methodology of *Boxplot based induction rules (BbIR)*, oriented to the automatic generation of conceptual descriptions of classifications that can support later decision-making is presented. A particular application to Waste Water Treatment Plants (WWTP) is in progress and results appear to be very promising, see Gibert and Roda (2000). *Boxplot based induction rules (BbIR)*, see Gibert and Pérez-Bonilla (2006), is a proposal to produce compact concepts associated to the classes, oriented to express the differential characteristic of every class in such a way that the user can easily understand which is the underlying classification criterion and can easily decide the treatment or action to be assigned to each class. Given a classification, the idea is to provide an automatic interpretation for it that supports the construction of intelligent decision support systems. The core of this process is the *BbD* method which is denoted to discretize numerical variables in such a way that particularities of the classes are elicited automatically analysis of conditional distributions, *Multiple boxplots*(Tukey (1977)), and it successfully the interpretation process used by most experts in they real.

References

GIBERT, K and PÉREZ-BONILLA, A. (2006). Automatic generation of interpretation as a tool for modelling decisions. In: *III International Conference on Modeling Decisions for Artificial Intelligence*, Tarragona, in press.

GIBERT, K. and RODA, I. (2000). Identifying characteristic situations in wastewater treatment plants. In: *Workshop BESAI (ECAI)*, **1**, 1-9.

TUKEY, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Keywords

CLUSTERING, DATA MINING, CLASS INTERPRETATION, CONDITIONED DISTRIBUTIONS, MULTIPLE BOXPLOT, DISCRETIZATION, CHARACTERIZING VARIABLE, INDUCTION RULES, CONCEPTS, VALIDATION, SYMBOLIC DATA ANALYSIS

Similarity in Retrospect

John C. Gower

Statistics Department
The Open University
Milton Keynes, MK7 6AA, U. K

Abstract. It is nearly 50 years since I first became interested in measures of similarity. At that time there were several questions that concerned me. A few hundred similarity coefficients later, most still do. The problem is not so much what is the best measure to use, but what is our coefficient intended to measure. Some questions that deserve consideration are:

i. What features should be included in a measure of similarity?
ii. Is there any sense in which we may regard feature choice as a form of sampling and, if so, from what population are we sampling?
iii. Should we allow for feature weighting and, if so, under what circumstances?
iv. What place do stochastic considerations have in the choice/development of measures of similarity?
v. What objects are being compared?
vi. Are objects unique or should they be regarded as samples from notional populations of objects?
vii. How are populations of objects defined?
viii. Can similarity coefficients be classified in some useful way?

I do not believe that any of these questions have simple answers but in some circumstances we may give partial answers. I shall try to address these and similar problems.

Keywords

SIMILARITY COEFFICIENTS, SAMPLING CASES, SAMPLING FEATURES, STOCHASTIC CLASSIFICATION, NON-STOCHASTIC CLASSIFICATION

kNN Versus SVM in the Collaborative Filtering Framework

Miha Grčar, Blaž Fortuna, Dunja Mladenič, and Marko Grobelnik

Jožef Stefan Institute,
Jamova 39, SI-1000 Ljubljana, Slovenia

Abstract. We present experimental results of confronting the k-Nearest Neighbor (kNN) algorithm with Support Vector Machine (SVM) in the collaborative filtering framework using datasets with different properties. While k-Nearest Neighbor is usually used for the collaborative filtering tasks, Support Vector Machine is considered a state-of-the-art classification algorithm. Since collaborative filtering can also be interpreted as a classification/regression task, virtually any supervised learning algorithm (such as SVM) can also be applied. Experiments were performed on two standard, publicly available datasets and, on the other hand, on a real-life corporate dataset that does not fit the profile of ideal data for collaborative filtering.

We conclude that on our datasets, kNN is dominant on the Jester dataset that has relatively low sparsity. On the two datasets with high to extremely high level of sparsity (EachMovie, the corporate dataset), kNN starts failing as it is unable to form reliable neighborhoods. In such case it is best to use a model-based approach, such as SVM classifier or SVM regression. Another strong argument for using the SVM approaches on highly sparse data is the ability to predict more ratings than with the variants of the memory-based approach.

References

BILLSUS, D., and PAZZANI, M. J. (1998): Learning Collaborative Information Filers. In: *Proceedings of the Fifteenth International Conference on Machine Learning*.

BREESE, J.S., HECKERMAN, D., and KADIE, C. (1998): Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*.

GRCAR, M., MLADENIC D., GROBELNIK, M. (2005): Applying Collaborative Filtering to Real-life Corporate Data. In: *Proceedings of the 29th Annual Conference of the German Classification Society (GfKl 2005)*, Springer, 2005.

HERLOCKER, J.L., KONSTAN, J.A., TERVEEN, L.G., and RIEDL, J.T. (2004): Evaluating Collaborative Filtering Recommender Systems. In: *ACM Transactions on Information Systems, Vol. 22, No. 1, 5–53*.

Keywords

USER PROFILING, COLLABORATIVE FILTERING, SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBORS, MACHINE LEARNING, WEB LOG MINING

Multidimensional Scaling of Histogram Dissimilarities

P.J.F. Groenen[1] and S. Winsberg[2]

[1] Econometrics Institute, Erasmus University, Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
email: groenen@few.eur.nl
[2] Predisoft, San Pedro, Costa Rica
email: SuzanneWinsberg@predisoft.com

Abstract. Multidimensional scaling aims at reconstructing dissimilarities between pairs of objects by distances in a low dimensional space. However, in some cases the dissimilarity itself is unknown, but the range, or a histogram of the dissimilarities is given. This type of data fall in the wider class of symbolic data (see Bock and Diday (2000)). We model a histogram of dissimilarities by a histogram of the distances defined as the minimum and maximum distance between two sets of embedded rectangles representing the objects. In this paper, we provide a new algorithm called Hist-Scal using iterative majorization, that is based on an algorithm, I-Scal developed for the case where the dissimilarities are given by a range of values ie an interval (see Groenen et al. (in press)). The advantage of iterative majorization is that each iteration is guaranteed to improve the solution until no improvement is possible. We present the results on an empirical data set on synthetic musical tones.

References

BOCK, H.H. and DIDAY, E. (1974): *Analysis of symbolic Data*. Springer, Berlin.

Keywords

MULTIDIMENSIONAL SCALING, HISTOGRAM DATA, SYMBOLIC DATA ANALYSIS, ITERATIVE MAJORIZATION, DISTANCE SMOOTHING, HIST-SCAL

Multigroup SVM through an Optimal Scaling Approach

Patrick J.F. Groenen, Georgi Nalbantov, and Cor Bioch

Econometric Institute, Erasmus University Rotterdam,
Rotterdam, The Netherlands

Abstract. Support vector machines (SVM) belong to the best prediction methods available for the two-group classification problem. It turns out that the SVM has a close connection to optimal scaling regression with the dependent variable having two groups and allowing the tied data within one group to be untied. This transformation is often called the primary approach to ties in the optimal scaling literature.

However, the classical SVM has no natural extension to the situation of more than one group. By using the optimal scaling approach, we propose an extension for the multigroup classification problem by using an ordinal transformation and untieing the ties. Because usually the grouping has no natural order, all different orderings of the groups are tried and the best retained. Therefore, in practical situations the number of groups cannot be too large. In our presentation, we show how this new method works and apply it to a multigroup classification problem.

Keywords

SUPPORT VECTOR MACHINES, OPTIMAL SCALING REGRESSION, MULTIGROUP CLASSIFICATION PROBLEM

Properties of Minimum Rigidity Graphs Associated with a Clustering System

Gentian Gusho[1 2]

[1] Département Logique des Usages, Sciences Sociales et de l'Information,
École Nationale Supérieure des Télécommunications de Bretagne,
Technopôle de Brest-Iroise - CS 83818,
29238 Brest cedex,
FRANCE

[2] TAMCIC,
U.M.R. CNRS 2872
(ENST Bretagne)
Gentian.Gusho@enst-bretagne.fr

Abstract. In this paper, we study a new approach in classification, *the realizations of a dissimilarity* (Brucker 2003) which is a binary clustering system (Barthélemy 2003) extracted from a dissimilarity. The realizations which represent the data as it is, provide a broader representation and interpretation of dissimilarity information compared to classical models. Finally, by using graph theory and, more precisely, the concept of the *minimum rigidity graph* associated to a clustering system, we show the relation between realizations of a dissimilarity and the indexed hierarchy induced by the associated sub-dominant ultrametric (Leclerc 1981).

References

BRUCKER, F (2003): Réalisations de dissimilarité. *Actes des Rencontres de la Société Francophone de Classification*, 7–10.

BARTHELEMY, J.-P. (2003): Classification binaire. *Actes des Rencontres de la Société Francophone de Classification*, 67-69.

LECLERC, B. (1981): Déscription combinatoire des ultramétriques. *Mathématiques et Sciences Humaines, 73, 5–37*.

Keywords

DISSIMILARITY, CLUSTERING SYSTEM, RIGIDITY, GRAPH

The Length and Breadth of Classification Science: The Case Study of Banking Fraud

David J. Hand

Department of Mathematics, Imperial College London
London SW7 2AZ, United Kingdom

Abstract. Classification is a fundamental concept.ă It lies at the very basis of human thought, when we group objects and events into classes so that we can process them effectively.ă It structures human relationships, when we allocate people into groups and their behaviour into types so that we know how to interact.ă It forms the foundations of science, when we construct theories from observations of naturally occurring regularities.ă This talk illustrates some of the ubiquity of classification concepts by looking at the case study of fraud detection in banking.ă In this situation, the aim is to classify each transaction, each account, and each account holder as fraudulent or non-fraudulent.ă Many classification tools have been applied to this problem, including both supervised and unsupervised methods, single and multiple class methods, and ranging from the earliest classification methods to be developed right through to the most recent innovations.

Keywords

CLASSIFICATION, FRAUD, BANKING

Some Applications of Combinatorial Methods Used for Cluster Verification

Bernard Harris

Department of Statistics, University of Wisconsin
1300 University Avenue, Madison, Wisconsin 53706, USA

Abstract. N independent identically distributed k-dimensional continuous random variables have been observed. For data satisfying this condition and satisfying certain regularity conditions, tests using graph theoretic methodology have been introduced for cluster verification. These tests are used for detection of spurious clusters, that is, clusters which are determined by some algorithm, but are simply a result of random variation. A summary of these methods is presented.

Two applications of these methods are discussed. In the first of these, we discuss a urn model for contamination of semiconductor materials by boron atoms. The same type of proximity analyses can be applied in this case. Basically, if two many contaminants are too close together, which is similar to the generation of clusters, then electronic equipment using these materials will not function satisfactorily.

The second application is to use these methods for detection of non-trivial clusters. This will be applied to detection of mixtures of normal distributions.

Keywords

CLUSTER VERIFICATION, BORON COMTAMINATION OF SEMICONDUCTOR MATERIALS, IDENTIFICATION OF MIXTURES

Network Representations of City-Block Models

Willem J. Heiser[1] and Laurence E. Frank[2]

[1] Department of Psychology,
Leiden University, P.O. Box 9555,
2300 RB Leiden, The Netherlands
[2] Department of Methods and Statistics,
Faculty of Social and Behavioral Sciences,
Utrecht University, P.O. Box 80140,
3508 TC Utrecht, The Netherlands

Abstract. City-block models for similarity always allow network representations that reproduce the same distances as the unique coordinate representation. A rule to construct such networks is given, based on additivity of city-block distances across sequences of intermediate points along monotonic trajectories in space. The paper also defines the concept of internal node, which helps in reducing the complexity of networks and in making them better interpretable. The general graph construction rule and definition of internal nodes also apply to the distinctive features model, the common features model (additive clustering), as well as to hierarchical trees, additive trees, and extended trees. Additivity is the key property that makes the city-block metric so versatile and causes a basic unity of dimensional, hierarchical and featural representations of similarity.

Keywords

MULTIDIMENSIONAL SCALING (MDS), ADDITIVE CLUSTERING, FEATURE MODELS, METRIC SEGMENT, PARTIAL ISOMETRY

Design of Dissimilarity Measures: A New Dissimilarity between Species Distribution Areas

Christian Hennig[1] and Bernhard Hausdorf[2]

[1] Department of Statistical Science, University College London
Gower St, London WC1E 6BT, United Kingdom
[2] Zoologisches Museum der Universität Hamburg
Martin-Luther-King-Platz 3, 20146 Hamburg, Germany

Abstract. In many situations, dissimilarities between objects cannot be measured directly, but have to be constructed from some known characteristics of the objects of interest, e.g. some values on certain variables.

From a philosophical point of view, the assumption of the objective existence of a "true" but not directly observable dissimilarity value between two objects is highly questionable. Therefore we treat the dissimilarity construction problem as a problem of the choice or design of such a measure and not as an estimation problem of some existing but unknown quantities.

Therefore, subjective judgment is necessarily involved, and the main aim of the design of a dissimilarity measure is the proper representation of a subjective or intersubjective concept (usually of subject-matter experts) of similarity or dissimilarity between the objects.

We give some guidelines for the choice and design of dissimilarity measures and illustrate some of them by the construction of a new dissimilarity measure between species distribution areas in biogeography, the so-called "geco coefficient". Species distribution data can be digitized as presences and absences in certain geographic units. As opposed to all measures already present in the literature, the geco coefficient introduced in the present paper takes the geographic distance between the units into account. The advantages of the new measure are illustrated by a study of the sensitivity against incomplete sampling and changes in the definition of the geographic units in two real data sets.

Keywords

BIOGEOGRAPHY, GECO COEFFICIENT, STABILITY, SUBJECTIVE DECISIONS

Computing Summaries of Time Series Databases with Clustering and Segmentation

Bernard Hugueney[1], Georges Hébrail[2], and Yves Lechevallier[3]

[1] Université PARIS-DAUPHINE
LAMSADE
Place du Maréchal de Lattre de Tassigny
75775 PARIS CEDEX 16
bernard.hugueney@lamsade.dauphine.fr,
http://www.lamsade.dauphine.fr/~hugueney
[2] GET-ENST Paris, Laboratoire LTCI - UMR 5141 CNRS
Département Informatique et Réseaux
46, Rue Barrault, 75634 Paris Cedex 13
georges.hebrail@enst.fr
[3] INRIA - Rocquencourt,
Domaine de Voluceau - Rocquencourt - B. P. 105
78153 Le Chesnay Cedex - France
Yves.Lechevallier@inria.fr

Abstract. With the evergrowing use of sensors for monitoring purposes and information systems usage logging, many databases are filled with very large amounts of temporal data. Exploratory analysis of those time series is therefore requiered in order to summarize the available data. There are two main approaches to the summarization of a set of time series. One is to represent the large number of time series by clustering them into a small number of homogeneous groups. The other is to segment the large number of data points of the time series into a smaller number of episodes, representing each of the time series by a piecewise constant model. What we propose is to create summaries of a set of time series though segmentation of clusters of time series. The whole data set can then be represented by a small number of piecewise constant models. The only parameters are the number of clusters and the total number of episodes. Using dynamic programming, we perform an optimal distribution of episodes amongst the clusters. This approach has been applied to both benchmark and real-world data-sets.

Keywords

DATA REPRESENTATION, TIME SERIES DATABASES, CLUSTERING, SEGMENTATION, DYNAMIC PROGRAMMING

A Method for Local Representation the Asymmetric Similarity Data Matrix

Tadashi Imaizumi

Department of Management & Information Sciences, Tama University,
4-1-1 Hijirigaoka, Tama-city, Tokyo, Japan, 206-0022

Abstract. In data analysis of similarity data matrix whose cell represents the degree of relationship between two objects, MDS(MultiDimensonal Scaling) or Cluster Analysis methods have been widely used.
As the number of objects is increasing, some model must be introduced since the overall representation of objects onto the common space is less reasonable. In these case, models and methods for lower dimensional representation whose minimize the loss function,

$$Loss(S, X) = \Sigma_{i,j}^{n} w_{ij}(\delta_{ij} - dist_{ij})^2 \quad (1)$$

will be useful. In this loss function, δ_{ij} is an observed dissimilarity between object o_i and o_j,$dist_{ij}$ is the corresponding distance and w_{ij} is a weighting function. the most simple weighting function is
if $\delta_{ij} < \delta_{cut}$, then $w_{ij} = 1$, else $w_{ij} = 0$.
We want to extract a local relationship objects as a cluster. And the triple of objects is a reasonable smallest cluster. And this indicates a principle to choose d_{cut} as maximizing the number of triple of which all distances δ_{ij}, δ_{jk} and δ_{ik} is smaller than δ_{cut}.
In this case, we need to check the average distance within a cluster of triple since δ_{cut} will be determined the number of triple only.
On the other hand, the asymmetric relationship of data matrix have useful information of objects. This suggests us that these cluster of triple should be defined by using six δs. In this talk, I will propose the the method for choosing the parameter δ_{cut} when we analyze the asymmetric similarity data matrix with Okada and Imaizumi's asymmetric MDS.

$$Loss(S, X)Loss(S, X) = \Sigma_{i,j}^{n} w_{ij}(\hat{m}_{ij} - m_{ij})^2 \quad (2)$$

where, $m_{ij} = dist_{ij} - r_i + r_j$.

References

OKADA, A and IMAIZUMI, T(1987):Nonmetric multidimensional scaling of asymmetric proximities.*Behaviormetrika, 21ĄD81–96*

Keywords

ASYMMETRIC SIMILARITY DATA, CLUSTER, TRIPLE

A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data

Antonio Irpino and Rosanna Verde

Facoltá di Studi Politici e per l'Alta Formazione Europea e Mediterranea
"Jean Monnet", Seconda Universitá degli Studi di Napoli,
Caserta, I-81020, Italy

Abstract. Symbolic Data Analysis (SDA) aims to to describe and analyze complex and structured data extracted, for example, from large databases. Such data, which can be expressed as concepts, are modeled by symbolic objects described by multivalued variables. In the present paper we present a new distance, based on the Wasserstein metric, in order to cluster a set of data described by distributions with finite continue support, or, as called in SDA, by "histograms". The proposed distance permits us to define a measure of inertia of data with respect to a barycenter that satisfies the Huygens theorem of decomposition of inertia. We propose to use this measure for an agglomerative hierarchical clustering of histogram data based on the Ward criterion. An application to real data validates the procedure.

References

AITCHISON, J. (1986): The Statistical Analysis of Compositional Data, New York: Chapman Hall.

BOCK, H.H. and DIDAY, E. (2000): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.

BILLARD, L., DIDAY, E. (2003): From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis *Journal of the American Statistical Association, 98, 462, 470-487*.

GIBBS, A.L. and SU, F.E. (2002): On choosing and bounding probability metrics, *International Statistical Review, 70, 419*.

IRPINO, A. and VERDE, R.(2005): A New Distance for Symbolic Data Clustering, *CLADAG 2005, Book of short papers, MUP, 393-396*.

MALLOWS, C. L. (1972): A note on asymptotic joint normality. *Annals of Mathematical Statistics, 43(2),508-515*.

WARD, J.H. (1963):Hierarchical Grouping to Optimize an Objective Function,*Journal of the American Statistical Association, vol. 58, 238-244*.

Keywords

HISTOGRAM DATA, WASSERSTEIN DISTANCE, EMPIRICAL DISTRIBUTION, MALLOW'S DISTANCE

Copula Methods in Financial Data Analysis

Krzysztof Jajuga

Department of Financial Investments and Risk Management,
Wrocław University of Economics, Ul. Komandorska 118-120, 53-345 Wroclaw

Abstract. As a rule, financial data analysis is performed by multivariate statistical and econometric methods. The paper presents the use of copula methods in analysis of financial data, especially in the form of financial time series. Copula analysis consists in separate analysis of marginal distributions and the dependence between variables.

The paper discusses possible use of copula methods in the following financial problems: market risk measurement, credit risk measurement, extreme value analysis and tail dependence.

In addition to theoretical survey, the paper presents some empirical examples.

Keywords

COPULA FUNCTION, EXTREME VALUE ANALYSIS, FINANCIAL RISK, TAIL DEPENDENCE

Classification of Slovenian Words from a Search Engine Index

Primož Jakopin

Corpus Laboratory, Fran Ramovš Institute of Slovenian Language ZRC SAZU
Ljubljana, Slovenia
primoz.jakopin@guest.arnes.si

Abstract. In the paper some problems that surfaced during the preparation of a list of Slovenian words (with POS data, at http://bos.zrc-sazu.si/besede.html), are described. The project started in 2004 with words from two non-overlapping monolingual dictionaries of Slovenian (93.000 and 178.000 headwords), both also available on the Internet; in 2005 it was expanded by 50.000 new words from an in-house 160 mil. word text corpus (http://bos.zrc-sazu.si/a_beseda.html), composed mainly of Slovenian newspaper, transcribed formal-speech and fiction texts.

The index of the leading Slovenian web search engine, NAJDI.SI, was used as the third source. It consists of (February 2006) 58.957.000 different tokens, of which 17.853.000 are letters-only, i.e. words (or better wordforms) in the usual sense; 13.900.000 have remained after conversion to lower-case. After the intersection with wordform list (3.945.000 different tokens), obtained from words and open class derivatives of the previously assembled 321.000-word list, 12.922.000 tokens have remained, candidates for new words.

The language is very inflected, a typical open class word lemma has 11 different derived wordforms, and there are lemmas with close to 100 derivatives, such as the adjective *star* (*old* in English, 85 wordforms). So quite often clusters of wordforms, connected to the same lemma, are to be found: *medijskoaktivistično*, *medijskodidaktični*, *medijskodidaktično*, *medijskofilozofska*, *medijskoizobraževalnih*, *medijskokritično*, *medijskonačrtovalna*, *medijskoodmevna*, *medijskoodmevni*, *medijskopedagoškim*, *medijskopopulistična*, *medijskopromocijski*, *medijskospecifične*, *medijskospecifičnih*, *medijskotehnični*, *medijskoteoretične* and *medijskozgodovinski*. In the main part of the paper the algorithm for lemma identification, based on n-gram probabilities for Slovenian, known prefixes and, before all, suffixes from inflectional paradigms.

Keywords

WORD STATISTICS, LEXICON, SLOVENIAN LANGUAGE, ENTROPY, LEMMATISATION

Dissimilarities Taking Values in a Poset

Melvin F. Janowitz

DIMACS
Rutgers University
96 Frelinghuysen Road
Piscataway, NJ 08854, USA

Abstract. When clustering is based on data that is produced by repeated trials or by measurements, it is natural to view each attribute as being a sample of a probability distribution. If one wants to cluster before simplifying the data, it is natural to define a dissimilarity coefficient (DC) taking values in the space of percentile functions. This was investigated in Janowitz and Schweizer (1989). There are other situations where it is natural for a DC to take values in a poset. But the possibilities for clustering functions are limited unless a modified view is given of the nature of a dissimilarity coefficient.

It will be argued that the proper setting is as follows: Given a finite nonempty set E and a poset L, a DC is a mapping d from ordered pairs of E to the order filters of L having the property that $d(x, y) = d(y, x)$ for all all $x, y \in E$. One does not assume even that $d(x, x) = 0$ for all x. The idea is that $d(x, y)$ picks out $\{h \in P\}$ at which the pair (x, y) is a candidate for clustering. These values naturally form an order filter of L. This idea allows ordinal clustering methods to be implemented in the given setting. It also provides a common framework in which both cluster analysis and formal concept analysis may coexist.

References

JANOWITZ, M.F. and SCHWEIZER, B. (1989): Ordinal and percentile clustering. *Mathematical Social Sciences, 18, 135-186*.

Keywords

DISSIMILARITY COEFFICIENT, POSET, FILTER, FORMAL CONCEPT ANALYSIS

Comparison of Teaching Mathematics: An application of Adapted Leaders Clustering Method

Barbara Japelj Pavešič

Educational Research Institute,
Gerbičeva 62, SI-1000 Ljubljana, Slovenia

Abstract. Mathematics as a science is culturally independent, but what about teaching mathematics? In an attempt to answer this question we have compared data on learning and teaching of mathematics collected in the Trends in International Mathematics and Science Study (TIMSS) in 1999 and 2003. These studies included about 60 countries, thousands of teachers and over a million students. The collected data can be organized in an ego-centered network with teachers as egos and their students as alters. We applied the Adapted Leaders Clustering Method (ALCM) to find similarities and differences between teachers over the world. When we first used ALCM on TIMSS 1999 data from 37 countries, we looked for similarities between teachers according to their teaching methods and relate discovered characteristics of teacher clusters to the achievement of students. We discovered that the resulting clusters of teachers have interesting common tracts depending on their regional origin and language background, which clearly indicated that the teaching of mathematics is not independent of cultural background.

We are going to present results of clustering of teachers from the latest TIMSS 2003 study when more countries were involved, a comparison to results from 1999, and to a result of the application of ALCM to a single country (Slovenia).

References

BOCK, H.H. and DIDAY, E. (2000): Symbolic Objects. In: H.-H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg.

HARTIGAN, J.A. (1975): *Clustering Algorithms*. Wiley, New York.

KORENJAK-ČERNE, S. (2002): Symbolic data analysis approach to clustering large ego-centered networks. In: Mrvar, A., Ferligoj, A.(eds.).*Developments in statistics*, Metodološki zvezki 17. 45–53. FDV, Ljubljana.

KORENJAK-ČERNE, S., BATAGELJ, V. (2002): Symbolic data analysis approach to clustering large datasets. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Data Analysis*. Springer, Berlin, 319–327.

KORENJAK-ČERNE, S., (2002): The program package CLAMIX.

Keywords

EGO-CENTERED NETWORKS, SYMBOLIC DATA, CLUSTERING, MEASUREMENT IN EDUCATION

Comparing What We Have Rather Than Developing Something New?

Henk A.L. Kiers

Heymans Institute, University of Groningen,
Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

Abstract. The majority of papers on data analysis that I encounter introduce new or refined methods for particular data analysis problems, and often compare the new developed method to one or a few existing ones. As a result, we have an enormous mass of techniques with all sorts of special features, but many of them, I fear, will never be used by anybody else than the proposer. Besides, a practical data analyst has enormous difficulty in deciding whether the new technique (s)he reads about should be used or not, because the comparison is made only to one or a few techniques, and not to all available techniques. For some problems, explicit comparisons have been made. E.g., for the simple classification problem, Hand (2004) indicated that in practice simple methods typically work just as well as more sophisticated methods.

Should we continue developing new methods, with new features in which indeed the new method has an edge compared to other methods, at least for certain kinds of data? Or should we focus more on global systematic comparisons of existing methods? To answer this, we should first consider what should be the criteria with respect to which methods are compared. Related to this, we should consider carefully how data on which comparisons are based, are to be chosen or constructed. There are no general answers to this, but I will discuss considerations on these issues, thus aiming to offer a start of a more general discussion on the topic.

References

HAND, D.J. (2004): Academic obsessions and classification realities: ignoring practicalities in supervised classification. In: D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds.): *Classification, Clustering and Data Mining Applications*. Springer, Berlin.

Keywords

COMPARISON OF TECHNIQUES, SIMPLICITY

Anti-Robinson Structures for Analyzing Three-Way Two-Mode Data

Hans-Friedrich Köhn

Department of Psychology, University of Illinois at Urbana-Champaign,
603 East Daniel St., Champaign, IL 61820, USA

Abstract. The decomposition of a square-symmetric proximity matrix into a sum of equally-sized matrices, restricted only to display a specific ordinal patterning, the (anti-)Robinson form, can be regarded as a non-metric analogue to spectral decomposition for identifying an ordinal low-rank approximation to a matrix. Due to their particular patterning, (anti-)Robinson matrix components lend themselves immediately to the further representation through secondary structures, either in the form of a (discrete non-spatial) ultrametric dendrogram or a (continuous spatial) unidimensional scale. As both structures are estimated through least-squares, a direct comparison of their fit is legitimate to decide whether the relationship between row and column objects, can be better represented through a discrete or a continuous model. We propose an extension of the ordinal square-symmetric matrix decomposition to the analysis of structural individual differences for three-way two-mode data. The task of modelling individual differences is addressed within a deviation-from-the-mean framework: based on the aggregated body of individual proximity data, an optimal ordinal matrix decomposition is identified that in a second analysis step is used as a frame of reference for the decomposition of the individual proximity matrices. For the reference as well as the individual decompositions, unidimensional scales and ultrametric tree representations can be constructed. An application to judgments of schematic face stimuli illustrates the method.

Keywords

THREE-WAY TWO-MODE PROXIMITY MATRICES, ANTI-ROBINSON FORM, ORDER CONSTRAINTS, QUADRATIC ASSIGNMENT, ITERATIVE PROJECTION, LEAST-SQUARES MATRIX APPROXIMATION, UNIDIMENSIONAL SCALING, ULTRAMETRIC TREE STRUCTURES

About Strain Force in a Capital and Robust Analysis of Planar Shape

Daniel Kosiorowski

Department of Statistics, Cracow University of Economics,
Rakowicka 25, 31-510 Cracow, Poland

Abstract. In this paper we present a statistical approach appealing to a statistical analysis of shape that allows us give some arguments for a theoretic concept according to which a capital stored in a certain system is described by the ability of this system to perturb a certain space of values, the flow of the capital is connected with intermolecular stresses in the substance of the capital. Within the statistical theory of shape, the shape of an object belonging to a certain class of objects is considered based on the concept of a landmark – a point which is characteristic of all the objects of the class under consideration and which corresponds with a certain specific substantive or mathematical feature of the objects. We use interpretation according to which the notion of an average shape could correspond to an stress function and a variability of shape could correspond to an amount of energy stored in the economic system. Relatively well known and appropriate for analysis of capital flows between economic sectors is planar analysis of shape. In this case complex arithmetic leads to a significant simplification. Namely among others we can introduce a distance between planar shapes on base of linear complex regression model. Stochastic models used within statistical theory of shape are too restrictive for economic, financial applications. Commonly used estimators of the average shape and the variability of shape are not robust. That facts motivate us to propose two robust modifications of Procrustes analysis referring to data depth concept. We propose a projection depth trimming of the observations with outlying residuals and a rank test of equity of two average shapes that uses an induced by projection depth ranking of tangent coordinates of shape. Theoretical considerations are illustrated by results of empirical studies concerning relation between stresses in the in the substance of the capital and flows of the capital between sectors of polish economy. Studies were conducted on base of data from five polish sector sub–indices during a one year period. Stocks of indices were treated as representative particles of the capital and were considered with respect to a daily increase and volume.

References

DYCKERHOFF, R. (2004): Data Depths Satisfying the Projection Property. *Allgemeines Statistisches Archiv , 88, 163-190*.

KENT, J. T. (1994): The Complex Bingham Distribution and Shape Analysis. *Journal of the Royal Statistical Society, Series B , 56, 285-299*.

Keywords

AVERAGE SHAPE, PROJECTION DEPTH, CAPITAL FLOW

A Statistical Framework for Assessing the Congruence Among Overlapping Trees and Detecting Common Spatial Patterns in Comparative Phylogeography

François-Joseph Lapointe

Département de sciences biologiques, Université de Montréal,
C.P. 6128, Succ. Centre-ville, Montréal, Québec, H3C 3J7, Canada

Abstract. Comparative phylogeography offers an interesting framework for studying and understanding the spatial patterns of genetic divergence of codistributed species. However, the field has been hampered by a lack of statistical rigor. Recently, Lapointe and Rissler (2005) have proposed a method for testing the congruence across multiple species and detect common geographical signals in the data. In the present paper, a statistical method based on Maximum Agreement Subtrees (Finden and Gordon 1985) will be presented for comparing trees defined on partially-overlapping sets of leaves. The distance-based multiple regression method of Legendre et al. (1994) will also be used to test for the presence of common geographical breaks in the genetic data. An application to the comparative phylogeography of codistributed species in California will be presented to demonstrate the use of these statistical approaches.

References

FINDEN, C. R. and GORDON, A. D. (1985): Obtaining Common Pruned Trees. *Journal of Classification, 2, 225–276.*

LAPOINTE, F.-J. and RISSLER, L. J. (2005):
Consensus, Congruence, and the Comparative Phylogeography of Codistributed Species in California.
The American Naturalist, 166, 290–299.

LEGENDRE, P. and LAPOINTE, F.-J. (1994): Modeling Brain Evolution from Behavior. *Evolution, 48, 1487–1499.*

Keywords

COMPARATIVE PHYLOGEOGRAPHY, CONGRUENCE, MAST, MULTIPLE REGRESSION, STATISTICAL TEST

Dependence and Interdependence Analysis for Interval-Valued Variables

Carlo Lauro[1] and Federica Gioia[2]

[1] Dipartimento di Matematica e Statistica
Università di Napoli Federico II
Complesso Universitario di Monte S. Angelo, Via Cinthia
I-80126 Napoli, Italy
[2] Dipartimento di Matematica e Statistica
Università di Napoli Federico II
Complesso Universitario di Monte S. Angelo, Via Cinthia
I-80126 Napoli, Italy

Abstract. Data analysis is often affected by different types of errors as: measurement errors, computation errors, imprecision related to the method adopted for estimating the data. The methods which have been proposed for treating errors in the data, may also be applied to different kinds of data that in real life are of interval type. The uncertainty in the data, which is strictly connected to the above errors, may be treated by considering, rather than a single value for each data, the interval of values in which it may fall: *the interval data*. The purpose of the present paper is to introduce methods for analyzing the *interdependence* and *dependence* among *interval-valued* variables. Statistical units described by interval-valued variables can be assumed as a special case of Symbolic Object (SO). In Symbolic Data Analysis (SDA), these data are represented as boxes. Accordingly, the purpose of the present work is the extension of the *Principal Component Analysis* to obtain a visualization of such boxes, on a lower dimensional space. Furthermore, a new method for fitting an *interval simple linear regression* equation is developed. With difference to other approaches proposed in the literature that work on scalar recoding of the intervals using classical tools of analysis, we make extensively use of the interval algebra tools combined with some optimization techniques.

Keywords

INTERVAL-VALUED VARIABLE, INTERVAL ALGEBRA, INTERVAL CORRELATION MATRIX, INTERVAL EIGENVECTORS, INTERVAL EIGENVALUES, INTERVAL REGRESSION LINE, VISUALIZATION

Machine Learning for Microarray Prediction in Clinical Research

Berthold Lausen

Department of Biometry and Epidemiology, University of Erlangen-Nuremberg,
Waldstr. 6, D-91054 Erlangen, Germany

Abstract. Microarray gene expression experiments can be used to identify a set of genes differentiating types of cancer and to improve classification of cancer tissue samples. For illustration I use microarray versus conventional prediction of lymph node metastasis in colorectal carcinoma. Recent developements for summarization of affymetrix probe level data are discussed. I illustrate possibilities to identify new sets of genes which discriminate colorectal cancer with and without lymph node metastatis. Moreover, I discuss bundling classifiers and other recent machine learning proposals for clinical prediction with affymetrix microarray data. Bundling classifiers allows the combination of several classification methods and avoids a method selection bias.

References

CRONER, R.S., PETERS, A., BRUECKL, W.M., MATZEL, K.E., KLEIN-HITPASS, L., BRABLETZ, TH., PAPADOPOULOS, J., HOHENBERGER, W., REINGRUBER, B., and LAUSEN, B. (2005): Microarray versus conventional prediction of lymph node metastasis in colorectal carcinoma. *Cancer, 104, 395–404*.

DETTLING, M. (2004): BagBoosting for tumor classification with gene expression data. *Bioinformatics 20(8), 3583–3593*.

HOCHREITER, S., CLEVERT, D.-A., and OBERMAYER, K. (2006): A new summarization method for affymetrix probe level data. *Bioinformatics 22(8), 943–949*.

HOTHORN, T., and LAUSEN, B. (2005): Bundling classifiers by bagging trees. *Computational Statistics and Data Analysis 49, 1068–1078*.

LAUSEN, B. (2002): Bioinformatics and classification: The analysis of genome expression data. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Data Analysis*. Springer, Berlin, 455–461.

LAUSEN, B., HOTHORN, T., BRETZ, F., and SCHUMACHER, M. (2004): Assessment of optimal selected prognostic factors. *Biometrical Journal 46, 364–374*.

Keywords

MICROARRAY GENE EXPRESSION, BUNDLING CLASSIFIERS

Symbolic Clustering of Large Datasets

Yves Lechevallier[1], Rosanna Verde[2], and Francisco de A.T. de Carvalho[3]

[1] INRIA, Domaine de Voluceau, Rocquencourt, 78153 Le Chesnay Cedex, France (yves.lechevallier@inria.fr)
[2] Dip. di Strategie Aziendali e Metod. Quantitative, Seconda Universitá di Napoli, Piazza Umberto I, 81043 Capua (CE), Italy (rosanna.verde@unina2.it)
[3] Centro de Informatica - CIn/UFPE, Av. Prof. Luiz Freire, s/n, Cidade, Universitaria, CEP 50740-540, Recife-PE, Brasil (fatc@cin.ufpe.br)

Abstract. We present an approach to cluster large datasets based on a dynamic clustering algorithm of symbolic data (SCLUST). A preliminary data reduction, using Kohonen Self Organizing Maps (SOM), is performed. As results, the individual measurements are replaced by micro-clusters which are grouped in a few clusters modeled by symbolic objects. By computing the extension of these symbolic objects, symbolic clustering algorithm allows discovering the natural classes. An application on a real data set shows the usefulness of this methodology.

References

Ambroise, C., Seże, G., Badran, F., Thiria, S. (2000): Hierarchical clustering of Self-Organizing Maps for cloud classification. *Neurocomputing, 30, 47–52*.

Bock, H. H. and Diday, E. (2000): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Springer, Heidelberg.

Breiman, L., Friedman, J.H., Oslhen, R.A., Stone, C.J. (1984): *Classification and regression trees*. Chapman & Hall/CRC.

Celeux, G. , Diday, E. , Govaert, G. , Lechevallier, Y. , Ralambondrainy, H. (1988): *Classification Automatique des Données : Environnement Statistique et Informatique*. Dunod, Gauthier-Villards, Paris.

Chavent, M., De Carvalho, F. A. T., Lechevallier, Y. and Verde, R. (2003). Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle. *Revue de Statistique Appliquée, v. LI, n. 4, p. 5-29*.

Michalski, R. S., Diday, E. and Stepp, R. E.(1981). A recent advance in data analysis: Clustering Objects into classes characterized by conjunctive concepts. In: Kanal L. N., Rosenfeld A. (Eds.): *Progress in Pattern Recognition*. North-Holland, 33–56.

Murtagh, F. (1995): Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Patterns Recognition Letters, 16, 399–408*.

Keywords

DYNAMIC CLUSTERING ALGORITHM, KOHONEN SELF ORGANIZING MAPS, SYMBOLIC DATA ANALYSIS, INTERVAL DATA, MODAL SYMBOLIC DATA, HAUSDORFF DISTANCE

A Consensus Approach Based on Frequent Groupings

Bruno Leclerc

Centre d'Analyse et de Mathématique Sociales,
École des Hautes Études en Sciences Sociales
54 bd Raspail, 75006 Paris, France

Abstract. We consider a *profile* (k-tuple) $\mathcal{P} = (\mathcal{C}_1, ..., \mathcal{C}_k)$ of classifications of a finite set E. Each classification is a family of subsets (classes) of E, and is assumed here to be a *Sperner family*, i.e. for any $i \in K = \{1, ..., k\}, C, C' \in \mathcal{C}_i$ and $C \subseteq C'$ implies $C = C'$ (partitions are such families). Given the profile $\mathcal{P}$, we define the *$\mathcal{P}$-grouping index* $g_{\mathcal{P}}$ on the power set 2^E of all the subsets of E by: for any $A \subseteq E$,

$$g_{\mathcal{P}}(A) = |\{i \in K : A \subseteq C \text{ for some } C \in \mathcal{C}_i\}|.$$

So, $g_{\mathcal{P}}(A)$ is the absolute frequency of finding the elements of A grouped inside a class when scanning the profile $\mathcal{P}$. A consensus method is associated to the previous index. Given a number t, with $0 < t \leq k$, we say that $A \subseteq E$ is a *frequent grouping* if $g_{\mathcal{P}}(A) \geq t$, and the *$(\mathcal{P}, t)$-frequent grouping consensus* of $\mathcal{P}$, denoted $FG_t(\mathcal{P})$ is the Sperner family of all the maximal frequent groupings.

Although they are somewhat natural, frequent groupings do not seem to have hold attention until now. They generalize the *frequent items* of data mining, obtained when each of the $\mathcal{C}_i$'s reduces to a single class. In relation with association rules extraction, frequent items have been extensively studied, especially on their algorithmic aspects. In this talk, we consider, on the one hand, the relations between frequent items and frequent groupings and, on the other hand, the basic properties of frequent groupings as a consensus rule.

References

HIPP, J., GÜNTZER, U. and NAKHAEIZADEH, G. (2000): Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations, 2/1, 58–64*.

BEN YAHIA, S. and MEPHU NGUIFO, E. (2004): Approches d'extraction de règles d'association basées sur la correspondance de Galois. *RSTI-ISI, 9/3-4, 23–55*.

Keywords

CONSENSUS, FREQUENT ITEMS, NONHIERARCHICAL METHODS

Tree Structured Prognostic Model for HCC Using the Gene Expression Data

Taerim Lee

Department of Information & Statistics, Korea national Open University
#169 Dongsung dong, Jongro ku, Seoul, Korea

Abstract. A tree-structured method for analyzing censored survival data is attractive for several reasons. The association between survival time and covariates is easy to interpret even if the covariates are numerous and there are interactions among them. Moreover a tree-structured method using gene expression data gives additional insight into the role of the covariates. Recent cDNA microarray technology makes it possible to find out large numbers of gene related with human cancers and to support accurate diagnosis of human cancers according to their several pathological judgements.

Recent progress in both diagnostic and therapeutic technique of hepatocellular carcinoma appears to improve the prognosis. The purpose of this study was designed to evaluate the prognosis of HCC in relation to treatment methods and their affecting gene and clinical factors by tree structured model. We could identify a feature set of 44 genes to distinguish the status of HCC and non-tumor liver tissues.

Our proposed survival tree model shows the genes related with clinical, pathological data and risk factors of HCC. These findings could be available to predict prognosis of HCC and give valuable information to justify the treatment strategy for clinician.

References

QIN L.F. and NG I.O.L. (2001): Expression of p27Kip1 and p21WAF1/CIP1 in primary hepatocellular carcinoma: Clinicopathologic cerrelation ans survival analysis. *Human Pathology, 32,778-785.*

TIBSHIRANI R., HASTIE T., NARASIMHAN B. (2002): Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America. USA, 99, 6567-6572*.

Keywords

HEPATOCELLULAR CARCINOMA, MICROARRAY, TREE STRUCTURED SURVIVAL MODEL

Benchmarking Cluster Algorithms

Friedrich Leisch

Institut für Statistik, Ludwig-Maximilians-Universität München,
Ludwigstraße 33, D-80539 München, Germany

Abstract. Benchmark studies are a popular tool to compare the performance of several competing algorithms for a certain statistical learning problem. We consider the case of comparing partitioning cluster analysis and use algorithm stability and average within-cluster sum of distances as performance measures. The main objective of this work is to port the general framework by Hothorn et al. (2005) for the design of benchmark experiments from the classification & regression setting to cluster analysis. By sampling from the empirical distribution of a given data set and clustering each of these bootstrap samples, a sample of iid observations of complete partitions is derived. Interesting characteristics of these partitions can then be examined using standard techniques from exploratory data analysis and inferential statistics, most importantly standard statistical test procedures. To demonstrate the usefulness in practice, the theoretical results are applied to benchmark studies based on artificial and real world data.

References

HOTHORN, T., LEISCH, F., ZEILEIS, A. and HORNIK, K. (2005): The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14/3, 675–699*.

Keywords

CLUSTER ANALYSIS, BENCHMARK STUDIES, BOOTSTRAPPING, PARTITION COMPARISON

Nonparametric Inference in Principal Components Analysis by Using Permutation Tests: Two Approaches Compared

Mariëlle Linting, Jacqueline Meulman, and Bart Jan van Os

Data Theory Group, Faculty of Social and Behavioral Sciences, Leiden University
Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands

Abstract. Principal components analysis (PCA) is a nonparametric method, frequently used to reduce a large number of observed variables to a smaller number of latent variables, called principal components. Despite the fact that PCA does not make distributional assumptions and is often used for exploratory research, it need not be deprived of inferential statistics. For instance, permutation tests may be applied to establish the statistical significance of elements of the PCA solution without making traditional statistical assumptions.

Permutation tests involve generating new data sets of the same size as the observed data set by randomly permuting the observed variables, causing the correlational structure of the data to be destroyed. For each parameter of interest, a value is calculated from each of the generated data sets, and all of these values form the null distribution of the parameter. The observed value of the parameter can then be compared to this null distribution, and p-values can be computed to establish the significance of the parameter.

When looking at the statistical significance of the PCA solution, we may focus on the variance accounted for (VAF) by the solution as a whole (indicated by the sum of the eigenvalues of the principal components), or on the significance of the contribution of separate variables (indicated by the component loadings). In previous research, both these issues were approached by permuting all of the variables in a data set simultaneously (Buja and Eyuboglu, 1992). However, we believe that it is theoretically more sensible to assess the contribution of one variable, given the structure of the other variables. Then, instead of the data set as a whole, the scores of one single variable are permuted, while keeping the other variables fixed. In the current study, we compare this new approach to the approach of Buja and Eyuboglu (1992) by means of an extensive simulation study.

References

BUJA, A. and EYUBOGLU, N. (1992): Remarks on parallel analysis. *Multivariate Behavioral Research, 27*, 509–540.

Keywords

PRINCIPAL COMPONENTS ANALYSIS, PERMUTATION TESTS, COMPONENT LOADINGS

Discovering Modules in Time-Series Gene Expression Data Using Biclustering

Sara C. Madeira[1 2 3] and Arlindo L. Oliveira[1 2]

[1] INESC-ID, Lisbon, Portugal
[2] Technical University of Lisbon, IST, Lisbon, Portugal
[3] University of Beira Interior, Covilhã, Portugal

Abstract. Several non-supervised machine learning methods have been used in the analysis of gene expression data. Recently, biclustering, a non-supervised approach that performs simultaneous clustering on the row and column dimensions of the data matrix, has been shown to be remarkably effective in a variety of applications. The advantages of biclustering (when compared to clustering) in the discovery of local expression patterns has been extensively studied and documented (Madeira and Oliveira, 2004). These expression patterns can be used to identify relevant biological processes possibly involved in regulatory mechanisms. Although, in its general form, biclustering is NP-complete, in the case of time-series expression data the interesting biclusters can be restricted to those with contiguous columns leading to a tractable problem.

In this context, we have recently proposed CCC-Biclustering (Madeira and Oliveira, 2005), an algorithm that finds and reports all maximal contiguous column coherent biclusters (CCC-biclusters) in time linear on the size of the expression matrix by processing a discretized matrix using string processing techniques based on suffix trees. Each expression pattern shared by a group of genes in a contiguous subset of time points is a potentially relevant biological process (module). However, discretization may limit the ability of the algorithm to discover biologically relevant patterns due to the noise inherent to most Microarray experiments. To overcome this problem we present a new algorithm that finds CCC-biclusters with up to a given number of errors per gene in the expression pattern that identifies the CCC-bicluster. These errors can, in general, be substitutions of a symbol in the expression pattern by other symbols in the alphabet (identifying measurement errors), or restricted to the lexicographically closer discretization symbols (identifying discretization errors).

References

MADEIRA S.C. and OLIVEIRA A.L. (2004): Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), 24–45.

MADEIRA S.C. and OLIVEIRA A.L. (2005): A linear time algorithm for biclustering time series expression data. In *Proc. of 5th Workshop on Algorithms in Bioinformatics*, Springer, LNCS/LNBI 3692, 39–52.

Keywords

BICLUSTERING WITH ERRORS, TIME-SERIES EXPRESSION DATA, MODULES, EXPRESSION PATTERNS, BIOLOGICAL PROCESSES

New Efficient Algorithm for Modeling Partial and Complete Gene Transfer Scenarios

Vladimir Makarenkov[1], Alix Boc[1], Charles F. Delwiche[2], Alpha Boubacar Diallo[1], and Hervé Philippe[3]

[1] Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), H3C 3P8, Canada,
[2] Cell Biology and Molecular Genetics, HJ Patterson Hall, Bldg. 073, University of Maryland at College Park, MD 20742-5815, USA.
[3] Département de biochimie, Faculté de Médecine, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, QC, H3C 3J7, Canada.

Abstract. Species evolution is a complex process comprising a number of important reticulation mechanisms such as horizontal (i.e. lateral) gene transfer, hybridization, and homoplasie. We describe a new algorithm allowing one to predict and visualize possible gene transfers events. The proposed algorithm relies either on a metric or topological optimization to estimate the probability of a horizontal gene transfer between any pair of edges in a species phylogeny. Species classification will be examined in the framework of the complete and partial gene transfer models.

References

DOOLITTLE, W. F. (1999): Phylogenetic classification and the universal tree. *Science, 284, 2124-2129*.

GUINDON, S. and GASCUEL, O. (2003): A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol., 52, 696-704*.

MAKARENKOV, V. and LECLERC, B. (1999): An algorithm for the fitting of a tree metric according to a weighted LS criterion. *J. of Classif., 16, 3-26*.

MAKARENKOV, V. (2001): reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics, 17, 664-668*.

MAKARENKOV, V., BOC, A. and DIALLO, A. B. (2004): Representing lateral gene transfer in species classification. Unique scenario. In: D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul (eds.): *Classification, Clustering and Data Mining Applications*. Springer Verlag, proc. IFCS 2004, Chicago 439-446

PAGE, R. D. M. and CHARLESTON, M. A. (1998): Trees within trees: phylogeny and historical associations. *Trends Ecol. Evol., 13, 356-359*.

WOESE, C., OLSEN, G., IBBA, M. and SÖLL, D. (2000): Aminoacyl-tRNA synthetases, genetic code, evolut. process. *Micr. Mol. Biol. Rev., 64, 202-236*.

Keywords

HORIZONTAL GENE TRANSFER, PHYLOGENETIC TREE, PHYLOGENETIC NETWORK, RETICULATE EVOLUTION, LEAST-SQUARES OPTIMIZATION, ROBINSON AND FOULDS TOPOLOGICAL DISTANCE, PARTIAL AND COMPLETE GENE TRANSFER

A Preliminary Approach to the Evaluation of MDS Unfolding Fit Measures

Ana Alexandra A. F. Martins[1] and Margarida G. M. S. Cardoso[2]

[1] Área Cientifica de Matemática, Instituto Superior de Engenharia de Lisboa,
Rua Conselheiro Emídio Navarro 1, 1950-062 Lisboa, Portugal
[2] Dep. of Quantitative Methods, Business School, ISCTE,
Av. das Forąs Armadas, 1649-026 Lisboa, Portugal

Abstract. The different methodologies in Multidimensional Scaling (MDS) are strongly conditioned by the type of input data. The problem of solution evaluation in MDS cannot thus be dissociated of the data characteristics.

The analysis of results in MDS has two perspectives. The first, of a more technical nature, concerns the evaluation of the model's adjustment in face of the available results (adjustment measures). A second important aspect is the results interpretability in the data's origin context. In practice, it frequently happens that these two perspectives collide, being difficult to achieve a balance between them.

A preliminary research is proposed in order to study the effects of alternative scales in MDS unfolding fit measures. An application is used to test the impact of alternative preferences scales.

The application reports the use of MDS unfolding techniques to visualize electoral results.

Keywords

MULTIDIMENSIONAL SCALING, UNFOLDING

New Developments in COSA: Clustering Objects on Subsets of Attributes

Jacqueline J. Meulman[1] and Jerome H. Friedman[2]

[1] Data Theory Group, FSW, and Mathematical Instute, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands
[2] Department of Statistics, Stanford University, 290 Serra Mall, Stanford CA 94305, USA

Abstract. The motivation for clustering objects on subsets of attributes (COSA) was given by consideration of data where the number of attributes is much larger than the number of objects. Obvious application is in systems biology (genomics, proteomics, and metabolomics). When we have a large numbers of attributes, objects might cluster on some attributes, and be far apart on all others. Common data analysis approaches in systems biology are to cluster the attributes first, and only after having reduced the original many-attribute data set to a much smaller one, one tries to cluster the objects. The problem here, of course, is that we would like to select those attributes that discriminate most among the objects (so we have to do this while regarding all attributes multivariately), and it is usually not good enough to inspect each attribute univariately. Therefore, two tasks have to be carried out simultaneously: cluster the objects into homogeneous groups, while selecting different subsets of variables (one for each group of objects). The attribute subset for any discovered group may be completely, partially or nonoverlapping with those for other groups. The notorious local optima problem is dealt with by starting with the inverse exponential mean (rather than the arithmetic mean) of the separate attribute distances. By using a homotopy strategy, the algorithm creates a smooth transition of the inverse exponential distance to the mean of the ordinary Euclidean distances over attributes. New insight will be presented for the homotopy strategy, and the weights that are crucial in the COSA procedure but that were rather underexposed as diagnostics in the original paper.

References

FRIEDMAN, J.H., and MEULMAN, J.J. (2004a). Clustering objects on subsets of variables (with discussion). *Journal of the Royal Statistical Society, Series B, 66, 815–849.*

FRIEDMAN, J.H., and MEULMAN, J.J. (2004b). *The COSA program in an R-environment*, available at http://www-stat.stanford.edu/~jhf/COSA.html.

Keywords

BIOINFORMATICS, FEATURE SELECTION, INVERSE EXPONENTIAL DISTANCE, SUBSET WEIGHTS, SUBSPACE CLUSTERING, TARGETED CLUSTERING

Empirical Study on the ICMP Traffic Data via the Internet

Hiroyuki Minami and Masahiro Mizuta

Information Initiative Center, Hokkaido University,
N11 W5, Kita-ku, Sapporo, JAPAN

Abstract. The Internet has become one of our life infrastructures. We wish we could utilize it through the ideal condition, say, enough network speed without any stress. However, we often realize slow accesses and sometimes tend to offer a complaint to an Internet Service Provider because we know that there are few operations to seek for this kind of problems.

A group of Internet operators in Japan is challenging to improve the effectiveness and quality about the Internet they provide to the customers. They have been collecting the periodical response time data, generated artificially. Their size is very huge, however, the data are very attractive for us since they have some interesting features. For example, the response time is not proportional to the real distance between hosts. There are many outliers in the data since some responses are much delayed due to heavy traffic jams and network incidents, e.g. invalid access, virus affection, Denial of Service, etc.

In this study, we apply some multivariate data analysis methods, including Functional Data Analysis technique to the data.

References

MINAMI, H. and MIZUTA, M. (2005): Statistical Approach on Internet Traffic Data. *Proc. of the 5th IASC Asian Conference on Statistical Computing, 109–112*.
CHEN, W.W.S. (Ed.) (2005). Statistical Methods in Computer Security. Marcel Dekker.

Keywords

INTERNET, FDA, MDS

Clustering in Different Perspectives: How Many Clusters are There?

Boris Mirkin

School of Computer Science and Information Systems, Birkbeck, University of London

Abstract. This talk is an attempt at structuring and systematising the development of clustering as a discipline. As the attendants to this conference are well aware, clustering is devoted to finding and describing cohesive or similar or homogeneous parts in data, the clusters. Common clustering structures are: partition, hierarchy, and a single cluster. Among distinguishable goals of clustering are data structuring, data describing, data generalizing, associating aspects, and information visualising.

At least four different frameworks can be distingushed in the literature on clustering as a data analysis discipline: classical statistics, machine learning, data mining, and knowledge discovery. In classical statistics, clustering is a method to fit a prespecified probabilistic model of the data generating mechanism. In machine learning, clustering is a tool for prediction. In data mining, clustering is a tool for finding patterns and regularities within the data. In knowledge discovery, clustering is a tool for updating, correcting and extending the existing knowledge; in this regard, clustering is but empirical classification.

Different frameworks may lead to different views on the same issues such as that of the optimal number of clusters. This is a very much important question in the classical statistics perspective and it is of minor importance in the knowledge discovery perspective.

The talk will focus on the data mining and knowledge discovery perspectives.

In data mining, the author advocates the so-called data recovery paradigm as described in (Mirkin 2005). A model for K-Means clustering within this paradigm leads to the concept of most contributing (anomalous) clusters, which can be used for addressing the issue of the number of clusters in several ways. One of them is restricting the user with significant clusters only (incomplete clustering) and another is the so-called intelligent K-Means, iK-Means (Mirkin 2005), to follow the one-by-one strategy of the principal component analysis for initialising the clustering process. In our experiments with Gaussian-mixture generated data, in many cases, iK-Means is the only winner among eight different procedures for determining K in K-Means, or, in other cases, trailing closely behind the other winners, the Kaufman-Rousseeeuw's silhouette width based approach or Hartigan's index based method.

In the knowledge discovery perspective, the number of clusters is to be determined according to the criterion of consistency with the knowledge of the domain. This approach is yet to be developed. We present a real-world example of the evolutionary analysis of protein families, in which the level of aggregation is determined with a two-step process. At the first step, the existing knowledge of protein function is utilised. The second, fine-tuning, step, involves the knowledge obtained from the aggregated protein families.

References

B. Mirkin (2005) *Clustering for Data Mining: A Data Recovery Approach*. Chapman and Hall/CRC.

Keywords

CLASSICAL CLUSTERING, DATA RECOVERY CLUSTERING, NUMBER OF CLUSTERS, INTERPRETATION

How Factorial Techniques Can Support Translation Processes in the Web Era

Michelangelo Misuraca

Dipartimento di Matematica e Statistica, Università "Federico II",
Via Cinthia - Complesso Monte S.Angelo, 80126 Napoli, Italy

Abstract. The approach to information sources has deeply modified in the Web era, compared with traditional media. It is possible to consider both a wider information diffusion (from a social, cultural and geographic viewpoint) and a differentiation of contents, relating to the subjective knowledge requirements.

Frequently it is possible to obtain sets of documents translated into different languages, known as *multilingual corpora*. In particular we have *parallel corpora* when the translation mate is an exact translation (e.g. United Nations or EU multilingual *corpora*), and *comparable corpora* when the translation mate is an approximate translation (e.g. e-documents on Internet).

The analysis of multilingual *corpora* has been discussed both in Textual Statistics (TS) and in Text Mining (TM). Particularly in TS has been faced up to deal with open-ended questions in multinational surveys (Lebart 1998); in TM has been faced up to develop language-independent representations of terms, in the frame of natural language and machine translation applications (Grefenstette, 1998).

Factorial techniques, as Latent Semantic Indexing or Correspondence Analysis, allow to graphically visualize the similarities/dissimilarities of documents in terms of used vocabulary, by using different kind of metrics. This paper aims at reviewing the problem of applying Procrustes rotations for comparing two *corpora* in different languages (Balbi and Misuraca, 2005, 2006), for evaluating the translation process in a Textual Data Analysis framework.

References

BALBI, S. and MISURACA, M. (2005): Procrustes Techniques for Text Mining. In: S. Zani and A. Cerioli (Eds.): *Book of the Short Papers, Meeting of the Classification and Data Analysis Group of the Italian Statistical Association* (CLADAG05). MUP, 37–40.

BALBI, S. and MISURACA, M. (2006): Rotated Canonical Correlation Analysis for Multilingual Corpora. In: *Actes de 8es Journées internationales d'Analyse statistique des Données Textuelles* (JADT06). [to appear]

GREFENSTETTE, G. (Ed.) (1998): *Cross Language Information Retrieval*. Kluwer Academic Publishers, London.

LEBART, L. (1998): Text mining in different languages. *Applied Stochastic Models and Data Analysis, 14, 323–334*.

Keywords

FACTORIAL ANALYSIS, PROCRUSTES ROTATIONS, TEXTUAL DATA

Evaluation of the Results of Functional Clustering

Masahiro Mizuta

Advanced Data Science Laboratory,
Information Initiative Center, Hokkaido University,
N.11, W.5, Kita-ku, Sapporo 060-0811, Japan

Abstract. In this paper, we deal with functional clustering. We have studied clustering methods for functional data including functional single linkage, functional k-means. There are two frameworks on functional clustering; results of functional clustering depended on the arguments or not. If we adopt the former framework, we can get groups of the objects with the methods. The result is not functional. Functional clustering methods under the latter framework produce *functional* grouping. Of course, which framework is better depends upon the functional data. We define a degree of the difference between two results: groups of objects and *functional* grouping of objects. The degree is useful for evaluating the results of functional clustering.

References

MIZUTA, M. (2003a): Hierarchical clustering for functional dissimilarity data. In: *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics*, Volume V, 223 – 227.

MIZUTA, M.(2003b): K-means method for functional data. In: *Bulletin of the International Statistical Institute, 54th Session, Book* 2, 69–71.

RAMSAY, J.O. & SILVERMAN, B.W. (2005): *Functional Data Analysis*. 2nd edition, New York: Springer-Verlag.

Keywords

FUNCTIONAL k-MEANS, FUNCTIONAL DATA ANALYSIS, FUNCTIONAL CLASSIFICATION DEPENDING ARGUMENTS

Resampling-Based Class Discovery from Microarray Data: Guidelines from Benchmarking Studies

Ulrich Möller

Leibniz Institute for Natural Product Research and Infections Biologie -
Hans Knöll Institute, Beutenbergstr. 11a, 07745 Jena, Germany

Abstract. Although not yet general practice data resampling (Lunneborg, 2000) is increasingly considered relevant in post-genomic (e.g., microarray) data analysis (cf. Michiels et al., 2005). Processing of many resamples multiplies the computation time. The user must choose among qualitatively different resampling schemes, where some have a crucial control parameter. Hence, the possible systematic error of this step that influences subsequent classification or modelling stages is of interest. However, related theoretical considerations and empirical results are sparse or lacking (as for unsupervised learning). We summarize a benchmark study of resampling performance in a clustering framework (Möller and Radke, 2006). Novel ideas for optimized resampling control are presented with results of microarray-based tumor class discovery. From this work clues are derived to configuring future applications.

References

LUNNEBORG, C.E. (2000): *Data analysis by resampling - concepts and applications*. Duxbury Press, Pacific Grove.

MICHIELS, S. KOSCIELNY, S. HILL, C. (2005): Prediction of cancer outcome with microarrays : a multiple random validation strategy. *Lancet, 365, 488–492*.

MÖLLER, U. and RADKE, D. (2006): Performance of data resampling methods for robust class discovery based on clustering. *Intelligent Data Analysis, 10/2, in press*.

Keywords

RESAMPLING, CLUSTERING, BENCHMARKING, MICROARRAY GENE EXPRESSION DATA

Regional Clusters and Socio-Economic Diversity in the European Union

Carlos M. F. Monteiro[1], João Oliveira Soares[1], and Cristina del Campo[2]

[1] CEG-IST, Instituto Superior Técnico, Universidade Técnica de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
[2] Dpto. de Estadíística e I.O. II, Universidad Complutense de Madrid, Campus de Somosaguas, 28223 – Madrid, Spain

Abstract. There are significant differences among European Union regions, which have been even heightened due to the recent enlargement in 2004. This paper aims to analyse this diversity and proposes a classification of European Regions that is adjusted to the different axes of socio-economic development and simultaneously is useful for the European regional policy purposes. The data used in this paper was published by the Eurostat and corresponds to the main statistical indicators of NUTS2 regions in the EU. Exploratory factor analysis was used to identify a smaller set of development dimensions. To search for groups of regions hierarchical and non-hierarchical clustering procedures were carried out on the factor scores. This methodology allowed the identification of clusters of socio-economic similarity which are confronted with the classes considered in the financial proposal of the European Commission (EC) for the period 2007–2013. It was found that each of the two main groups of the EC classification, convergence regions and competitiveness and employment regions, comprises at least two significantly different groups of regions, which differ not only by their average income but also by other indicators. The two other groups, phasing-in and phasing-out regions, also seem to lack homogeneity, being spread through different clusters.

Keywords

REGIONAL DEVELOPMENT, REGIONAL INDICATORS, CLUSTER ANALYSIS, FACTOR ANALYSIS, EUROPEAN POLICY

Finding Meaningful and Stable Clusters Using Local Cluster Analysis

Hans-Joachim Mucha

Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS),
D-10117 Berlin, Germany

Abstract. Let us consider the problem of finding clusters in a heterogeneous, high-dimensional setting. Typically, a cluster analysis model is applied to reach this aim. As a result, often ten or more clusters are detected in a heterogeneous data set. Usually, one can observe that the clusters have a quite different stability. Some of them are very stable. Thus, they can be reproduced and confirmed to a high degree, for instance, by simulations based on random resampling techniques. They are both homogeneous inside and well separated from each other. Moreover, sometimes they are located far away from the main body of the data like outliers. On the other side, hidden and tight neighboring clusters are more difficult to detect and they cannot be reproduced to a high degree.

The idea of this paper is to perform local clusterings subsequent to the usual cluster analysis. Here the following two main questions arise. Is it possible to improve the stability of some of the clusters? Are there new clusters that are not yet detected by global clustering? The paper presents a methodology for such an iterative clustering that can be a useful tool in discovering stable and meaningful clusters. To define stability with respect to individual clusters, measures of correspondence between subsets of some finte set are used. This is the middle part of a three level build-in validation of stability (Mucha (2004)). The proposed methodology is used successfully in the field of archaeometry. Here, without loss of generality, it is applied to hierarchical cluster analysis. The improvements of local cluster analysis will be illustrated by means of multivariate graphics.

References

MUCHA, H.-J. (2004): Automatic Validation of Hierarchical Clustering. In: J. Antoch (Ed.): *Proceedings in Computational Statistics, COMPSTAT 2004, 16th Symposium*. Physica-Verlag, Heidelberg, 1535–1542.

Keywords

CLUSTER ANALYSIS, JACCARD COEFFICIENT, WARD'S METHOD, HIERARCHICAL CLUSTER ANALYSIS, STABILITY, RESAMPLING

Evaluating Different Approaches to Measuring the Similarity of Melodies

Daniel Müllensiefen and Klaus Frieler

Institute of Musicology, University of Hamburg,
Neue Rabenstr. 13, D-20354 Hamburg, Germany

Abstract. We describe an empirical approach to evaluating similarity measures for the comparision of melodies. The approach taken here is based on a comparision of a variety of algorithmic similarity measures with experimentally gathered rating data from human music experts. First we give an overview of melodic similarity measures found in the literature and sketch a general construction scheme. Next we present our experimental setting and summarize the main results, along with the construction of an optimised measure on the basis of a linear model. In the following section we extensively discuss the nature of similarity and distance measures for melodies and other musical material by comparision of human and algorithmic approaches. We claim that the concept of 'similarity' is more adequate for cognitive processes as opposed to the notion of 'distance'. Futhermore, we point out that there exists a high context-dependency of human similarity judgments, although in a given context judgments of human experts reveal high inter- and intra-personal reliability, whichs makes algorithmic modelling an fairly successful enterprise.

References

MÜLLENSIEFEN, D. (2004): *Variabilität und Konstanz von Melodien in der Erinnerung. Ein Beitrag zur musikpsychologischen Gedächtnisforschung.* PhD work, University of Hamburg.

MÜLLENSIEFEN, D. and FRIELER, K. (2004a): Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments. *Computing in Musicology, 13, 147–176.*

MÜLLENSIEFEN, D. and HENNIG, CH. (2005): Modeling Memory for Melodies. In: *Proceedings of the 29th Annual Conference of the German Society for Classification (GfKl).* Springer, Berlin

STEIBECK, W. (1982): *Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse.* Bärenreiter, Kassel.

UITDENBOGERD, A.L. (2002): *Music Information Retrieval Technology.* PhD thesis, RMIT University Melbourne Victoria, Australia

Keywords

MELODIC SIMILARITY, SIMILARITY MEASURES, SIMILARITY PERCEPTION, LINEAR MODEL

Ultrametrics in Data Analysis, Quantum Physics, and Computational Logic

Fionn Murtagh

Dept. Computer Science, Royal Holloway, University of London,
Egham, Surrey TW20 0EX, United Kingdom

Abstract. An ultrametric or tree distance is used in hierarchical clustering. P-adic algebra, expressing ultrametric topology, is used in the physics of the early Universe, and in quantum statistics. In semantic analysis, spherically complete ultrametric spaces play an important role in the proof of computability of recursive systems.

Recent applications to data analysis have included: (i) computational implications of pervasive ultrametricity resulting from sparse and possibly high-dimensional spaces; (ii) textual analysis based on inherent, local hierarchical structure; (iii) time series analysis, again based on inherent, local hierarchical structure; and (iv) new algorithms for search and matching in high dimensional massive data sets.

We will review recent progress in this interface area and show how it leads to new insights on information-computation.

Keywords

HIERARCHICAL CLUSTERING, TEXT ANALYSIS, TIME SERIES CLUSTERING, PROXIMITY SEARCH, NEAREST NEIGHBORS, COMPUTATIONAL COMPLEXITY, DATA CODING, HIGH DIMENSIONAL DATA ANALYSIS

A Review on Block Clustering under the Mixture Approach

Mohamed Nadif[1] and Gérard Govaert[2]

[1] LITA, Université de Metz, Ile du Saulcy, 57045 Metz Cedex, France
mohamed.nadif@univ-metz.fr
[2] HEUDIASYC, UMR CNRS 6599, Université de Technologie de Compiègne,
BP 20529, 60205 Compiègne Cedex, France
gerard.govaert@utc.fr

Abstract. The basic principle of block clustering methods is to make permutations of a set of objects I and a set of variables J in order to construct a correspondence structure on $I \times J$. The advantage of such methods is that they distill data matrix into a simpler data matrix having the same structure. Moreover, they process large data sets as far less computation is required than for processing the two sets I and J separately.

Knowing that the use of the models of mixture is of a great interest in classification, we have recently proposed a mixture model taking into account the block clustering problem on the both sets. We have developed the block EM algorithm as part of the maximum likelihood and fuzzy approaches (Govaert and Nadif, 2005), and the block CEM algorithm as part of the classification maximum likelihood approach (Govaert and Nadif, 2003). These algorithms are based on the alternated application of EM or CEM on intermediate data matrices. These methods have several advantages, for example, they generalize classical methods on binary and contingency table proposed by Govaert (1984) and offer new methods more adapted on real data. We will discus their bheaviour from synthetic data and study their performance on real data.

References

GOVAERT G., NADIF, M.: An EM algorithm for the block mixture model, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2005) 643-647.

GOVAERT G., NADIF, M.: Fuzzy clustering to estimate the parameters of block mixture models, Soft Computing, **10**(5),(2006) 415-422.

GOVAERT G., NADIF M.: Clustering with Block Mixture Models, Pattern Recognition, **36**, (2003) 463-473.

GOVAERT, G.: Algorithme de classification d'un tableau de contingence, in Data analysis and informatics 3, E. Diday et al., eds, North-Holland, Amsterdam, (1984) 223-236.

Keywords

CLUSTERING, MIXTURE MODELS, EM ALGORITHM

Career Mobility over the Life Course among Women in Japan

Miki Nakai

Department of Social Sciences, College of Social Sciences, Ritsumeikan University, 56-1 Toji-in Kitamachi, Kyoto 603-8577 Japan

Abstract. The aim of the present paper is to examine the career dynamics among women in Japan. Asymmetric multidimensional scaling is used to develop better understanding of the mechanisms by which intragenerational job mobility occurs, especially women's early and mid-career changes in career mobility. Utilizing the data from the nationally representative survey in Japan in 1995, occupational mobility tables between specific ages at intervals of five, ten and twenty years for each woman are analyzed so that we can find out the cumulative effects of career shift on women's career development over the life course, as well as opportunity structures in labour market. Result reveals that transition among occupational categories varies according to the organizational environments in which the women participate and their life stages. Results also suggest that asymmetric career mobility increase in women's mid-career over time, specifically in their late 30s and early 40s, suggesting that introduction of appropriate measures related to work-life balance, especially in small organizations, and mid-career professional training among women at earlier stage after motherhood seem to be necessary to utilize more of female workforce in the era of gender equality.

References

NAKAI, M. and AKACHI, M. (2000): Labour Market and Social Participation. In: K. Seiyama (Ed.): *Gender, Market, and Family*. University of Tokyo Press, Tokyo, 111–131. (in Japanese)

OKADA, A. and IMAIZUMI, T. (1997): Asymmetric Multidimensional Scaling of Two-Mode Three-Way Proximities. *Journal of Classification, 14, 195–224*.

ROSENFELD, R. A. (1992): Job Mobility and Career Processes. *Annual Review of Sociology, 18, 39–61*.

SATO, Y. (1998): Change in Job Mobility Patterns in the Postwar Japanese Society. In: Y. Sato (Ed.): *Social Mobility and Career Analysis, 45–64*. (in Japanese)

Keywords

CAREER MOBILITY, LABOUR MARKET, ASYMMETRIC MULTIDIMENSIONAL SCALING, GENDER ROLE

Binary Classification with Support Hyperplanes

Georgi I. Nalbantov[1,2], Jan C. Bioch[2], and Patrick J.F. Groenen[2]

[1] Erasmus Research Institute of Management,
Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[2] Econometric Institute, Erasmus University Rotterdam, The Netherlands

Abstract. A new approach to the binary classification task is proposed, called: Support Hyperplanes (SH). In the linearly separable case, a point lies on the SH class-separating surface if the following three conditions are met: (1) the point is equally distant from two hyperplanes; (2) these two hyperplanes are consistent with the data; and (3) the distance from the point to (any) one of the hyperplanes is maximal.

One way to view SH is as a generalization of the state-of-art Support Vector Machines (SVM). In SVM classification, a point lies on the SVM class-separating surface if in addition to the three SH conditions, a fourth one is added, namely: (4) the two hyperplanes should be parallel. SH can be extended to handle the linearly non-separable case by means of kernels and/or introduction of slack variables. We discuss some theoretical links between SH and other methods such as SVM, Nearest Neighbor and Version Spaces, and assess the performance of SH on several medium-sized problems.

References

BURGES, C. (1998): A tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery 2, 121–167*.

MÜLLER, K., MIKA, S., RÄTSCH, G., TSUDA, K., and SCHÖLKOPF, B. (2001): An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks 12(2), 181–201*.

VAPNIK, V.N. (1995): *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Keywords

BINARY CLASSIFICATION PROBLEM, SUPPORT VECTOR MACHINES, IMPLICIT DECISION SURFACE

Multicriteria Ordered Clustering

Philippe Nemery and Yves De Smet

Service de Mathématiques de Gestion,SMG, Université Libre de Bruxelles
Boulevard du Triomphe CP 210-01, 1050 Brussels, Belgium

Abstract. In the multicriteria decision aid context, a lot of attention has been paid to assign objects to predefined ordered or not ordered groups, called respectively 'classes' or 'categories'. A number of approaches have been proposed to tackle these problems. These are mainly characterized by the definition of 'central' or 'limit' objects for these groups. To our knowledge, not enough attention has been paid to elicit ordered groups : to obtain what we will call 'ordered clusters' (ordered groups of objects) corresponding to the decision makers preferences (this will be referred as a 'multicriteria ordered clustering' problem).
Starting from a pairwise comparison of the objects that have led to a valued preference matrix without making any assumption about the way these preference degrees have been computed. An homogeneity and a coherence factor to evaluate the quality of the ordered partition will be defined. The problem of eliciting 'ordered clusters' thus will be considered as an optimization problem and a basic heuristic, inspired by the tabu search, will be presented to solve it. To validate our approach, we will present our results both on artificial problems and on data related to a financial problem.

References

ANASTASSIOU, T. and ZOPOUINIDIS, C. (1997) : Country risk assessment : a multicriteria analysis approach. *The Journal of Euro-Asian Management. 3(1), 51-73*

COSSET, J.T. and SISKOS, Y. and ZOPOUNIDIS, C. (1992) : Evaluating country risk : A decision support approach. *Global Finance Journal. 3(1), 79-95*

FIGUEIRA, J. and DE SMET, Y. and BRANS, J.P. (2003) : Promethee for MCDA Classification and Sorting problems: Promethee TRI and Promethee CLUSTER. *submitted for publication.*

FIGUEIRA, J. and TERVONEN, T. and ALMEIDA-DIASZ RISO, A. and SALMIMEN, L.P. (2004) : A Parameter Stability Analysis Method for ELECTRE TRI. *NATO Advanced Research Workshop*,Thessaloniki, Greece

ROY,B. (1985) : Méthodologie multicritére d'aide á la décision, *Gestion, Série : Production et techniques quantitatives appliquées á la gestion*, Ed. Economica

Keywords

MULTICRITERIA CLUSTERING, MULTICRITERIA DECISION AID, DATA ANALYSIS, APPLICATIONS

Investigation into Genome-Related and Nanotechnology Research at Grants-in-Aid in JAPAN

Masaki Nishizawa and Yuan Sun

National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8640, Japan

Abstract. Japanese Government and Council for Science and Technology Policy are promoting "Life Sciences", "Information and Communication Technology", "Environmental Sciences" and "Nanotechnology and Materials" field as Prioritized Areas on the 2nd Science and Technology Basic Plan (FY2001-FY2005)[1] under the Science and Technology Basic Low.

The system of Grants-in-Aid for Scientific Research[2] from Ministry of Education, Culture, Sports, Science and Technology of Japan is one of the oldest ones, which is the funding system for researchers belonging to universities and institutes in Japan. The fund was allotted to each researcher by peer review under the application for their own research projects.

In this Grants-in-Aid, it is very interesting to see how the adoption situation of the research of these prioritized area has changed. We measure strength of a related level each other by comparing "Field feature keyword" that has been developed so far[3][4] and the research subjects and the applied detail research fields in the Grants-in-Aid. Especially in this analysis, the relation between a field related to the genome, a nanotechnology in the prioritized area and the applied research subjects and their applied detail research fields are examined by using this method, and the secular distortion of the adoption situation and the mutual relation are examined by using Correspondence Analysis etc.

References

The Science and Technology Basic plan:
http://www8.cao.go.jp/cstp/english/basic/index.html

Grant-in-Aid for Scientific Research:
http://www.jsps.go.jp/english/e-grants/grants.html

M. NISHIZAWA, Y. SUN, M. YANO (2001): A Study on Informatics-Related Research Fields at Japanese Universities, *NII Journal, No.2, 45–58*.

MASAKI NISHIZAWA, YUAN SUN(2004): Research Fields related to Information Science under a new classification in Japan. *Proceedings on 4th Inernational Conference on University Evaluation and Research Evaluation, Whan University, China, 27 September 67–73*.

Keywords

PRIORITIZED AREAS, GENOME-RELATED RESEARCH, NANOTECHNOLOGY, FIELD FEATURE KEYWORD, RESEARCH FIELD, DATABASE

An Asymmetric Cluster Analysis Study of Relationships among Japanese Prefectures in Marriage

Akinori Okada

Department of Industrial Relations, School of Social Relations,
Rikkyo (St. Paul's) University, 3-34-1 Nishi Ikebukuro,
Toshima-ku Tokyo, 171-8501 Japan

Abstract. In the present study relationships among 47 Japanese prefectures in marriage are analyzed by utilizing the asymmetric cluster analysis (Okada, 2000; Okada and Iwamoto, 1996). The data analyzed in the present study consist of frequencies of the marriage between any two Japanese prefectures obtained from a matchmaking service company from 2000 to 2005 (Tanaka, 2005), which also had been analyzed by the external analysis of the asymmetric multidimensional scaling (Okada, 2006). The data can be represented as a 47×47 table. The (j, k) element of the table represents the frequency of the marriage where the female comes from prefecture j and the male comes from prefecture k. The total number of the marriage is 10,115. A dendrogram of the prefectures is obtained from the analysis. The dendrogram of the asymmetric cluster analysis tells that the geographical location among prefectures governs the relationships strongly, which has been ignored in the asymmetric multidimensional study. Because the asymmetric multidimensional scaling was done externally by giving the map (location) of the prefectures, we cannot evaluate the effect of the geographical location of prefectures in the analysis.

References

OKADA, A. (2000): An asymmetric cluster analysis study of car switching data. In: K. W. Gaul, O. Opitz, and M. Schader, (Eds.): *Data Analysis: Scientific Modeling and Practical Application*. Springer, Berlin, 495–504.

OKADA, A. (2006): Regional Closeness in Marriage among Japanese Prefectures. *Abstracts of the German Japanese Workshop on March 7–8, 2006 in Berlin*.

OKADA, A. and IWAMOTO, T. (1996): University Enrollment Flow among the Japanese Prefectures: A Comparison Before and After the Joint First Stage Achievement Test by Asymmetric Cluster Analysis. *Behaviormetrika, 23, 169–185*.

TANAKA, T. (2005): IT Jidai no Kokkon Joho Sabisu [Matchmaking Service Using Information Technology]. *ESTRELA, No. 134, 27–32*.

Keywords

ASYMMETRY, CLUSTER ANALYSIS, MATCH MAKING, REGIONAL RELATIONSHIPS

Determining the Number of Components in Mixture Regressions of Normal Data

Ana Oliveira-Brochado[1] and Francisco Vitorino Martins[2]

[1] Faculty Of Economics, University of Oporto,
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
[2] Faculty of Economics, University of Oporto,
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal

Abstract. The aim of this work is to determine how well criteria designed to help the selection of the adequate number of mixture components perform in mixture regressions of normal data. We address this research question based on results of an extensive experimental design. The simulation experiment compares several criteria, including information criteria and classification-based criteria; we intend to include several criteria not considered in previous studies. Eight data characteristics are manipulated in this full factorial design, namely: true number of segments; mean separation between segments; number of individuals; number of observations per individual; number of predictors; measurement level of predictors; error variance and minimum segment size. The performance of the segment retention criteria is evaluated by their success rates; we also investigate the influence of experimental factors and their levels on success rates.

References

ANDREWS, R. L. and CURRIM, L. S. (2003): Recovering and Profiling the True Segmentation Structure in Markets: an Empirical Investigation. *International Journal of Research in Marketing, 20, 177–192*.

BROCHADO, A. and MARTINS, F.V. (2005): Assessing the number of components in mixture models: a review. FEP Working paper N. 184.

HAWKINS, D. S., ALLEN, D. M. and STEMBERG, A. J. (2001):Determining the number of components in mixtures of linear models. *Computational Statistics and Data Analysis, 38, 15–48*.

MCLACHAN, G. and PEEL, D. (2000): *Finite Mixture Models*. Wiley.

WEDEL, M. and DESARBO, W. S. (1995): A Mixture Likelihood Approach for Generalized Linear Models. *Journal of Classification, 12, 21–55*.

Keywords

NUMBER OF MIXTURE COMPONENTS, EXPERIMENTAL DESIGN, MIXTURE REGRESSION MODEL

How Important are ICT Services for Economic Growth?

W. Polasek and R. Sellner

IHS Vienna,
Stumpergasse 56, 1060 Vienna, Austria
E-mail: polasek @ ihs.ac.at

Abstract. The research question focuses at a thorough analysis of productivity growth in EU service sectors. Recent studies have drawn attention to the higher growth rates in the USA since the mid 1990s, particularly in industries that produce ICT, or that make intensive use of ICT technologies. Findings from research into the barriers to productivity in the ICT-intensive industries will be used to suggest future options for economic policy aimed at higher productivity growth.

As a basic first stage in the analysis, the facts concerning productivity in the EU member states (and relative to the data on productivity in the US) will be presented. Productivity growth rates for individual service industries in individual EU member states are calculated, and their development of the period 1980 to 2002 is observed. In this way the relative strengths of countries in particular types of service is made transparent.

Following on from the primary description of relevant (ICT-intensive) sectors, causes for the delayed deployment of ICT in Europe compared to the US since the 1990s and the relative low growth in productivity will be studied in detail. The expected effects on further liberalisation in the internal market for services will be assessed. Further research will be conducted on the structure and policies in those EU member states and branches that do indeed show relatively high productivity increases, in order to suggest possible actions to be taken in less productive branches. Regional differences in productivity and employment in ICT-intensive branches will be further highlighted and possible trade-offs between productivity increases and product quality, or between productivity and employment will be investigated. Special attention will also be given in a part of the report to other factors that may influence the development of labour productivity in ICT-intensive service sectors. An output of the analysis will be to list barriers to productivity growth and economic policies that should be focused on in order to improve productivity.

Central to the analysis is an empirical investigation of the link between ICT investments and sectoral development over time of productivity and employment in the EU member states (where data is available) and in the USA. Based on the quantitative results it will be possible to estimate the effects of the completion of the internal market for services. The role of other factors (e.g. Innovation, infrastructure, degree of privatisation) will be discussed on the basis of current literature.

Keywords

PRODUCTIVITY GROWTH, COMPARISON OF USA AND EU, ICT INVESTMENTS, SERVICE SECTORS, PANEL ESTIMATION

Recent Advances in Model-Based Clustering: Variable Selection and Social Networks

Adrian E. Raftery[1,2]

[1] Center for Statistics and the Social Sciences, University of Washington, Box 354320, Seattle, WA 98195-4320
[2] UTIA, Prague

Abstract. Cluster analysis is the automated search for groups of related observations in a dataset. Model-based clustering bases cluster analysis on a finite mixture probabilistic model, allowing inference to be put on a formal statistical basis. This leads to maximum likelihood and Bayesian estimation of the model parameters, assessment of uncertainty about group classifications, formal inference about the number of groups present and the best clustering models, as well as robust methods for dealing with outliers.

I will describe a method for deciding which variables should be used for clustering. This recasts the variable selection problem as a model selection one, leading to a solution based on approximate Bayes factors. In experiments, we found that removing irrelevant variables often improved performance, and led to more parsimonious clustering models and easier visualization of results.

I will also describe the application of model-based clustering to social network data. Network models describe ties among interacting units or actors. Network data often exhibit transitivity, meaning that two actors that have ties to a third actor are more likely to be tied than actors that do not, homophily by attributes, meaning that actors with similar values of variables are more likely to be tied, and clustering. Interest often focuses on finding clusters of actors or ties, and the number of groups in the data is typically unknown. We propose a new model, the Latent Position Cluster Model, under which the probability of a tie between two actors depends on the distance between them in an unobserved Euclidean "social space," and the actors' locations in the latent social space arise from a mixture of distributions, each one corresponding to a cluster. It models transitivity, homophily by attributes and clustering simultaneously, and does not require the number of clusters to be known. The model makes it easy to simulate realistic networks with clustering, potentially useful as inputs to models of more complex systems of which the network is part, such as epidemic models of infectious disease.

This is joint work with Nema Dean, Mark Handcock, Jeremy Tantrum and Chris Fraley.

Spatial Hierarchical and Pyramidal Clustering Software

Mohamed Cherif Rahal and Edwin Diday

LISE-CEREMADE,Université Paris IX Dauphine,
Place du Maréchal de Lattre de Tassigny,
75775 PARIS 16

Abstract. Progress in knowledge discovery and clustering Large Databases has raised the necessity of visualizing sets of data on an aggregated form (Symbolic data). In this paper we propose pyramidal and hierarchical clustering software (SPHICS) which is based on an agglomerative clustering algorithm. Given a set to be clustered (Symbolic or classical data), we place each unit on a node of a grid. The aim of SPHICS (Spatial Pyramidal and HIerarchical Clustering Software), is to find homogeneous overlapped subsets (the pyramidal case) or non-overlapped ones (hierarchical case). As input data our software accepts XML, SDS (Symbolic SODAS data format), or native classical data, and uses several kinds of dissimilarity measures which change according to the type of data, for example Hausdorff, Euclidian, Manhattan , Minkowsky for interval and numerical data, Khi square, de Carvalho , L1-Norm for categorical, histograms, diagrams and non-numerical data. For the representation of the results we use spatial graphical and interactive tools using the OpenGL library, zoom, rotation, histograms of contribution of variables. Finally we compare the grid and its associated spatial clustering produced by our software to the grid produced by the Self Organizing Maps of Kohonen.

References

JOHNSON, S.C. (1967) : Hierarchical clustering schemes. *Psychometrika 32 pp. 241-254*

DIDAY,E. (2004): Spatial Pyramidal Clustering Based on a Tessellation. In: D. Bank and al. (Eds.): *International Federation of Classification Societies 2004*. Springer Verlag, Chicago, 105–120.

PAK, KK. (2005): Classifications Hiérarchiques et Pyramidales Spatiales et nouvelles techniques d'interprétation. *Thèse de Doctorat, Univ. Paris Dauphine Paris*.

RAHAL, MC. PAK, KK. DIDAY, E. (2005): Elagage et aide à l'interprétation symbolique et graphique d'une pyramide. In: N.vincent and Al (Eds.) *Extraction et gestion des connaissances EGC 2006*.Cepadus, Paris, 183–185.

Keywords

HIERARCHICAL AND PYRAMIDAL CLUSTERING, DISSIMILARITIES, SPATIAL CLASSIFICATION, SYMBOLIC DATA ANALYSIS, KOHONEN MAPS

Some New Criteria for Hierarchical Agglomerative Clustering and for Discriminant Analysis. Applications to Interval Data

Jean-Paul Rasson and François Roland

Department of Mathematics, Statistics Unit, University of Namur,
Rempart de la Vierge 8, B-5000 Namur, Belgium

Abstract. All the clustering methods try to reach the same goal : to establish whether or not objects of given set fall naturally into groups (or classes, or clusters) such that objects in the same group are similar to one another and different from objects in other groups. Hierarchical agglomerative clustering provides a wide family of clustering methods. They perform as follow. At the start, all the objects form their own singleton classes. Then at each step, the most similar pair of existing classes are merged.The process stops when it only remains a single class containing all objects.

The crucial problem with agglomerative methods lies in choosing the right way to define the dissimilarity between objects and between classes. Traditional methods based on euclidean or mahalanobis distances are not efficient nor reasonable when the dimension grows.

We propose a new and very original way to define the dissimilarity between objects. We postulate that the dissimilarity between objects must depend on the Lebesgue measure of the hypervolume sustended by the multidimensional density between objects. To avoid computational problems, two objects will be said to be similar if all the sustended areas (on each axes) between the respective coordinates of the two objects are small. In this case indeed, the hypervolume sustended by the multidimensional density will be small. This assertion not reversible.

Using this original dissimilarity measure, a new agglomerative clustering algorithm was developped. At the same time, the new dissimilarity measure was combined with a k nearest neighboor algorithm and a new discriminant analysis algorithm was achieved.

The clustering algorithm and the discriminant analysis one were applied with succes in the two following domains : the early prediction of entreprise bankruptcy and the image analysis. In some cases , stratification before discriminant analysis carried out a 20% gain of well classified data.

We present here this new efficient rule which can be said 'counter-intuitive'. We developp why this dissimilarity seems to be natural and we define it. We also explain how empirical distribution functions can be used to make considerable computational improvements. As examples, we show how the new clustering algorithm can be applied to interval data when the middle and half-lenght representation is used.

References

BOCK, H.H. and DIDAY, E (2000): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Springer-Verlag.

CHAVENT, M. and LECHEVALLIER, Y. (2002): Dynamical clustering of interval data: Optimization of an adequacy criterion based on hausdorff distance. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Data Analysis*. Springer, Berlin, 53–60.

RASSON, J.-P. and PIRÇON, J.-Y. and ROLAND, F (2003): Stratification before Discriminant Analysis : A Must ? Proceedings of the 27th Annual Conference of th GfKl. In: D. Baier and K.-D. Wernecke: *Innovations in Classification, Data Science, and Information Systems* 54–60.

Keywords

HIERARCHICAL AGGLOMERATIVE CLUSTERING, DISSIMILARITY, DENSITY ESTIMATION, DISCRIMINANT ANALYSIS, INTERVAL DATA

Simultaneous Models for Clustering and Reduction

Roberto Rocci[1] and Maurizio Vichi[2]

[1] Department SEFeMEQ, University of "Tor Vergata",
Via Columbia 2, 00133 Roma, ITALY
[2] Department of Statistics, Probability and Applied Statistics,
University "La Sapienza", P.le A.Moro 5, 00185 Roma, ITALY

Abstract. In many applications principal component analysis is applied before clustering to reduce the dimension of the data. This sequential procedure, named "tandem analysis", has been criticized by several authors (see for example Chang, 1983), because the reduction step could remove some information about the clustering structure of the data. To overcome this problem, some authors proposed methodologies for simultaneous reduction and clustering of two-way data in the context of the least squares approach (Bock, 1987; De Soete & Carroll, 1994; Vichi & Kiers, 2001).
In this paper those proposals are compared with the sequential procedures and reinterpreted as simultaneous versions of them. They are also analyzed and compared from an empirical and theoretical point of view. From the comparisons emerges the possibility to build a unique loss function able to fit models ranging from *k*-means to principal component analysis passing trough several new and old simultaneous models.

References

BOCK, H.H. (1987): On the interface between cluster analysis, principal components, and multidimensional scaling. In: H. Bozdogan, A.J. Gupta (Eds.), *Multivariate statistical modelling and data analysis, Proceedings of Advances Symposium on Multivariate Modelling and Data Analysis, Knoxville, Tennessee, May 15-16, 1986*. Reidel Publishing Co., Dordrecht, 17–34.

CHANG, W. (1983): On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics, 32, 267–275*.

DE SOETE, G., CARROLL, J.D. (1994) : *k*-means Clustering in a Low-dimensional Euclidean Space. In: E. Diday et al. (Eds.): *New Approaches in Classification and Data Analysis*. Springer, Heidelberg, 212–219.

VICHI, M., KIERS, H.A.L. (2001): Factorial *k*-means analysis for two-way data. *Computational Statistics and Data Analysis, 37, 49–64*.

Keywords

PRINCIPAL COMPONENTS, CLUSTERS, SIMULTANEOUS CLUSTERING AND REDUCTION

Objects Grouping Based on Entropy

Iwona Roeske-Slomka

Department Statistics and Demography
University of Economics in Poznan
Poland

Abstract. Research objects can be for example companies, households or other units. Their activities or functioning are characterized by participation structure ratios which can reflect the production participation of individual goods in overall production or the expenses structure of households etc. Each object is characterized by the same number, the same type of participation structure ratios. The entropy concept can be determined by the following formula:

$$H_e = -\Sigma_i p_i log_2 p_i$$

where p_i means sequence structure ratio. Each object of research has its own determined empirical entropy value: H_e. The top limit which H_e can reach is determined by:

$$H_{max} = log_2 n$$

where n-number of structure ratios. The so-called grouping interval (k) is determined by the following formula:

$$k = \frac{{}_{max}H_e - {}_{min}\ H_e}{n}$$

To the first group will belong the objects which empirical entropy is not higher than: ${}_{min}H_e + k$; to the second than: ${}_{min}H_e + 2k$; to the third: ${}_{min}H_e + 3k$, etc. This methods is illustrated by poster presentation with reference to 12 hypothetical objects where each one is determined by four structure index. In consequence four groups of objects were separated. In the first group three objects were included. In the second four. In the third three and in the last: two.

Keywords

OBJECTS, RATIOS OF STRUCTURE PARTICIPATION, ENTROPY, INTERVAL OF GROUPING

Identifying and Classifying Social Groups: A Machine Learning Approach

Matteo Roffilli and Alessandro Lomi

University of Bologna, Italy

Abstract. The identification of social groups remains one of the main analytical themes in the analysis of social networks and, in more general terms, in the study of social organization. Traditional network approaches to group identification encounter a variety of problems when the data to be analyzed involve two-mode networks, i.e., relations between two distinct sets of objects with no reflexive relation allowed within each set. In this paper we propose a relatively novel approach to the recognition and identification of social groups in data generated by network-based processes in the context of two-mode networks. Our approach is based on a family of learning algorithms called Support Vector Machines (SVM). The analytical framework provided by SVM provides a flexible statistical environment to solve classification tasks, and to reframe regression and density estimation problems. We explore the relative merits of our approach to the analysis of social networks in the context of the well known "Southern women" (SW) data set collected by Davis Gardner and Gardner. We compare our results with those that have been produced by different analytical approaches. We show that our method, which acts as a data-independent preprocessing step, is able to reduce the complexity of the clustering problem enabling the application of simpler configurations of common algorithms.

References

BOORMAN S. and WHITE H. (1976): Social Structure from Multiple Networks II. Role Structures. *American Journal of Sociology 81: 1384-1446*.

BORGATTI S. P. and EVERETT M. G. (1997): Network analysis of 2-mode data. *Social Networks, 19/3, 243-269*.

BREIGER R. (1974): The Duality of Persons and Groups. *Social Forces, 53, 181-90*.

DOREIAN P., BATAGELJ V. and FERLIGOJ A. (2005): *Generalized blockmodeling*. Cambridge University Press.

VAPNIK V. (1995): *The Nature of Statistical Learning Theory*. Springer Verlag.

Keywords

TWO-MODE NETWORK DATA, BLOCKMODELING, NOVELTY DETECTION, MACHINE LEARNING, SVM

Dissimilarities for Web Usage Mining

Fabrice Rossi[1], Francisco De Carvalho[2], Yves Lechevallier[1], and Alzennyr Da Silva[1 2]

[1] Projet AxIS, INRIA Rocquencourt, Domaine de Voluceau,
Rocquencourt, B.P. 105, 78153 Le Chesnay cedex – France
[2] Centro de Informatica - CIn/UFPE
Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brasil

Abstract. The obtention of a set of homogeneous classes of pages according to the browsing patterns identified in web server log files can be very useful for the analysis of organization of the site and of its adequacy to user needs. Such a set of homogeneous classes is often obtained from a dissimilarity measure between the visited pages defined via the visits extracted from the logs. There are however many possibilities for defined such a measure. This paper presents an analysis of different dissimilarity measures based on the comparison between the semantic structure of the site identified by experts and the clustering constructed with standard algorithms applied to the dissimilarity matrices generated by the chosen measures.

References

CELEUX, G., DIDAY, E., GOVAERT, G., LECHEVALLIER, Y. and RALAMBONDRAINY H. (1989): *Classification Automatique des Données*. Bordas, Paris.

CHEN C. (1998): Generalized similarity analysis and pathfinder network scaling. *Interacting with Computers, 10, 107–128.*

GOWER, J. and LEGENDRE, P. (1986): Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification, 3, 5–48*.

HUBERT, L.and ARABIE, P. (1985): Comparing partitions. *Journal of Classification, 2, 193–218*.

ROSSI, F., EL-GOLLI, A. and LECHEVALLIER Y. (2005): Usage guided clustering of web pages with the median self organizing map. In: *Proceedings of XIIIth European Symposium on Artificial Neural Networks (ESANN 2005)*. Bruges (Belgium), 351–356.

TANASA, D. and TROUSSE, B. (2004a): Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems, 19(2), 59–65*.

TANASA, D. and TROUSSE, B. (2004b): Data preprocessing for wum. *IEEE Potentials, 23(3), 22–25*.

VAN RIJSBERGEN, C.J. (1979): *I*nformation Retrieval. Butterworths.

Keywords

CLUSTERING, DISSIMILARITY MEASURES, WEB USAGE MINING

Classification of Regional Labour Markets for Purposes of Research and of Labour Market Policy

Felix Rüb, Daniel Werner, and Katja Wolf

Institute for Employment Research (IAB),
Regensburger Str. 104, D-90478 Nuremberg, Germany

Abstract. In many countries labour market policy has to deal with fairly large and persistent regional labour market disparities. In the case of Germany, parts of the country are affected by a deep unemployment crisis whereas others show nearly full employment. Since these disparities cannot be reduced to only one dimension a classification system of labour markets was developed. The criterion of this system was the identification of the "regional disadvantage" for the success of labour market policy. This new classification is at the level of districts (NUTS 3 regions).

To optimise the results a two-step classification method was applied. The first step included regression analyses to identify the exogenous determinants of the success of labour market policy. In the second step, different types of labour markets are determined from a specific variant of cluster analysis which used the weighted variables identified as significant in the first step. This classification is used in the Federal Employment Agency for benchmarking reasons. Besides that, the new classification obtained could also be employed in research, for example in the evaluation of labour market policy.

References

BLIEN, Uwe; HIRSCHENAUER, Franziska; ARENDT, Manfred; BRAUN, Hans Jürgen; GUNST, Dieter-Michael; KILCIOGLU, Sibel; KLEINSCHMIDT, Helmut; MUSATI, Martina; ROSS, Herrmann; VOLLKOMMER,Dieter; WEIN, Jochen (2004): *Typisierung von Bezirken der Agenturen für Arbeit*. Zeitschrift für Arbeitsmarktforschung / Journal for Labour Market Research 2/2004,S. 146-175.

Keywords

REGIONAL LABOUR MARKETS, CLASSIFICATION

Assumptions Behind the Statistics In Authorship Attribution Studies: A Search for Valid Tests

Joseph Rudman

Department of English, Carnegie Mellon University,
Pittsburgh, Pennsylvania 15213, USA

Abstract. Controversy surrounds the methodology used in non-traditional authorship attribution studies (those studies that make use of the computer, stylistics, and statistics). (RUDMAN, 1998) (RUDMAN, 2003) One problem is that many of the most commonly used statistical tests have assumptions about the input data that do not hold for the primary data of these attribution studies, the textual data itself (e.g. normal distributions, randomness, independence). "...inappropriate statistical methods.... In particular, asymtotic normality assumptions have been used unjustifiably, leading to flawed results." (DUNNING, p. 61) (See also CLAYMAN, p. 386) "Assumptions such as the binomiality of word counts or the independence of several variables chosen as markers need checking." (MOSTELLER and WALLACE, p. 280)

This paper looks at some of the more frequently used tests (e.g. chi-square) and then at the questions, "Are there assumptions behind various tests that do not appear in the studies?" and "Does the use of statistical tests whose assumptions are not met invalidate the results?" After presenting the disagreements in the literature, I hope to engage the house in discussing and coming to a consensus on these disagreements.

References

CLAYMAN, D.L. (1992): Trends and Issues in Quantitative Stylistics. *Transactions of the American Philological Association (1974-), 122, 385–390*.

DUNNING, T. (1993): Accurate Methods for the Statistics of Suprise and Coincidence. *Computational Linguistics, 19.1, 61–74*.

MOSTELLER, F. and WALLACE, D.L. (1984): *Applied Bayesian and Classical Inference: The Case of* The Federalist *Papers*. Springer-Verlag, New York.

RUDMAN, J. (1998): The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities, 31, 351–365*.

RUDMAN, J. (2003): Cherry Picking in Nontraditional Authorship Attribution Studies. *Chance, 16.2, 26–32*.

Keywords

ASSUMPTIONS, AUTHORSHIP ATTRIBUTION, LINGUISTICS, STATISTICAL METHODS, STYLISTICS

A Local Maximum Likelihood Estimator for Zero-inflated Count Data

José António Santos[1] and M. Manuela Neves[2]

[1] ISEGI, New University of Lisbon,
Campus de Campolide, 1070-312 Lisboa, Portugal
[2] Department of Mathematics, ISA, Technical University of Lisbon,
Tapada da Ajuda, 1349-017 Lisboa, Portugal

Abstract. The Poisson regression is the basic framework to deal with count data. An overwhelming property of this model is equaldispersion.
Likelihood-based smothers were founded on the penalized likelihood concept. Then was suggested the local likelihood concept which bring about a more robust and smoother estimator and extended the nonparametric regression analysis to maximum likelihood-based regression models. Later on this concept was extended to the kernel smoothing and local polynomial kernel regression framework.
A local maximum likelihood estimator based on Poisson regression is presented as well as its asymptotic bias, variance and distribution. This semi-parametric estimator is intended to be an alternative to the Poisson and Zero-inflated Poisson regression models that does not depend on regularity conditions and model specification accuracy. Some simulation results are presented based on Poisson and zero-inflated Poisson simulated data as well as some real data cases.

References

FAN, J., FARMEN, M. and GIJBELS, I. (1998): Local Maximum Likelihood Estimation and Inference. *Journal of the Royal Statistical Society B, 60, 591–608*.

FAN, J., HECKMAN, N. and WAND, M. (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-likelihood Functions. *Journal of the American Statistical Association, 90: 141-50*.

STANISWALIS, J. (1989). The Kernel Estimate of a Regression Function in Likelihood-based Models. *Journal of the American Statistical Association, 84: 276-83*.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association, 81: 96-103*.

TIBSHIRANI, R. and HASTIE, T. (1987). Local Likelihood Estimation. *Journal of the American Statistical Association, 82: 559-67*.

Keywords

CONFIDENCE INTERVALS, COUNT DATA, LOCAL MAXIMUM LIKELIHOOD, OVERDISPERSION, POISSON REGRESSION, SEMI-PARAMETRIC REGRESSION, SIMULATION, ZERO-INFLATED COUNT DATA

Clustering for Mixed Data Using Sphereical Representation

Yoshiharu Sato

Division of Computer Science,
Graduate School of Information Science and Technology,
Hokkaido University,
Kita-ku, Kita 14, Nishi 9, 060-0814 Sapporo, Japan

Abstract. In the data whose attribute values can be mixed binary and continuous (say mixed data), traditional cluster analysis has the problem in mixture of the distance between binary data and the distance between continuous data. The author has been proposed a transformation binary data into a directional data (spherical representation of binary data), in the previous conference, from the considerations of the similarity measures and the natural definition of the descriptive measures.
Additionally, in this report, we propose the method for spherical representation of q-dimensional continuous data such that the q-dimensional spherical distance relation correspond to the distance relation between continuous data as much as possible. In this case, the region of the q-dimensional sphere is confine to positive quadrant. Then the original region of q-dimensional continuous data should be scaling properly (scaling up or down) This method is basically identical with the metric multidimensional scaling on a sphere.
Then, the mixed data are represented on the sphere, that is, an object whose attributes takes mixed values is represented as a directional data. Using the descriptive measures on a sphere, we can execute usual k-means method for mixed data. Finally, the performance of this clustering is evaluated by the concrete data.

References

MACQUEEN, J. (1967): Some methods for classification and analysis of multivariate observations. In: L.M.Le Cam & J.Neyman (Eds.): *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.

MARDIA, K. (1972): *Statistics of Directional Data*. Academic Press.

Keywords

K-MEANS METHOD, MULTIDIMENSIONAL SCALING, SPHERICAL DESCRIPTIVE MEASURES

A Comparison of Distance Measures for Clustering Time–Course Microarray Data

Theresa Scharl[1] and Friedrich Leisch[2]

[1] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10/1071, A-1040 Vienna, Austria
[2] Department of Statistics, University of Munich, Ludwigstraße 33, D-80539 München, Germany

Abstract. Clustering time–course microarray data is an important tool to find co–regulated genes and groups of genes with similar temporal or spacial expression patterns. Depending on the distance measure and cluster algorithm used different kinds of clusters will be found.

In a simulation study on various datasets the influence of cluster algorithm and distance measure used is investigated. The quality–based cluster algorithm QT–Clust (Heyer et al. (1999)) is compared to a stochastic variant of the original algorithm (Leisch (2006)) using different distance measures. For that purpose the stability and sum of within cluster distances are evaluated. Finally stochastic QT–Clust is compared to the well–known k–means algorithm. The main focus is not to find the "best" combination of cluster algorithm and distance measure but to get a deeper understanding for what is going on when different methods are used.

References

HEYER, L.J., KRUGLYAK, S. and YOOSEPH, S. (1999): Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research, 9, 1106-1115*.

LEISCH, F. (2006): A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis, Accepted for publication*.

Keywords

CLUSTER ANALYSIS, DISTANCE MEASURES, TIME–COURSE MICROARRAY DATA

Revealing the Nature of Interactions in Two-Way Two-Mode Data by Constrained Simultaneous Clustering Models

Jan Schepers and Iven Van Mechelen

Department of Psychology, Katholieke Universiteit Leuven,
Tiensestraat 102, B-3000 Leuven, Belgium

Abstract. In several domains of science, researchers often deal with two-way two-mode data that are considered to be values of a criterion variable, for all combinations of the levels of two categorical predictor variables. A major challenge then is to capture the nature of the interaction between the row and column variables in the prediction of the criterion. In this context the distinction between several, qualitatively different, types of interactions can be most relevant, including that between disordinal, single ordinal and double ordinal interactions (Mokken & Lewis, 1982, Hager & Westermann, 1983). The identification of the exact nature of the interaction, however, often is troubled by the complexity of the data (usually the number of rows and/or columns is large). In this paper, we will discuss constrained versions of a recently introduced two-mode clustering model (Schepers & Van Mechelen, 2005) that implies a categorical reduction of both the rows and the columns of a data matrix, the constraints being such that different types of (dis)ordinal interactions can be captured. By comparing the fit of models that include different constraints, the nature of the interaction present in the data may be revealed.

References

SCHEPERS, J., and VAN MECHELEN, I. (2005): Hierarchical classes modeling of real-valued data. *Paper presented at IMPS 2005, Tilburg, The Netherlands*.

MOKKEN, R.J., and LEWIS, C. (1982): A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6, 417–430*.

HAGER, W., and WESTERMANN, R. (1983): Ordinality and disordinality of statistical interaction in two-way ANOVA. *Archiv für Psychologie, 135, 341–359*.

Keywords

TWO-MODE CLUSTERING, TYPES OF INTERACTIONS

Mining Association Rules in Folksonomies

Christoph Schmitz[1], Andreas Hotho[1], Robert Jäschke[1,2], Gerd Stumme[1,2]

[1] Knowledge & Data Engineering Group, Department of Mathematics and Computer Science, University of Kassel, Wilhelmshöher Allee 73, D–34121 Kassel, Germany, http://www.kde.cs.uni-kassel.de
[2] Research Center L3S, Expo Plaza 1, D–30539 Hannover, Germany, http://www.l3s.de

Abstract. Social bookmark tools are rapidly emerging on the Web. In such systems users are setting up lightweight conceptual structures called folksonomies. These systems provide currently relatively few structure. We discuss in this paper, how association rule mining can be adopted to analyze and structure folksonomies, and how the results can be used for ontology learning and supporting emergent semantics. We demonstrate our approach on a large scale dataset stemming from an online system.

Keywords

FOLKSONOMIES, SOCIAL BOOKMARK SYSTEMS, TAGGING, DATA MINING, ASSOCIATION RULES, EMERGENT SEMANTICS, ONTOLOGY LEARNING, WEB MINING

Classification in Marketing Science

Sören W. Scholz, Ralf Wagner, and Reinhold Decker

Business Administration and Marketing, Bielefeld University
Universitätsstraße 25, 33615 Bielefeld, Germany

Abstract. Pawlak (1982) claims that knowledge by itself is deep-seated in the classificatory abilities of human beings in the domain of marketing science.

This paper explores the role of clustering and classification algorithms in marketing beyond the prominent application of market segmentation (c.f. Wagner et al. (2005)). The study is based on a sample of more than 1,900 articles chosen from international academic marketing journals. We apply a hierarchical self organizing map and a bi-secting k-means (cf. Scholz and Decker (2005)) for text classification to show the relation between the use of cluster algorithms and the implications for practitioners argued in the above articles. Moreover, we evaluate the clusters by means of annotating quality ratings of documents assigned to each cluster.

The study highlights the application gaps of classification and clustering with respect to different areas of marketing science and, therefore, hints to weak empirical foundations. Implications for the marketing education (cf. Wagner (2005)) are drawn, which are relevant for both, the students when choosing their courses and the lecturers when designing or updating their course offers.

References

PAWLAK, Z. (1982): *Rhough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer, Dordrecht.

SCHOLZ, S.W. and DECKER, R. (2005): *Automated Topic Detection and Tracking in Environmental Scanning: Identifying Hot Spots in Marketing*. Manuscript under review.

WAGNER, R. (2005): Mining Promising Qualification Patterns. In: D. Baier and K.-D. Wernecke (Eds.): *Innovations in Classification, Data Science, and Information Systems*. Springer, Berlin, 249–256.

WAGNER, R., SCHOLZ, S.W., and DECKER, R. (2005): The Number of Clusters in Market Segmentation, in: D. Baier, R. Decker, and L. Schmidt-Thieme (Eds.): *Data Analysis and Decision Support*. Berlin, Springer, 157-176.

Keywords

BI-SECTING K-MEANS, HIERARCHICAL SOM, MARKETING SCIENCE, TEXT MINING, TOPIC DETECTION AND TRACKING

Generalisation Analysis as a Foundation for Classification

John Shawe-Taylor

Centre for Computational Statistics and Machine Learning
University College London
Gower Street, London WC1E 6BT
England

Abstract. The presentation will review frequentist approaches to assessing the performance of classifiers on new data. This provides a statistical view that rests on the relatively agnostic assumption that data are drawn independently from an unknown but fixed distribution. The analysis is concerned with assessing the generalisation, that is the performance of a classifier on unseen data drawn from the same distribution. The implications of these results for well-known learning algorithms such as Support Vector Machines will be reviewed. Extensions of the approach to subspace methods such as (kernel) principal components analysis will be discussed as well as some results for clustering algorithms. In both cases we are again concerned with the usefulness of the deduced simplifications when handling new data. The emphasis will be on the framework that the analysis provides for designing principled algorithms that infer underlying structure from samples of data.

Keywords

GENERALISATION, FREQUENTIST, SUPPORT VECTOR MACHINES, PRINCIPAL COMPONENTS ANALYSIS, CLUSTERING

Interpoint Distances

Alexander M. Shurygin

Department of Computer Mathematics and Cybernetics, Moscow State University,
Vorobjevy Gory, MSU, Moscow, Russia

Abstract. Let independent observations $x_1, \ldots, x_n \in Q \subseteq \mathbb{R}^p$, $p \geq 1$, in domain Q have distribution absolutely continuous with respect to the Lebesque measure. Then interpoint distances $r_{\alpha,\beta} = \|x_\alpha - x_\beta\|$, $\alpha < \beta = 2, \ldots, n$, between them are asymptotically (when $n \to \infty$) pairwise independent.

The fact formulated above was proved by the author. It gives possibility to consider interpoint distances instead of points and so to increase the sample size in $n/2$ times in some problems that can be reduced to interpoint distances. To such problems belong the goodness–of–fit tests, problems of classification and mathematical prediction of strong earthquakes. Solutions constructed with the help of interpoint distances are much more precise than solutions using points themselves.

References

SHURYGIN, A.M. (2005): Pattern Recognition by Interpoint Distances. *Transl. of Rep. Russian Academy of Sciences, 405/5*.

Keywords

INTERPOINT DISTANCES, CLASSIFICATION, MATHEMATICAL PREDICTION OF STRONG EARTHQUAKES

Theory of Stable Estimation

Alexander M. Shurygin

Department of Computer Mathematics and Cybernetics, Moscow State University,
Vorobjevy Gory, MSU, Moscow, Russia

Abstract. Let independent random variables $x_1, \ldots, x_n \in \mathcal{X} \subset \mathbb{R}^1$ have the probability density function (pdf) $f := f(x, \theta)$ dependent on unknown parameter θ that ought to be stably estimated. The minimum contrast estimator $\hat{\theta}$ is solution of the equation

$$\Sigma\psi(x_i, \hat{\theta}) = 0, \tag{1}$$

where $\psi(\cdot)$ is the score function. The limit distribution of $\sqrt{n}(\theta - \hat{\theta})$ is normal $N(0, V(f, \psi))$, where asymptotic variance

$$V(f, \psi) = \mathrm{E}\,\psi^2(x, \theta)/[\mathrm{E}\,\dot{\psi}(x, \theta)]^2$$

and $\dot{\psi}(x, \theta) = (\partial/\partial\theta)\psi(x, \theta)$.

The Lagrange derivative of the variance with respect to pdf can be a measure of instability of the estimator. It has minimum W_* at the estimator of maximum stability with score function

$$\psi_*(x, \theta) = c\,(\partial/\partial\theta)f(x, \theta) + \beta f(x, \theta), \quad \beta : \ \mathrm{E}\,\psi_* = 0.$$

The ratio $\mathrm{stb}\theta = W_*/W(f, \psi)$ is called stability of estimator $\hat{\theta}$ from (1), it varies from zero to one.

Considering two characteristics of an estimator, its efficiency and stability, we can choose an estimator with good both characteristics. Such is the radical estimator that has the score function

$$\psi_r(x, \theta) = [(\partial/\partial\theta)\log f(x, \theta) + \beta]\sqrt{f(x, \theta)}, \quad \beta : \ \mathrm{E}\,\psi_r = 0.$$

Functional

$$R(f, \psi) = [\int_{\mathcal{X}} \psi^2(x, \theta)\sqrt{f(x, \theta)}\,dx]/[\mathrm{E}\,\dot{\psi}(x, \theta)]^2$$

reaches its minimum R_* at the radical estimator. Radicalness of an estimator from (1) $\mathrm{rad}\hat{\theta} = R_*/R(f, \psi)$ is a convenient measure of estimator applied utility.

References

SHURYGIN, A.M. (1994): Variational Optimization of the Estimator Stability. *Automation and Remote Control, 55/11, 1611–1622* .

Keywords

ROBUSTNESS, STABILITY OF ESTIMATORS

A New Hierarchical Clustering Method based on Graph Theory

Helena Brás Silva[1], Paula Brito[2], and Joaquim Pinto da Costa[3]

[1] Department of Mathematics, Polytechnic School of Engineering of Porto (ISEP), Portugal
[2] School of Economics / LIACC, University of Porto, Portugal
[3] Department of Applied Mathematics / FC & LIACC, University of Porto, Portugal

Abstract. Using some concepts of graph theory, we propose a new hierarchical clustering method, where the dissimilarity measure between clusters is based on an index of evaluation of the relative isolation between clusters. Let $\Omega = \{x_1, x_2, \ldots, x_n\}$ be the data set of n elements to be clustered. We define the graph $G(V, E)$ on the data set, assigning to each vertex of V an element of the data set and there is an edge joining two different vertices of the graph if their dissimilarity is greater than some control parameter. Let $C_1, C_2, \ldots, C_k$ be k disjoint clusters and let $|C_i| = n_i$, the number of elements of cluster C_i, for $i = 1, \ldots, k$. The *index of evaluation of the relative isolation between clusters*, Δ_{ij}, measures the isolation between two clusters of a partition and is given by: $\Delta_{ij} = n_i n_j - \sum_{v \in C_i} d_G^{C_j}(v)$ where $d_G^{C_j}(v)$ is the degree of vertex $v \in C_i$ as concerns cluster C_j.

This relative isolation index between clusters may be used to define a dissimilarity measure between clusters. We define $\delta(C_i, C_j)$ as a *dissimilarity measure between clusters* as: $\delta(C_i, C_j) = \begin{cases} 0 & \text{if } i = j \\ 1 - \dfrac{\Delta_{ij}}{n_i n_j} & \text{if } i \neq j \end{cases}$ where Δ_{ij} is the relative isolation index between clusters C_i e C_j. Using this dissimilarity measure as aggregation index, we define a new ascending hierarchical clustering algorithm. In particular, this dissimilarity between clusters is used to define a hierarchical based optimization of the clustering method proposed in BRÁS SILVA, H. et.al.. This approach is illustrated by some simulated and real examples.

References

BRÁS SILVA, H., BRITO, P. and PINTO da COSTA, J.: A Partitional Clustering Algorithm Validated by a Clustering Tendency Index based on Graph Theory. *Pattern Recognition.* To appear.

Keywords

UNSUPERVISED LEARNING, HIERARCHICAL CLUSTERING, DISSIMILARITY BETWEEN CLUSTERS

Using MCMC as a Stochastic Optimization Procedure for Musical Time Series

Katrin Sommer and Claus Weihs

Department of Statistics, University of Dortmund,
D-44221 Dortmund, Germany

Abstract. Based on a model of Davy and Godsill (2002) we describe a general model for time series from monophonic musical sound to estimate the pitch. The model is a hierarchical Bayes Model which will be estimated with MCMC methods. All the parameters and their prior distributions are motivated individually. For parameter estimation an MCMC based stochastic optimization is introduced. In a simulation study it will be looked for the best implementation of the optimization procedure.

References

DAVY, M. and GODSILL, S.J. (2002): Bayesian Harmonic Models for Musical Pitch Estimation and Analysis. *Technical Report 431*, Cambridge University Engineering Department.

LIGGES, U., WEIHS, C., HASSE-BECKER, P. (2002): Detection of Locally Stationary Segments in Time Series. In: Härdle, W., Rönz, B. (Hrsg.): *COMPSTAT 2002 - Proceedings in Computational Statistics - 15th Symposium held in Berlin, Germany*. Heidelberg: Physica, 285–290.

McGill University Master Samples. McGill University, Quebec, Canada. `http://www.music.mcgill.ca/resources/mums/html/index.htm`

WEIHS, C. and LIGGES, U. (2006): Parameter Optimization in Automatic Transcription of Music. In Spiliopoulou, M., Kruse, R., Nürnberger, A., Borgelt, C. and Gaul, W. (eds.): *From Data and Information Analysis to Knowledge Engineering*. Springer, Berlin, 740 – 747.

Keywords

MCMC, STOCHASTIC OPTIMIZATION PROCEDURE, MUSICAL TIME SERIES

Adjusting Incorrect Classifications of Items in Subtests: Oblique Multiple Group Method or Confirmatory Common Factor Method

Ilse Stuive, Henk A.L. Kiers, Marieke E. Timmerman, and
Jos M.F. ten Berge

Heymans Institute (DPMG)
University of Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
The Netherlands

Abstract. Psychological tests often consist of items that can be meaningfully classified in subtests. Confirmatory factor analysis is an often used method to verify whether or not an expected classification is supported by the observed data. Frequently, the expected classification is not supported by the data, meaning that the classification used is probably incorrect. Researchers then often want to adjust the currently used classification such that a different classification of items in subtests is obtained which is supported by the data.

In the current study, two confirmatory factor analysis methods are compared on the quality of their suggestions to adjust incorrect classifications. Of the two methods, the confirmatory common factor method is the most often used in practice. However, previous research reported a poor quality of the suggested adjustments provided by this method. Therefore, this method will be compared with a less often used but promising method, the oblique multiple group method. To investigate the quality of the suggestions by these methods, simulated data were constructed for which the Şcorrectť classification was known. After using an incorrect classification on the data, it was investigated whether or not the adjustments suggested by both methods resulted in the recovery of the Şcorrectť classification.

Keywords

CONFIRMATORY FACTOR ANALYSIS, INCORRECT CLASSIFICATION, OBLIQUE MULTIPLE GROUP METHOD, CONFIRMATORY COMMON FACTOR METHOD

The Classification of Journals in Citation Database of Japanese Papers (CJP) by Keyword Analysis

Yuan Sun and Masaki Nishizawa

National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8640, Japan

Abstract. For policy analyses or research evaluation, citation database is a useful tool and its information about the size or impact of R &D activity in scientific disciplines or fields is especially essential sometime. In this study, a new scheme for subject classification of source journals in Citation Database of Japanese Papers (CJP) by keyword analysis is described that could update old classification which has been used several decades and does somewhat not suit the on-going situation. CJP is produced by National Institute of Informatics (NII), Japan, and contains 916,735 papers from approximately 1590 Japanese academic society journals in science, engineering, agriculture and medicine during 1995-2004. In this study, we make a collection of characteristic keywords of each journal based on authorsÕ keywords of each paper appeared in that journal. By comparing characteristic keyword groups of the journals, we reckon the similarities between journals and classify journals furthermore. At the same time, we apply the same approach to specify the characteristic keywords of each research field used for the Grants-in-Aid for Scientific Research in Japan, and classify the journals of CJP into research categories of the Grants-in-Aid. The results will be shown and discussed further. We expect this classification scheme to allow journals in CJP or other databases to be classified rapidly and transparently, for changes in their research level with time.

References

M. NEGISHI, Y. SUN, K. SHIGI (2004): Citation database for Japanese papers: a new bibliometric tool for Japanese academic society. *Scientometrics, Vol.60, No.3, pp.333–351*.

Y. SUN, M. NISHIZAWA (2002): Investigation into Genome-related Research at Japanese Universities by Keyword Analysis. *the 8th Conference of the International Federation of Classification Societies Societies on Data Science Classification and Related Methods. Cracow, Poland: IFCS 2002, July 16-19, p.171*.

Keywords

CITATION DATABASE, SOURCE JOURNAL, CLASSIFICATION, KEYWORD ANALYSIS

Simulation Study of Asymmetric K-Medoids Clustering Algorithms for Dissimilarity Data

Akinobu Takeuchi[1] and Hiroshi Yadohisa[2]

[1] Department of Humanities and Social Scicences, Jissen Women's University, Tokyo 191-8510, JAPAN
[2] Department of Culture and Information Science, Doshisha University, Kyoto 610-0349, JAPAN

Abstract. The k-means clustering algorithm is usually used for analyzing multivariate data, which can be represented in the Euclidean space. On the other hand, the k-medoids clustering algorithm is proposed for analyzing dissimilarity data that only contains information on the dissimilarities between objects. The k-medoids algorithm is determined depending on the definitions of the 1) initial cluster centers, 2) dissimilarity between the object and the cluster center, and 3) determination of the cluster centers, in the same manner as that for the k-means algorithm.

As an extension of the k-medoids algorithm, Takeuchi and Yadohisa (2006) proposed new asymmetric k-medoids algorithms for dissimilarity data. The asymmetric algorithms are also defined depending on 1), 2), and 3). In addition, the asymmetry of the data causes to extend an arbitrariness to define the algorithms. In such cases, the dissimilarities between two objects, which is expressed by the definitions of 1), 2), and 3), are asymmetric; therefore, the definitions of 1), 2), and 3) should be modified to consider the asymmetry between the objects. In such a situation, the algorithm can be defined in many ways.

In this paper, we evaluate the asymmetric k-medoids clustering algorithm by simulation studies. Artificial data are created with a) different types of data dispersion, b) varying numbers of clusters, and c) different degrees of asymmetry. The criteria for evaluation are i) the sum of the squared error ratios, ii) Spearman's rank correlation coefficient, and iii) the cophenetic correlation coefficient for the asymmetric clustering algorithms.

References

TAKEUCHI, A. and YADOHISA, H. (2006): Asymmetric k-medoid clustering algorithms for dissimilarity data: Abstracts of *2nd German Japanese Symposium on Classification ADVANCES IN DATA ANALYSIS AND RELATED NEW TECHNIQUES & APPLICATIONS*.

Keywords

ASYMMETIC DISSIMILARITY DATA, CLUSTERING, k-MEANS

Application of Categorical Principal Component Analysis to the Classification of Antiretroviral Drug Usage

Shinobu Tatsunami[1], Rie Kuwabara[1], Masashi Taki[1], Junichi Mimaya[2], and Akira Shirahata[3]

[1] Collaboration of Medical Statistics Unit, Radioisotope Reseach Institute and Department of Pediatrics, St. Marianna University School of Medicine, Kawasaki, Japan 216-8511
[2] Division of Hematology and Oncology, Children's Hospital of Shizuoka, Shizuoka, Japan 420-8660
[3] Department of Pediatrics, University of Occupational and Environmental Health, Japan, Kitakyushu, Japan 807-8555

Abstract. We analyzed the combinations of drugs used in antiretroviral therapy. Data of drug usage were obtained from 556 Japanese patients with coagulation disorders (Tatsunami et al, 2002) that were treated between 2000 and 2004. We used numbers of nucleoside reverse transcriptase (RT) inhibitors (NRTI), non-nucleoside RT inhibitors (NNRTI), and protease inhibitors (PI). Three dichotomous variables, meaning the usage/non-usage of combined drugs, usage of only NRTs, and usage of only two drugs, were also included. Variables were subjected to categorical principal component analysis (CATPCA). Annual changes in the pattern of drug combinations were expressed as scatter plots of object points from CATPCA in bubble charts. A total 40 combination patterns were identified, and CATPCA extracted four major patterns. A noticeable time-series change was the increasing usage of one combined drug containing two PIs with two drugs of NRTs. Use of this combination was not observed in 2000, however, its annual fraction of use from 2001 to 2004 was 2%, 8%, 13% and 14%, respectively. Values of medical markers among patients receiving this combination in 2004 were satisfactory. No other remarkable changes were observed over the five years. With increasing applicable drugs, a numerical classification of combination patterns should prove useful.

References

TATSUANMI S., TAKI M., SHIRAHATA A., MIMAYA J. and YAMADA K. (2003): Number of People in Japan with Coagulation Disorders: 2001 Update, International Journal of Hematology, 77/1, 96-98.

Keywords

CATPCA, HIV, THERAPY

A One-sided View of Classification: A Food Science Perspective

Deirdre Toher[1,2], Gerard Downey[1], and Thomas Brendan Murphy[2]

[1] Teagasc, Ashtown Food Research Centre,
Dublin 15, Ireland
[2] Department of Statistics, School of Computer Science and Statistics,
Trinity College Dublin, Dublin 2, Ireland

Abstract.
In food authenticity the main concern is whether a food is as described or not. If the food is not what it claims to be, its actual composition or origin is often of little concern, or the sample would be sent for further, more expensive and time consuming, testing. As such, a one-sided view of classification is taken: only one group is of actual interest.

The group of interest is modelled using model-based classification techniques that impose a parsimonious structure on the covariance matrix. All other observations are modelled as Poisson noise, as such observations may be from a wide variety of possible groups.

Variable selection is incorporated into the modelling procedure by using 5-fold cross validation on Brier's score values. A greedy selection process, whereby the variable with the lowest Brier's score is added at each stage, is implemented. This is supplemented by a variable removal process.

A one-sided approach as described above gives sufficient flexibility when samples from outside the group of interest may be extremely diverse. This method has been applied to a variety of food science applications. Forina *et al.* (1982) olive oil data is used as an example of the application of this procedure.

References

FORINA, M., ARMANINO, C., LANTERI, S. and TISCORNIA, E. (1982): Classification of olive oils from their fatty acid composition. In: H. Martens and H. Russwurm Jr (Eds.): *Food Research and Data Analysis* (1983). Applied Science, London, 189–214.

Keywords

FOOD AUTHENTICITY, MODEL-BASED CLUSTERING, ONE-SIDED CLASSIFICATION

Empirical Analysis of Attribute-Aware Recommendation Algorithms with Variable Synthetic Data

Karen H. L. Tso and Lars Schmidt-Thieme

Computer-based New Media Group (CGMN),
Department of Computer Science, University of Freiburg,
George-Köhler-Allee 51, 79110 Freiburg, Germany
{tso,lst}@informatik.uni-freiburg.de

Abstract. Recommender Systems (RS) have helped achieving success in E-commerce. Delving better RS algorithms has been an ongoing research. However, it has always been difficult to find adequate datasets to help evaluating RS algorithms. Public data suitable for such kind of evaluation is limited, especially for data containing content information (attributes). Previous researches have shown that the performance of RS rely on the characteristics and quality of datasets. Although, a few others have conducted studies on synthetically generated data to mimic the user-product datasets, datasets containing attributes information are rarely investigated. In this paper, we review synthetic datasets used in RS and present our synthetic data generator that considers attributes. Moreover, we conduct empirical evaluations on existing hybrid recommendation algorithms and other state-of-the-art algorithms using these synthetic data and observe the sensitivity of the algorithms when varying qualities of attribute data are applied to the them.

References

HERLOCKER, J., KONSTAN, J., BORCHERS, A., and RIEDL, J. (1999): An Algorithmic Framework for Performing Collaborative Filtering. In: ACM SIGIR'99. ACM press.

MARLIN, B., ROWEIS, S. and ZEMEL, R. (2005): Unsupervised Learning with Non-ignorable Missing Data. In: 10th International Workshop on Artificial Intelligence and Statistics, 222-229.

SARWAR, B.M., KARYPIS, G., KONSTAN, J.A. and RIEDL, J. (2000): Analysis of recommendation algorithms for E-commerce. In: 2nd ACM Conference on Electronic Commerce. ACM, New York. 285-295.

TSO, H.L.K. and SCHMIDT-THIEME, L. (2005): Attribute-Aware Collaborative Filtering. In: 29th Annual Conference of the German Classification Society 2005, Magdeburg, Germany.

Keywords

RECOMMENDER SYSTEMS, COLLABORATIVE FILTERING, SYNTHETIC DATA, ATTRIBUTE-AWARE, HYBRID RECOMMENDER SYSTEMS

An Overview of Mixed Data Analysis Based on Gower's Coefficient of Similarity

Lan Umek[1], Luka Bresciani[1], and Luka Kronegger[2]

[1] Graduate program on Statistics, University of Ljubljana,
Kongresni trg 10, SI-1000 Ljubljana, Slovenia
[2] Faculty of Social Sciences, University of Ljubljana,
Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia

Abstract. Several approaches for mixed data analysis exist. The most important contribution on the field of mixed data was given by J.C. Gower (1971). He introduced the first (dis)similarity measure that works on different types of variables. In the talk an overview of further developements of this measure will be presented. Special attention will be given to the development of variable weights (e.g. Cox and Cox, 2000) and to the use of updated versions of Gower's coefficient in clustering algorithms (e.g. Friedman and Meulman, 2004).

References

COX, T.F., COX, M.A. (2000): General weighted Two way Dissimilarity Coefficient. *Journal of Classification, 17, 101–121*.

FRIEDMAN, J.H., MEULMAN, J.J.(2004): Clustering objects on subsets of attributes. *Journal Of The Royal Statistical Society Series B, 66(4), 815–849*.

GOWER, J.C. (1971): A general coefficient of similarity and some of its properties. *Biometrics, 27, 857–871*.

Keywords

SIMILARITY COEFFICIENT, CLUSTERING, MIXED DATA

Simultaneous Clustering Methods Using the Max Operator: The Hierarchical Classes Approach

Iven Van Mechelen

Department of Psychology, Katholieke Universiteit Leuven,
Tiensestraat 102, B-3000 Leuven, Belgium

Abstract. Hierarchical classes models constitute a family of clustering models that imply a simultaneous overlapping clustering of the different sets of entities involved in a data set under study. The models imply a reconstruction of the data, making use of the Max operator. A distinctive characteristic of the models is that they include a representation of quasi-order (implication-type) relations for each of the sets of entities under study.
In this talk the hierarchical classes family will be briefly introduced, making use of a few illustrative examples. Subsequently, a state-of-the-art overview will be presented of research on the family that will include a summary of the most recent modeling and algorithmic developments.

References

CEULEMANS, E. and VAN MECHELEN, I. (2005): Hierarchical classes models for three-way three-mode binary data: Interrelations and model selection. *Psychometrika, 70, 461–480.*

VAN MECHELEN, I., BOCK, H.-H. and DE BOECK, P. (2004): Two-mode clustering methods: A structural overview. *Statistical Methods in Medical Research, 13, 363–394*.

Keywords

BICLUSTERING, HIERARCHICAL CLASSES MODELS, MULTIWAY ANALYSIS, SIMULTANEOUS CLUSTERING

Biclustering Methods for Microarray Gene Expression Data: Towards a Unifying Taxonomy

Iven Van Mechelen

Department of Psychology, Katholieke Universiteit Leuven,
Tiensestraat 102, B-3000 Leuven, Belgium

Abstract. To identify the biological mechanisms underlying microarray gene expression data, biclustering methods often have appeared to be most useful. Such methods yield data clusters that imply both a cluster of genes and a cluster of tissues or conditions, which may relate to underlying cellular processes, pathways or motifs. The biclustering domain, however, is very heterogeneous in several respects, and as such has never been easily accessible. As a way out for this problem, recently two structured overviews of biclustering methods have been proposed, one stemming from a bioinformatics perspective (Madeira and Oliveira, 2004), and one stemming from the perspective of traditional cluster analysis (Van Mechelen, Bock, and De Boeck, 2004). In this paper, I will integrate and extend these two structured overviews, which will yield a unifying taxonomy of biclustering methods. The basic principles underlying this taxonomy may facilitate the dialogue between data analysts/bioinformaticians and biological researchers in order to select, for a particular microarray gene expression data set at hand, suitable methods to retrieve biologically relevant biclusters.

References

MADEIRA, S.C. and OLIVEIRA, A.L. (2004): Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics, 1, 24–45*.

VAN MECHELEN, I., BOCK, H.-H. and DE BOECK, P. (2004): Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research, 13, 363–394*.

Keywords

MICROARRAY GENE EXPRESSION DATA, BICLUSTERING, UNIFYING TAXONOMY

Approximation of Globally Optimized Trees

Bart Jan van Os and Jacqueline J. Meulman

Department of Education, Data Theory Group, Leiden University,
P.O. Box 9555, 2300 RB Leiden, The Netherlands

Abstract. Since their introduction, classification and regression trees have acquired popularity in many scientific fields. Such trees are usually constructed by growing the tree from the top, while optimizing each new split locally given previously acquired splits. As a result, these trees may not be the best trees possible with respect to the misclassification rate or the residual sum of squared error in the terminal nodes.

Here, the optimization of a classification tree will be pursued through a global objective function defined on the classes formed by the terminal nodes, with size constraints on the tree, such as restricting the maximum depth of the tree or the maximum number of leaves. In practice, trees for different size restrictions are constructed in the same process and a cross-validation procedure is used to choose the best sized tree.

An exact exponential time Dynamic Programming algorithm for solving this problem has been proposed, that solves non trivial problems of, say, 10 predictors and a sample size of 500. Such an approach will be been shown to reduce generalized prediction error, sometimes while reducing the size of the best trees. For larger problems, however, the algorithm easily becomes infeasible. We will present strategies that reduce the search space, while upholding the global optimizing strategy, such that the resulting algorithm can be characterized as a heuristic approximation algorithm. In particular, the search space can be reduced by discretizing the predictor space and by doing a depth-first search for upper levels of the tree. Examples will be given.

References

BREIMAN, L AND FRIEDMAN, J. H., AND OLSHON, R. A. AND STONE, C. J. (1984): *Classification and Regression Trees*. Wadsworth, Belmont, CA.

VAN OS, B.J. (2001): *Dynamic Programming for Partitioning in Multivariate Data Analysis*. Leiden University.

Keywords

CLASSIFICATION TREES, GLOBAL OPTIMIZATION, DYNAMIC PROGRAMMING, EXACT ALGORITHM, HEURISTIC

Properties and Performances of Shape Similarity Measures

Remco C. Veltkamp[1] and Longin Jan Latecki[2]

[1] Dept. Computing Science, Utrecht University
Padualaan 14, 3584 CH Utrecht, The Netherlands
Remco.Veltkamp@cs.uu.nl
[2] Dept. of Computer and Information Sciences, Temple University
Philadelphia, PA 19094, USA
latecki@temple.edu

Abstract. This paper gives an overview of shape dissimilarity measure properties, such as metric and robustness properties, and of retrieval performance measures. Fifteen shape similarity measures are shortly described and compared. Since an objective comparison of their qualities seems to be impossible, experimental comparison is needed. The Motion Picture Expert Group (MPEG), a working group of ISO/IEC (see `http://www.chiariglione.org/mpeg/`) has defined the MPGE-7 standard for description and search of audio and visual content. A region based and a contour based shape similarity method are part of the standard. The data set created by the MPEG-7 committee for evaluation of shape similarity measures (Bober, 1999;Latecki, 2000) offers an excellent possibility for objective experimental comparison of the existing approaches evaluated based on the retrieval rate. Their retrieval results on the MPEG-7 Core Experiment CE-Shape-1 test set as reported in the literature and obtained by a reimplementation are compared and discussed. To compare the performance of similarity measures, we built the framework SIDESTEP – Shape-based Image Delivery Statistics Evaluation Project, `http://give-lab.cs.uu.nl/sidestep/`.

References

M. BOBER, J. D. KIM, H. K. KIM, Y. S. KIM, W.-Y. KIM, and K. MULLER. Summary of the results in shape descriptor core experiment. ISO/IEC JTC1/SC29/WG11/MPEG99/M4869, 1999.

L. J. LATECKI, R. LAKAEMPER, U. ECKHARDT. Shape descriptors for non-rigid shapes with a single closed contour. Proc. CVPR, 2000, 424-429.

Keywords

SHAPE SIMILARITY, EVALUATION

Exploratory Analysis of Uterine Electromyographic Data from Pregnant Sheep

Gaj Vidmar[1], Krešimir Matković[2], Branimir Leskošek[1], and Drago Rudel[1,3]

[1] IBMI, Faculty of Medicine, University of Ljubljana,
Vrazov trg 2, SI-1000 Ljubljana, Slovenia
[2] VRVis Research Center for Virtual Reality and Visualization, Ltd.,
Donau-City-Strasse 1, A-1220 Vienna, Austria
[3] MKS Electronic Systems Ltd.,
Rožna dolina c. XVII/22b, SI-1111 Ljubljana, Slovenia

Abstract. We present an overview of analyses of data from ten years of research (Rudel, 2002; Leskošek, Pajntar & Rudel, 1998) on uterine smooth-muscle activity in pregnant sheep as a model for humans. Electromiography (EMG) was performed at the horn and the cervix of the uterus in 35 sheep. The signals were processed in time and frequency domain yielding root-mean-square (RMS) and median frequency (MF) over one-minute periods as the data for further analyses. Research setup and addressed issues comprised normal course of EMG activity with approaching labor, effects of mild electric stimulation, effects of labor accelerating or decelerating medication, and EMG activity during and shortly before and after labor. The gathered data were difficult to analyze because of interrupted time series, large amounts of data from small number of subjects, huge intra- and inter-subject variability, low signal-to-noise ratio and lack of experimental control over potentially relevant factors. We employed various methods including pixelization-based (Levy, 2004) visualization of data quality and chronology of the entire research, 3D spectral plots (over time), robust descriptive graphics (boxplots of raw and aggregated data, local regression smoothing), linear modeling of transformed data (with mixed-model ANOVAs) and interactive visualization of large datasets (with the ComVis tool).

References

LESKOŠEK, B., PAJNTAR, M. and RUDEL, D. (1998): Time/frequency analysis of the uterine EMG in pregnancy and parturition in sheep. In: R. Magjarević (Ed.): *Biomedical Measurement and Instrumentation – BMI'98.* KoREMA, Zagreb, 3–106–109.

LEVY, P. (2004): The Case View, a Generic Method of Visualization of the Case Mix. *International Journal of Medical Informatics, 73/9-10, 713–718.*

RUDEL, D. (2002): In-Vivo Electrical Stimulation of Uterine Smooth Muscles in Sheep. 2$^{\mathrm{nd}}$ *European Medical & Biological Engineering Conference EMBEC02, IFMBE Proceedings, 3/1, 796–797.*

Keywords

EMG, LABOR, SHEEP, EDA, VISUALIZATION

Partitioning by Particle Swarm Optimization

Mario Villalobos-Arias[1] and Javier Trejos-Zelaya[2]

[1] CIMPA, Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica. Fax: +(506) 207 4397. E-Mail: mvillalo@cariari.ucr.ac.cr.
[2] Same address E-Mail: jtrejos@cariari.ucr.ac.cr.

Abstract. We deal with the problem of partitioning a set of objects described by p numerical variables, by the minimization of the within variance or inertia, using the *particle swarm optimization* (PSO) heuristics. The PSO is based on the use of a set of agents or particles that correspond to states of an optimization problem, and makes moves in a numerical space of the agents towards an optimal position, according to some rules. We propose a clustering algorithm using PSO, establishing a set of agents—partitions defined by a set of K centroids in a p–dimensional space; each centroid is associated to a cluster by the allocation of objects to the closest centroid. Centroids will move in the space according to the principles of the PSO and each partition is redefined by the allocation step. Hence, the PSO will consist on the move, of random intensity, of K centroids (or particles) for each p–dimensional agent on the direction of the overall best agent, limiting the move on the direction of a vector called velocity in order to avoid divergence. Results are compared to those obtained with other techniques, such as genetic algorithms, simulated annealing, tabu search, ant colonies, Grasp, k-means and Ward's hierarchical methods.

Keywords

AGENT–BASED OPTIMIZATION, HEURISTICS, COMBINATORIAL OPTIMIZATION, LOCAL MINIMA, CLUSTERING, CLASSIFICATION

Capturing Unobserved Heterogeneity in PLS Path Modeling

Vincenzo Esposito Vinzi[1] and Laura Trinchera[1,2]

[1] Department of Mathematics and Statistics, University of Naples "Federico II"
Via Cintia 26, Complesso Monte S. Angelo, 80126 Napoli, Italy
[2] Electricité de France, R&D, Clamart, France

Abstract. PLS (Partial Least Squares) approach to structural equation modeling (Tenenhaus et al. 2005) is used to estimate both impacts between adjacent latent variables and case-wise factor scores. A typical research issue is the identification of distinctive case segments by capturing the so called unobserved heterogeneity of individual behaviors. In the classical maximum likelihood approach, segmentation can be achieved, among others, by finite mixture models and the expectation-maximization (EM) algorithm (Jedidi et al. 1997). Data is generated by a mixture of two or more populations characterized by different covariance structures.

This solution seems inappropriate for PLS because of hard methodological assumptions incoherent with the soft modeling PLS spirit. Hahn et al. (2002) propose the finite mixture partial least squares (FIMIX-PLS) approach regarding the latent variable scores predicted in PLS. Conceptual and methodological problems arise and demand for further research and new methods (Esposito Vinzi et al. 2004): the measurement model is assumed to be the same for all groups, the multivariate normality for the estimated latent variables is assumed, the EM algorithm might fall in local optima. The obtained results will be assessed in terms of differences between yielded models and an ex post or implicit analysis will characterize the identified segments in terms of concomitant variables and membership probabilities.

References

ESPOSITO VINZI, V., LAURO, C. and AMATO, S. (2004): PLS Typological Regression: Algorithmic, Classification and Validation Issues. In: M. Vichi, P. Monari, S. Mignani and A. Montanari (Eds.): *New Developments in Classification and Data Analysis*. Springer–Verlag, Berlin, 133–140.

HAHN, C., JOHNSON, M., HERRMANN, A. and HUBER, F. (2002): Capturing Customer Heterogeneity using a Finite Mixture PLS Approach. *Schmalenbach Business Review, 54, 243–269*.

JEDIDI, K., HARSHANJEET, S.J. and DE SARBO, W.S. (1997): STEMM: A General Finite Mixture Structural Equation Model. *J. of Classif., 14, 23–50*.

TENENHAUS, M., ESPOSITO VINZI, V., CHATELIN, Y.M. and LAURO, C. (2005): PLS Path Modeling. *Comput. Stat. and Data Analysis, 48, 159–205*.

Keywords

PARTIAL LEAST SQUARES, STRUCTURAL EQUATION MODELS, RESPONSE BASED SEGMENTATION, TYPOLOGICAL REGRESSION

Comparing Optimal Individual and Collective Assessment Procedures

Hans J. Vos[1], Ruth Ben-Yashar[2], and Shmuel Nitzan[2]

[1] Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands
[2] Department of Economics, Bar-Ilan University, 52900 Ramat-Gan, Israel

Abstract. This paper focuses on the comparison between the optimal cutoff points set on single and multiple tests in predictor-based assessment, that is, assessing applicants as either suitable or unsuitable for a job. Our main result specifies the condition that determines the number of predictor tests, the collective assessment rule (i.e., aggregation procedure of predictor tests' recommendations) and the function relating the tests' assessment skills to the predictor cutoff points. The model advocated here has been applied earlier successfully by Ben-Yashar and Nitzan (2001) to economics where organizations face the comparable problem of deciding on approval or rejection of investment projects. A team of n decision makers has to decide which ones of a set of projects are to be accepted so as to maximize the team's common expected utility. The proposed collective assessment procedure is illustrated by an empirical example of assessing candidates as either suitable or unsuitable for a traineeship by means of the Assessment Center (AC) method. Optimal individual and collective predictor cutoff points will be compared.

References

BEN-YASHAR, R. and NITZAN, S. (2001): Investment Criteria in Single and Multi-Member Economic Organizations. *Public Choice, 109, 1–13*.

Keywords

OPTIMAL ASSESSMENT, INDIVIDUAL ASSESSMENT, COLLECTIVE ASSESSMENT, BAYESIAN DATA ANALYSIS

Patterns of Associations in Finite Sets of Items

Ralf Wagner

Business Administration and Marketing, Bielefeld University
Universitätsstraße 25, 33615 Bielefeld, Germany

Abstract. Mining association rules is well established in quantitative business research literature and makes up an up-and-coming topic in marketing practice. However, reducing the analysis to the assessment and interpretation of a few selected rules does not provide a complete picture of the data structure revealed by the rules.

This presentation introduces a new approach of visualizing relations between items by assigning them to a rectangular grid with respect to their mutual association. The visualization task leads to a quadratic assignment problem (Cela (1997)) and is tackled by means of a genetic algorithm (Blum & Roli (2003)). The methodology is demonstrated by evaluating a set of rules describing marketing practices in Russia (Wagner (2005)).

References

BLUM, C. and ROLI, A. (2003): Metaheuristics in Combinatorial Optimazation: Overview and Conceptual Comparision, *ACM Computing Survey, 35/3, 268-308*.

CELA, E. (1997): *The Quadratic Assignment Problem: Theory and Algorithms*, Kluwer, Dordrecht.

WAGNER, R. (2005): Contemporary Marketing Practices in Russia, *European Journal of Marketing, Vol. 39/1-2, 199–215*.

Keywords

ASSOCIATION RULES, GENETIC ALGORITHM, PATTERN RECOGNITION, VISUALIZATION

Local Models in Register Classification by Timbre

Claus Weihs, Gero Szepannek, Uwe Ligges, Karsten Luebke, and Nils Raabe

Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany
e-mail: weihs@statistik.uni-dortmund.de

Abstract. Investigating a data set containing different sounds of several instruments suggests that local modelling may be a promising approach to take into account different timbre characteristics of different instruments. For this reason, some basic ideas towards such a local modelling are realized in this paper yielding a framework for further studies.

Keywords

CLASSIFICATION, LOCAL MODELLING, MUSIC

CHIC-Model: A Global Model for Coupled Binary Data

Tom Wilderjans, Eva Ceulemans, and Iven Van Mechelen

Onderzoeksgroep kwantitatieve en persoonlijkheidspsychologie, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium

Abstract. Often problems result in the collection of coupled data, which consist of different N-way N-mode subdatasets that have one or more modes in common. To reveal the structure underlying such data an integrated modeling strategy, with a single set of parameters for the common mode(s), that is estimated based on the information in all subdatasets, may be most appropriate. Such a strategy implies a global model, consisting of different N-way N-mode submodels, and a global loss function that is a (weighted) sum of the loss functions associated with the different submodels. In this talk such a global model for an integrated analysis of a three-way three-mode binary data array and a two-way two-mode binary data array that have one mode in common is presented. A Simulated Annealing algorithm to estimate the model parameters is described and evaluated in a simulation study. An application of the model to real psychological data is discussed.

Keywords

COUPLED DATA, BINARY DATA, HIERARCHICAL CLASSES, MULTIWAY ANALYSIS

The Geographical Variation of Cervical Cancer Screening Efficiency in Slovenia

Vesna Zadnik, Tina Žagar, and Maja Primic Žakelj

Epidemiology and Cancer Registries, Institute of Oncology Ljubljana,
Zaloška 2, SI-1000 Ljubljana, Slovenia

Abstract. Organized cervical cancer screening programmes (OCCSP) aim to reduce cervical cancer incidence rate, as the disease could be diagnosed in the premalignant stage, and in addition, to increase the percentage of patients diagnosed in the early stage of disease. OCCSP was introduced in Slovenia recently. The intention of our study was to estimate the geographical variability of OCCSP efficiency in Slovenia.

Standardized incidence ratios for each of 58 SloveniaŠs administrative units (AU) were calculated for three time periods: 1996-1998 (no organized screening), 1999-2001 (pilot OCCSP in central and costal part of the country) and 2002-2004 (nationwide OCCSP). The effectiveness of screening is monitored by the ratio of patients diagnosed in the early stage and those diagnosed at later stages. As the data on early stage patients by 58 AU are scarce, the raw data were smoothed by hierarchical Bayesian model with a conditionally autoregressive prior.

Our study demonstrated the early benefit of the national OCCSP all over the Slovenia. The gain was more pronounced in the costal and central part of the country, where the pilot program of screening was introduced three years before the nationwide screening. In the north-eastern part more intensive promotion of the programme is needed.

References

SANKILA, R., DEMARET, E., HAKAMA, M., LYNGE, E., SCHOUTEN, L.J. and PARKIN, D.M. (Eds.) (2000): *Evaluation and Monitoring of Screening Programmes.* European Commission, Brussels-Luxembourg.

CARLIN, B.P. and LOUIS, T.A. (2000): *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, New York.

PRIMIC-ŽAKELJ, M., ZADNIK, V. and ŽAGAR, T. (2006): Ali se učinki programa ZORA že kažejo v incidenci invazijskega raka materničnega vratu. *Prvi slovenski kongres o cervikalni patologiji s kolposkopskim tečajem, 9-11. marec 2006, Kranjska gora.* Ljubljana: Združenje za ginekološko onkologijo, kolposkopijo in cervikalno patologijo: Ginekološka klinika, Klinični center, 2006.

Keywords

SPATIAL ANALYSIS, BAYESIAN SMOOTHING, CERVICAL CANCER, SCREENING ESTIMATION

Research Groups' Social Capital: A Clustering Approach

Petra Ziherl[1], Hajdeja Iglič[2], and Anuška Ferligoj[3]

[1] CATI d.o.o.,
Tržaška 2 , SI-1000 Ljubljana, Slovenia
[2] Faculty of Social Sciences, University of Ljubljana,
Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia
[3] Faculty of Social Sciences, University of Ljubljana,
Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia

Abstract. Our purpose in the article is to study the characteristics of research group's social capital. We proceed from the theoretical distinctions made in the literature on social capital, such as weak against strong ties, structural holes against cohesion and homogeneity against heterogeneity of a group. We assume that research group differ systematically according to the social capital they posses, which has an impact on their members' scientific performance.

Social capital of research groups is conceptualized in terms of complete networks. We use the data from the Slovenian study of academic research groups conducted in 2003/2004. Research groups include PhD students, their supervisors, and other researchers. They are representative of Slovenian research group, which include PhD students under the *junior researchers* program.

We explore the variation in the research groups' social capital by using clustering approach. The analysis reveals three types of research group's social capital: weak social capital, strong social capital of bonding kind, and strong social capital of bridging kind.

In the last part of the paper, we observe the scientific performance of PhD students in the obtained clusters. The results shows that students who are involved in research groups with bridging capital show significantly better performance than students in the groups with either bonding or weak social capital.

References

BURT, R.S. (1992): *Structural Holes*. MA: Harvard University Press, Cambridge.

GRANOVETTER, M.S. (1973): The strength of weak ties. *American Journal of Sociology, 78, 1360–1380.*

LIN, N. (1999): Building a Network Theory of Social Capital. *Connections, 22, 28–51.*

MCFADYEN, M.A. and CANNELLA, A.A. (2004): Social Capital and Knowledge Creation: Diminishing Returns of the Number and Strenght of Exchange Relationships. *Academy of Management Journal, 47, 735–746*.

Keywords

SOCIAL CAPITAL, RESEARCH GROUP, PhD STUDENTS

Clustering Time Varying Data of European Advertising Expenditures

Vesna Žabkar[1] and Katarina Košmelj[2]

[1] Faculty of Economics, University of Ljubljana, Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia
[2] Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, SI-1000 Ljubljana, Slovenia

Abstract. The advertising industry is an important component of economic activities of a certain country. European countries account for around one-quarter of the total global advertising spending (European Adspend Trends, 2005). Our study empirically examines the European advertising expenditure data for 29 European countries in the period from 1994 until 2004. Advertising expenditure on the country level includes aggregate value of advertising expenditure in the press (newspapers and magazines), television, radio, outdoor and cinema.

The objective of our work was to classify European countries according to similarity in advertising expenditure trends and to predict their behavior in near future. Analysis of advertising expenditure data across countries and longer time period is problematic due to differences in local currencies and exchange rates fluctuations during the examined time period. We overcame these problems considering a new variable: the ratio of adverting expenditure (Adspend) and gross domestic product (GDP), both were given in local currency and current prices. This variable presents the proportion of Adspend in GDP and allows comparison over countries and time.

The distance between two counties was calculated as the distance between two time-series (Košmelj, Batagelj, 1990). The time-dependent weights were derived from aggregate advertising expenditure growth rates in Europe in the time period examined. Two methodologies were applied: cluster analysis and multivariate scaling. Results reveal that we can present the European countries according to their time varying advertising expenditure levels in two dimensions, clustered in four groups with meaningful explanation. The analysis finds applicability in forecasting for aggregate advertising expenditure levels in specific countries.

References

European Adspend trends (2005). *International Journal of Advertising, 22, 149–152*.

KOŠMELJ K. and BATAGELJ, V. (1990): Cross-Sectional Approach for Clustering Time Varying Data. *Journal of Classification, 7/1, 99–109*.

World Marketing Data and Statistics (2006). www.euromonitor.com/womdas.

Keywords

TIME-DEPENDENT DISTANCE, GENERALIZED WARD METHOD, ADVERTISING EXPENDITURE

Evaluation of Generalized Blockmodeling and REGE on Regular Equivalence

Aleš Žiberna

Faculty of Social Sciences, University of Ljubljana,
Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia

Abstract. In the talk, the results of a comparison of different approaches to generalized blockmodeling (Doreian et. al, 2005; Žiberna, 2005a) and some versions of REGE (White, 2005; Žiberna, 2005a) on artificially generated valued networks that should be approximately ($max-$)regularly equivalent will be presented. All compared approaches are implemented in the R (R Development Core Team, 2005)package `blockmodeling` Őiberna(2006).

The networks were generated based on known ($max-$)regular blockmodel and partition using different beta distributions. The obtained partitions were compared to the original (known) partition using Rand index, adjusted for chance.

References

DOREIAN, P., BATAGELJ, V. and FERLIGOJ, A.(2006): *Generalized Blockmodeling*. Cambridge University Press, New York.

R Development Core Team (2005): R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. `http://www.R-project.org`.

WHITE, D.R.(2005): REGGE (web page). `http://eclectic.ss.uci.edu/~drwhite/REGGE/` (12.5.2005).

ŽIBERNA, A. (2005a): Generalized Blockmodeling of Valued Networks. *Unpublished paper (Presented on Sunbelt XXV Conference)*.

ŽIBERNA, A. (2005b): Direct and Indirect Methods for Regular Equivalence in Valued networks. *Unpublished paper (Presented on Applied Statistics 2005)*.

ŽIBERNA, A. (2006): `blockmodeling`: An R package for Generalized and classical blockmodeling of valued networks. `http://www2.arnes.si/~aziber4/blockmodeling/`.

Keywords

BLOCKMODELING, GENERALIZED BLOCKMODELING, REGE, REGULAR EQUIVALENCE, SIMULATIONS, VALUED NETWORKS

Authors

Key words

Social Program

A number of social events are scheduled for this year's meeting.

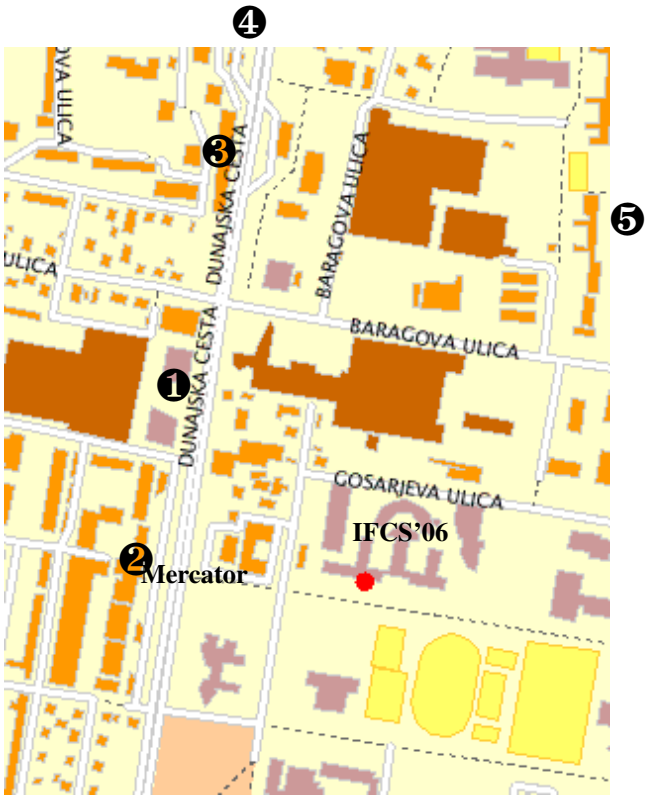
- July 25. 6:00-8:00 PM: Welcome reception at the Faculty of Social Science (Kardeljeva pl. 5, Ljubljana) (sponsored by SPSS).
- July 26. 6:30-8:00 PM: Reception given by prof.dr. Jure Zupan, the minister of higher education, science and technology (Villa Podrožnik)
- July 26. 8:00 PM: IFCS Council dinner, Restaurant Smrekarjev hram (for Council members only).
- July 27. 2:00 PM – approximately 9:00 PM: Half day excursion to lake Bled
- July 28. 7:00 PM: Conference dinner, Grand hotel Union (tickets 50.00 euro each, drinks included)

Conference dinner tickets can be purchased with registration. The cost of the other social program activities is included with registration for the meeting.

Recommended Restaurants

Around Faculty of Social Sciences

1. Self-Service Restaurant Glažuta
   Address: Dunajska 119, **Tel.:** + 386 (0)1 565 43 26, **Gsm:** + 386 (0)31 357 751
2. Self-Service Restaurant Mercator
   Address: Dunajska cesta 107
3. Špagetarija Favola, Italian food
   Address: Dunajska cesta 129, **Phone:** +386 (0)1 566 20 94
4. Sofra, Bosnian food
   Address: Dunajska cesta 145, **Phone:** +386 (0)1 565 68 00
5. Piramida
   Address: Vojkova 71, **Phone:** +386 (0)1 534 9881



Ljubljana Area

(€ denotes lower prices, €€ average prices, and €€€ higher prices)

5. Pri Škofu €
   Address: Rečna ulica 8, **Phone:** +386 (0)1 426 45 08
6. Pod Lipo €
   Address: Borštnikov trg 3, **Phone:** +386 (0)1 422 41 10
7. Pri Mraku €
   Address: Rimska cesta 4, **Phone:** +386 (0)1 421 96 00
8. Pod Rožnikom €
   Address: Cesta na Rožnik 18, **Phone:** +386 (0)1 251 34 46
9. Operna klet €
   Address: Župančičeva ulica 2, **Phone:** +386 (0)1 252 70 03
10. Foculus, pizzeria €
    Address: Gregorčičeva ulica 3, **Phone:** +386 (0)1 251 56 43
11. Napoli, pizzeria €
    Address: Prečna ulica 7
12. Ljubljanski dvor, pizzeria €
    Address: Dvorni trg 1, **Phone:** +386 (0)1 251 65 55
13. Gostilna Sokol €€
    Address: Ciril-Metodov trg 18, **Phone:** +386 (0)1 439 68 55
14. Pri Kovaču €€
    Address: Pot k Savi 9, **Phone:** +386 (0)1 537 12 44
15. Restavracija Grm €€
    Address: Hajdrihova ulica 16, **Phone:** +386 (0)1 425 25 00
16. Pen Club €€
    Address: Tomšičeva ulica 12, **Phone:** +386 (0)1 251 41 60
17. Hiša kulinarike Manna €€€
    Address: Eipprova ulica 1a, **Phone:** +386 (0)1 283 52 94
18. Gostilna As €€€
    Address: Čopova ulica 5a, **Phone:** +386 (0)1 425 88 22; +386 (0)1 252 72 37
19. Hana €€€
    Address: Tbilisijska ulica 85, **Phone:** +386 (0)1 256 14 02
20. Maxim €€€
    Address: Trg republike 1, **Phone:** +386 (0)1 476 69 80
21. JB Restaurant €€€
    Address: Miklošičeva cesta 17, **Phone:** +386 (0)1 433 13 59; +386 (0)1 474 72 19


Navje
Železna
Hacquetova
Vilharjeva
Robbova
DUNAJSKA
Smoletova
Parmova
Livarska
Likozarjeva
Kurilniška
Pivovarniška
Malgajeva
Jesenkova
Ruska
Medvedova
CELOVSKA
Trg Osvobodilne fronte
Masarykova
Tivolska
Gosposvetska
Slovenska
Dvořakova
Prežihova
Voglajnova
Kersnikova
Pražakova
Trdinova
Miklošičeva
Kolodvorska
Tavčarjeva
Argentinski park
Dalmatinova
Reslјeva
Slomškova
Metelkova
Tabor
TABOR
Kotnikova
Čufarjeva
Komenskega
Vidovdanska
Ilirska
Puharjeva
Štefanova
Župančičeva
Cankarjeva
Nazorjeva
Copova
Trubarjeva
Mala
Predna
Petkovškovo nabrežje
Ljubljanica
Poljanski nasip
Kapiteljska
Poljanska
Kristanova
Prešernov trg
Wolfova
Kongresni trg
Ciril-Metodov trg
Kopitarjeva
Prešernova
Šubičeva
Valvasorjeva
Tomšičeva
Beethovnova
Erjavčeva
Gregorčičeva
Igriška
Rimska
TRŽAŠKA CESTA
Aškerčeva
Mirje
Murnikova
Bogišičeva
Barjanska
Dvorni trg
Novi trg
Stari trg
Reber
Mestni trg
Cankarjevo nabr.
Gallusovo nabrežje
Levstikov trg
Gornji trg
Osojna pot
Ul. na Grad
Na Stolbi
Lončarska steza
Strossmayerjeva
Ulica 11. talcev
Ul. stare pravde
Pavla
Zarnikova
Streliška
C. slovenskih kmečkih uporov
Strmi pot
Karlovška
Zoisova
KRAKOVO
Krakovska
Krakovski nasip
Križevniška
Križanke
Gosposka
Emonska
Vrtna
Kladezna
Rečna
Ljubljanica
Grudnovo
Zvezdarska
Rožna
Hrenova
Vozarski pot
Žvonarska
Prijateljeva
Tesarska
Praprotnikova
Janežičeva
Merčnikova nabrežje
Cimpermanova
Žabjak
Gradaška
Elpprova
TRNOVO
Rožka
0
100
200 m

Sponsors