EUROPEAN CONFERENCE ON DATA ANALYSIS

Data Science: Foundations, Methods and Applications

2 – 4 September 2015

Book of Abstracts

bcs
British
Classification
Society

GfKl
Gesellschaft für
Klassifikation e.V.

TPS
Polskie
Towarzystwo
Statystyczne

University of Essex

European Conference on Data Analysis

Data Science: Foundations, Methods and Applications

September 2 – 4, 2015

University of Essex, Colchester, United Kingdom

Imprint

European Conference on Data Analysis 2015, Book of Abstracts

Editors: Berthold Lausen and Adi Florea

University of Essex
Department of Mathematical Sciences, Wivenhoe Park, Colchester CO4 3SQ, UK

Conference Chair: Berthold Lausen

Scientific Program Chairs:
Daniel Baier (*Gesellschaft für Klassifikation*), Andreas Geyer-Schulz (*Archives of Data Science*), Sabine Krolak-Schwerdt (*European Association for Data Science*), Fionn Murtagh (*British Classification Society*), Józef Pociecha (*Sekcjy Klasyfikacji i Analizy Danych PTS*)

Local Organisation Committee: Berthold Lausen (Mathematical Sciences, Essex), Adi Florea (Mathematical Sciences, Essex), Aris Perperoglou, (Mathematical Sciences, Essex), Hongsheng Dai (Mathematical Sciences, Essex), Andrew Harrison (Mathematical Sciences, Essex), Klaus McDonald-Maier (Computer Science & Electronic Engineering, Essex), Abdel Salhi (Mathematical Sciences, Essex)

Scientific Program Committee:
Daniel Baier (University of Bayreuth, Germany), Bernd Bischl (LMU Munich, Germany), Hampersum Bozdogan (University of Tennessee, USA), Carlos Cuevas Covarrubias (Universidad Anahuac, Mexico), Hongsheng Dai (University of Essex, UK), Reinhold Decker (Bielefeld University, Germany), Peter Flach (University of Bristol, UK), Andreas Geyer-Schulz (Karlsruhe Institute of Technology, Germany), Andrew Harrison (University of Essex, UK), Christian Hennig (University College London, UK), Irmela Herzog (The Rhineland Commission for Archaeological Monuments and Sites, Bonn, Germany), Eyke Hüllermeier (University of Paderborn, Germany), Krzysztof Jajuga (Wrocław University of Economics, Poland), Hans Kestler (University of Jena, Germany), Sabine Krolak-Schwerdt (University of Luxembourg, Luxembourg) Udo Kruschwitz (University of Essex, UK), Berthold Lausen (University of Essex, UK), Friedrich Leisch (University of Natural Ressources and Life Sciences, Vienna, Austria), Peter Lynn (University of Essex, UK), Klaus McDonald-Maier (University of Essex, UK), Fionn Murtagh (University of Derby, UK), Akinori Okada (Tama University, Tokio, Japan), Aris Perperoglou (University of Essex, UK), Anatoliy Pilyavskyy (Lviv Academy of Commerce, Ukraine), Józef Pociecha (Cracow University of Economics, Poland), Abdel Salhi (University of Essex, UK), João M. C. Santos Silva (University of Essex, UK), Willi Sauerbrei (University of Freiburg, Germany), Matthias Schmid (University of Bonn, Germany), Lars Schmidt-Thieme (University of Hildesheim, Germany), Frank Scholze (Karlsruhe Institute of Technology, Germany), Myra Spiliopoulou, (Otto-von-Guericke-Universität Magdeburg, Germany), Alfred Ultsch (University of Marburg, Germany), Ali Ünlü (Technical University Munich, Germany), Jeroen K. Vermunt (Tilburg University, Netherlands), Maurizio Vichi (University of Rome, Italy), Claus Weihs (Technical University Dortmund, Germany), Adalbert Wilhelm (Jacobs University Bremen, Germany), David Wishart (University of St. Andrews, UK), Taha Yasseri (University of Oxford, UK), Zhi-Hua Zhou (Nanjing University, China)

We are grateful for the support of our sponsors
Profusion (www.profusion.com)
Institute for Analytics and Data Science (www.essex.ac.uk/iads/)

EUROPEAN CONFERENCE ON DATA ANALYSIS 2015

Book of Abstracts

September 2 – 4, 2015
in Colchester
United Kingdom

Table of Contents

Welcome Note of the President of International Federation of Classification Societies

Dear Presidents of the *British*, *German* and *Polish Classification Societies*, distinguished Researchers, ladies and gentlemen, dear friends,

On behalf of the *International Federation of Classification Societies* it is a great privilege and an honor for the president of *IFCS* to open the third European Conference on Data Analysis, here in Colchester, at the University of Essex. First of all, on behalf of the Societies of the Federation, I would like to express my great pleasure in welcoming all participants to this Symposium.

This is a wonderful initiative of collaboration among the British, German and Polish Classification Societies, with the title *Data Science: Foundations, Methods and Applications*.

The IFCS in 1996 in Kobe, and also 1998 in Rome was the first scientific Association in the world to use the term "Data Science" in the title of the conference, with the speech of Chikio Hayhashi entitled "*What is Data Science? Fundamental Concepts and a Heuristic Example"*, thus, qualifying the pioneering research on Data Science proposed by several Classification Societies.

Data Science is not only a synthetic concept to connect Statistics and Computer Science, but represents a modern way to do research by taking advantage of the technological innovations and the automatic production of large volume of data, due to the computers and internet revolutions. Thus, it is very important that three well established Classification Societies explore the foundations, the innovative methods and the relevant applications in Data Science. In fact, in view of the rapidly advancing of science and technology and their research frontiers, with the quick spread of the knowledge in the information society, the Scientific Societies must have the capacity to intercept innovative and modern research. Meetings can no longer be the traditional media to give-and-get information with the outside world, because this is efficiently done through Internet without moving researchers. Modern conferences must detect and discuss deeply emerging and pioneering topics, becoming international laboratories to strengthen collaborations. To achieve this aim we need to reinforce our international network, parallelizing scientific activities and giving to each National Classification Society an international role in promoting collaboration among researchers by privileging international cooperation. I am firmly convinced that the ECDA Conference represents a wonderful "medium" to detect modern research topics and continuous collaboration among Classification Societies. The three Associations have organized sessions on emerging ideas where, in the recent years, members of

……………………………………………………………………………
Societies have achieved relevant theoretical, methodological and/or applied advances, in order to disseminate and discuss their outstanding scientific contributions.

This year we celebrate the 30th anniversary of the founding of IFCS. After 30 years the scope and purposes of the Federation -to promote mutual communication, cooperation and interchange of views among scientists interested in theories and practices of data analysis and classification in a wide range of applications as possible- is still very modern and we are glad that these purposes of mutual communication and cooperation have been so clearly adopted by the ECDA Conference.

I now pass to the acknowledgments. Very special thanks go to the program committee formed by highly appreciated experts who collaborated over the past year to create the best possible program format and content. In particular, I wish to congratulate with the Conference Chair Berthold Lausen and the Scientific Program Chairs Daniel Baier, Andreas Geyer-Schulz, Sabine Krolak-Schwerdt, Fionn Murtagh, Józef Pociecha for the hard work done and the competent support to finally set up the outstanding program. Special thanks go to the Local Organizing Committee for the excellent organization of the conference.

I would like to close my speech by expressing my sincere wishes for the success of the symposium and for all participants to discover new opportunities in the growing areas of research.

We invite you to enjoy the paper presentations during the next three days, and we hope that the discussions will be useful to increase international collaboration among researchers.

Maurizio Vichi

President of International Federation of Classification Societies

Welcome Note of the President of British Classification Society

Dear Participants,

Welcome to the European Conference on Data Analysis 2015, co-hosted by the *British Classification Society* (*BCS*), *Gesellschaft für Klassifikation e.V.* (*GfKl*) and *Sekcja Klasyfikacji i Analizy Danych* (*SKAD*) and held at the University of Essex from 1st to 4th September 2015.

ECDA 2015 was organised by Professor Berthold Lausen of the Department of Mathematical Sciences, University of Essex under the auspices of the British Classification Society, and attended by over 135 participants from 19 countries who have to choose from more than 100 invited and contributed lectures or workshops.

The sessions were organized under the headings *Data Analysis*, *Data Science*, *Clustering*, *Classification*, *Machine Learning and Knowledge Discovery*, *Library and Information Science*, *Finance and Economics*, *Digital Humanities and Social Sciences*, *Geosciences and Archaeology*, *Marketing*, *Musicology*, *Machine Learning and Knowledge Discovery*, *Outliers in Classification*, *Life Sciences*, *Mathematical Foundations of Data Science*, *Education*, *Big Data*, *Engineering, Logistics and Optimisation*.

The conference dinner will be held at Colchester Castle, the biggest Keep ever built in the United Kingdom and the largest that remains in existence throughout Europe.

I would like to thank all the speakers for their contributions, the many colleagues who reviewed the papers beforehand, those who chair sessions, and all who helped with the organization of this exceptional conference on data analysis.

David Wishart

President of British Classification Society

Welcome Note of the Section of Classification and Data Analysis of the Polish Statistical Society

Dear Participants,

On Behalf of *SKAD* (*Section of Classification and Data Analysis* of the *Polish Statistical Society – PTS*) it is my great pleasure to welcome you to the European Conference on Data Analysis 2015 in Colchester, United Kingdom.

I hope that this conference, as joint venture of GfKl, BCS and SKAD will be interesting, fruitful and inspiring for you.

I wish you productive discussions and a pleasant stay in the nice, old town Colchester.

Józef Pociecha

Welcome Note of the President of German Classification Society

Dear Participants,

Welcome to the European Conference on Data Analysis 2015 (ECDA 2015) *Data Science: Foundations, Methods and Applications* at the University of Essex, Colchester, UK. The British, German and Polish Classification Societies are co-hosting the conference. After the very successful ECDA conferences in Luxembourg (2013) and Bremen (2014) we are delighted to welcome the delegates from 19 different countries: Austria, Belgium, Canada, Cyprus, Germany, Finland, Greece, Italy, Luxembourg, Morocco, Mexico, New Zealand, Poland, Portugal, South Africa, Tunisia, United Kingdom, United States of America and Switzerland. The conference includes the *39th Annual Meeting of the German Classification Society* and the *Librarians and Information Science (LIS) Workshop*.

We are looking forward to the:

- keynotes of the conference
  - Hendrik Blockeel, Leuven: *Declarative data analysis*
  - Ulf Brefeld, Darmstadt: *Capturing user behaviour*
  - Andrzej Dudek, Wroclaw: *Cluster analysis in the XXI century, new algorithms and tendencies*
  - Arthur Gretton, London: *Kernel nonparametric tests of homogeneity, independence and multi-variable interaction*
  - Janette McQuillan, London: *Crowdsourcing classifications to accelerate cancer research*
  - Christine Müller, Dortmund: *Data depth*
  - Iris Pigeot, Bremen: *Challenges in the statistical analysis of longitudinal data*
  - Claus Weihs, Dortmund: *Efficient global optimization: motivation, variation, and application*
  - Stefan Wrobel, Bonn: *Data analytics in a networked world*
- five invited, 23 contributed sessions and the LIS workshop.

We hope that you enjoy the social program offer, visiting Colchester and beyond.

I like to thank my fellow *presidents* and *board members* of the *BCS*, *GfKl* and *SKAD* for organising the conference together, all members of the *scientific program committee* and *scientific program chairs* for their valuable work and important contribution. Moreover, I would like to thank the *University of Essex*, *Event Essex*, *Department of Mathematical Sciences* and the members of the *local organising committee* for the support and for making the conference possible; as well, our sponsors *Profusion* and *Institute for Analytics and Data Science* for supporting the conference.

I wish you all a fruitful, enjoyable and successful conference.

Berthold Lausen

Conference Chair and President of the Gesellschaft für Klassifikation

Fourth Joint Statistical Meeting of the Deutsche Arbeitsgemeinschaft Statistik "Statistics under one umbrella"

62. Biometrisches Kolloquium

Pfingsttagung der Deutschen Statistischen Gesellschaft

40. Jahrestagung der Gesellschaft für Klassifikation

DAG STAT
DEUTSCHE ARBEITSGEMEINSCHAFT STATISTIK

14. – 18. March 2016
Georg-August-Universität Göttingen
www.uni-goettingen.de/dagstat2016

DAGStat 2016

Vierte gemeinsame Tagung der Deutschen Arbeitsgemeinschaft Statistik "Statistik unter einem Dach"

vom 14.3. - 18.3. 2016 im zentralen Hörsaalgebäude (ZHG) der Georg-August Universität Göttingen

Fourth Joint Statistical Meeting of the Deutsche Arbeitsgemeinschaft Statistik "Statistics under one Umbrella"

14. - 18. March 2016 at the Zentrales Hörsaalgebäude (ZHG) of the Georg-August University Göttingen

62. Biometrisches Koloquium

Pfingsttagung der deutschen statistischen gesellschaft

40. Jahrestagung der Gesellschaft für Klassifikation

Registration and Abstract Submission for DAGStat2016 is now open.

Abstract submission for LIS workshop is now open.

The Best Paper Award honours the author(s) of a paper of exceptional merit dealing with a subject related to the Society's research activities.

The GfKl Award Jury consisting of Eyke Hüllermeier, Sabine Krolak-Schwerdt, Myra Spiliopoulou, Claus Weihs and Berthold Lausen have nominated Dr. Angelos Markos, Democritus University of Thrace, for the ECDA 2014 Best Paper Award.

The award consisting of a certificate and a €1000 cheque will be handed over during the Opening Ceremony. The *Laudatio* of the awardee will be given by Professor Hans Kestler, Chair of the Scientific Program Committee ECDA 2014, followed by a short presentation of the awarded paper.

Incremental Generalized Canonical Correlation Analysis

Angelos Markos[1] and Alfonso Iodice D'Enza[2]

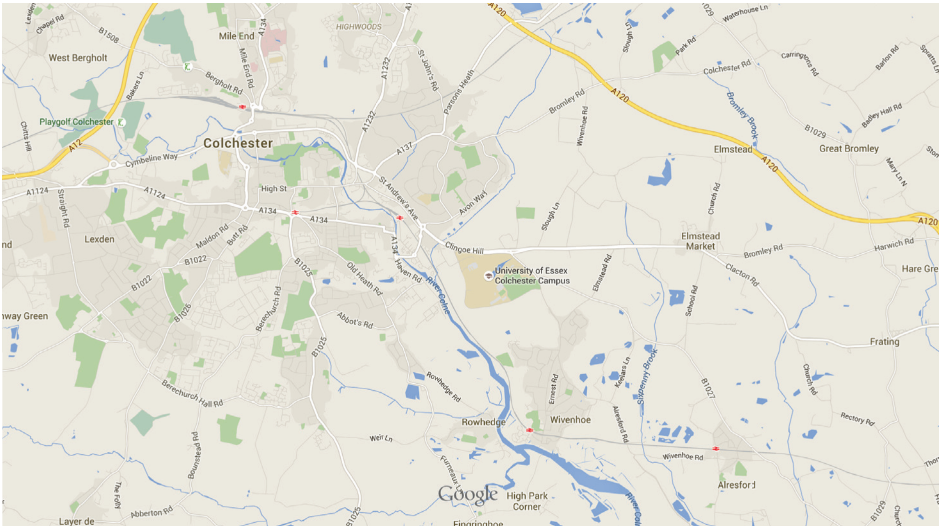
[1]Democritus University of Thrace, Department of Primary Education, 68100, Alexandroupoli, Greece *amarkos@eled.duth.gr*
[2]University of Cassino and Southern Lazio, Department of Economics and Law, Folcara 03043, Cassino FR, Italy *iodicede@unicas.it*

Abstract. Generalized canonical correlation analysis (GCANO) is a versatile technique that allows the joint analysis of several sets of data matrices through data reduction. The method embraces a number of representative techniques of multivariate data analysis as special cases. The GCANO solution can be obtained noniteratively through an eigenequation and distributional assumptions are not required. The high computational and memory requirements of ordinary eigendecomposition makes its application impractical on massive or sequential data sets. The aim of the present contribution is two-fold: *i)* to extend the family of GCANO techniques to a split-apply-combine framework, that leads to an exact implementation; *ii)* to allow for incremental updates of existing solutions, which lead to approximate yet highly accurate solutions. For this purpose, an incremental SVD approach with desirable properties is revised and embedded in the context of GCANO, and extends its applicability to modern big data problems and data streams.

The ***European Conference on Data Analysis 2015*** (ECDA 2015) is hosted by University of Essex one of the most internationally diverse universities in the United Kingdom. According to *Research Excellence Framework* (*REF 2014*) the University ranks in the top 20 UK universities for research excellence and the top 5 for social science research.

The University main campus is situated on the Eastern edge of Colchester, within Wivenhoe Park, less than a mile from the town of Wivenhoe and two miles from the town of Colchester (the oldest recorded Roman town in Britain).

Address: Wivenhoe Park,
Colchester CO4 3SQ,
United Kingdom



Kartendaten © 2015 Google 1 km

Fig. 1: University of Essex location

Finding your way to Colchester Campus:

- **by plane**
  - *London Stansted*: use the 133 coach service. Alternatively, you can take the coach service National Express.
  - *London Heathrow*: use the coach service to Colchester operated by National Express. Alternatively, take an underground train to Liverpool Street Station in London. You will need to take the Piccadilly line, eastbound, to Holborn, and then the Central line, eastbound, to Liverpool Street Station. Or, you can take a Heathrow Express train to Paddington Station in London and then get on the Underground (eastbound on the Circle line) to Liverpool Street Station. From Liverpool Street Station you will need to take a train to Colchester.
  - *London Gatwick*: use the coach service to Colchester operated by National Express. Alternatively, you can take a Gatwick Express train to Victoria Station in London, and then either a taxi or an underground train (eastbound on the Circle line) to Liverpool Street Station. From Liverpool Street Station you will need to take a train to Colchester.
  - *London Southend*: you can take a train to Shenfield. From Shenfield take a train to Colchester.
  - *London Luton*: use the coach service to Colchester operated by National Express. Alternatively, you can take the Thameslink train, which will take you to Farringdon. Join the Underground at Farringdon and take, either the Circle Line, the Metropolitan Line or the Hammersmith and City Line, eastbound, to Liverpool Street Station. From Liverpool Street Station you will need to take a train to Colchester.

- **by train**
  - trains run between London Liverpool Street and Colchester Station (also known as North Station) at approximately half-hourly intervals. The journey takes under an hour. Services also connect with Colchester from Norwich, Ipswich, Felixstowe, Harwich and Clacton/Walton.

Finding your way to Colchester Campus:

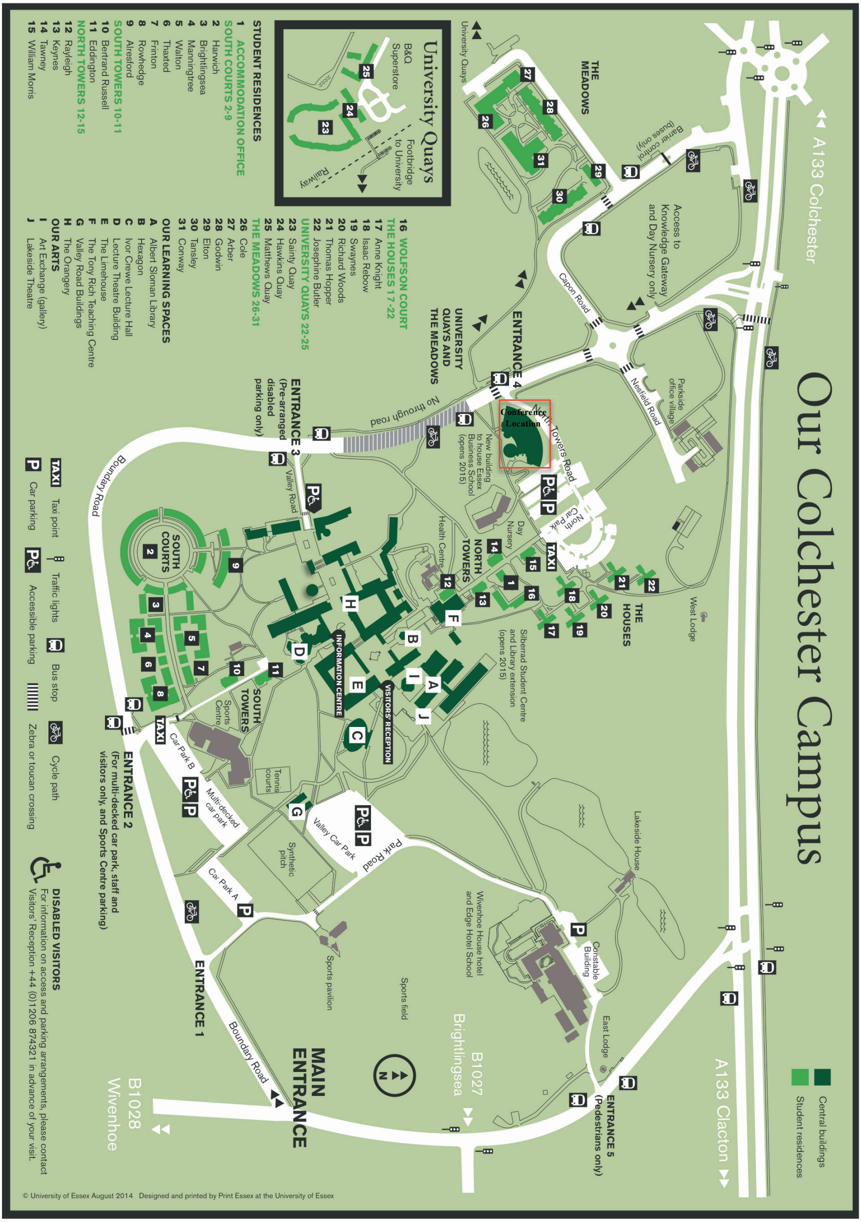
- **by bus**
  - to get to University of Essex, the bus services 61, 62, 74/74A, 75, 76 and 87 run through our campus connecting it to and from Colchester Station, Colchester Town Centre and Wivenhoe. They are operated by, First or Hedingham.

- **by taxi**
  - there are taxi ranks at Colchester North Station, Colchester Bus Station and in the Town Centre. The journey to the University normally takes about ten minutes and you should ask for North Towers.
    We strongly recommend you to book your taxi in advance.
    - *Towncar Minicabs*: +44 1206 515515
    - *Five Sevens*: +44 1206 577777
    - *Panther Cabs*: +44 1206 52552

- **by car**
  - Colchester can be reached by car either via:
    - **A12**, which links up with the M25 from the South (for Colchester take A133 exit)
    - **A14**, which links up with the M1/M6 from the North (for Colchester take A1232 exit)

    Please follow roadside signage to the University. You should park in the North Towers Car Park.
    For more information about parking spaces please visit: http://www.essex.ac.uk/staff/parking/

Conference Location
Campus Map



- The Opening Ceremony, plenary, semi-plenary lectures, invited and contributed sessions, as well the LIS' 2015 workshop, take place in the building housing the *Essex Business School* (*EBS*) located on University Campus. All rooms are situated on *Level 2* and fitted with the latest state-of-art audio-video (AV) facilities.

- A **rehearsal and preparation room** (*EBS 2.48*) is available for you during the conference. Additionally, *EBS 2.68* can be used for meetings.

- **Internet access** – ***eduroam*** is available on campus. If your institution is one of the participating organisations in *eduroam* you can directly login with your university email address and password.
  If you have any problems or queries with connecting to the Internet please contact our *IT helpdesk* located on the ground floor of the *Silberrad Student Centre*.

SOCIAL PROGRAM

European Conference on Data Analysis 2015

Tuesday, 01/September/2015	
8.00pm – 9.00pm	**Welcome Reception** (*Foyer, EBS*)
9.00pm – 10.30pm	**Informal Get Together** (Student Union Bar [*SU bar*])

Wednesday, 02/September/2015	
7.30pm – 9.00pm	**Flavour of Scotch Whisky** (*EBS 2.34*) Chair: Dr. David Wishart

Thursday, 03/September/2015	
6.00pm – 7.30pm	**Roman Walk** Departure by bus at 5.40 from North Towers
7.00pm – 8.00pm	**Access to Exhibition in Colchester Castle**
8.00pm – 10.30pm	**Conference Dinner** at **Colchester Castle** Departure by bus at 10.50pm to University accommodations

Please note:

The ***Conference fee*** for all participants includes:

- conference kit;
- welcome reception;
- all refreshments breaks;
- all lunches;
- Wednesday event “Flavour of Scotch Whisky”

Tuesday, 1st of September 2015

Welcome Reception
8pm – 9pm

- The welcome reception will take place on Tuesday 1st of September between **8pm and 9pm** at the **conference venue** (*Essex Business School*, *EBS*) in the **foyer area** (drinks & canapés are provided). Free to attend to all conference delegates, this will be a fantastic opportunity to meet and reconnect with colleagues from around the world.

Informal Get Together
9pm – 10.30pm

- For those who want to stay around a bit longer there is the opportunity to move in the *Students' Union Bar* (SU Bar) on the University of Essex Campus, located on *Square 3* (the costs are not included in the conference fee).

Wednesday, 2nd of September 2015 **7.30pm, Room:** ***EBS 2.34***

We'll tak a cup of kindness yet – the Flavour of Scotch Whisky

with **Dr. David Wishart**

Of all the spirits available today, single malt whiskies display the greatest diversity of flavour and heritage. David Wishart describes the production methods that influence the flavour, and concludes with a taste of several classic single malt whiskies that span his complete flavour spectrum.

He will guide you through the history and romance of Scotch whisky, from the *aqua vitae* of pre-Reformation monks, the *elixir* of medieval alchemists, to the hedonistic *uisge beatha* of remote Scottish crofts, and the taverns on Edinburgh's Royal Mile.

The heritage of Scotch whisky is evoked in the poems of Burns, the parties of George IV, Queen Victoria, Robert Stevenson's king of drinks, and the art of Landseer and Wilkie.

Today, due to variable peating and cask selection and preparation, the flavour of malt whisky is more widely diverse than ever. David describes his unique classification of Scotch malt whiskies by flavour, with several fine malt whiskies to taste. David's talk and tasting is sponsored by the whisky producers of Scotland.

Dr. David Wishart, author of *Whisky Classified: Choosing Single Malts by Flavour*, Third Edition 2012, is a Keeper of the Quaich and was Research Fellow at the School of Management, University of St Andrews.

Thursday, 3rd of September 2015

Coaches will be provided to take all those who pre-booked one of the two events to Colchester town centre and to return everyone who wants back to the University Campus at the end of the evening.

Roman walk
6pm

Fig.2: Castle Park

At **5.40pm** we will depart by bus to the Colchester town centre. From there, for those who booked the walk, some qualified guided tours will take you on an approximately 90-minute walk to discover the most beautiful corners and attractions of the city centre, including the picturesque Castle Park, the Dutch Quarter and the High Street, only minutes away from each other, but each with a very different story to tell.

Buses depart from University North Towers at 5.40pm.

Thursday, 3rd of September 2015

Exhibition opens in Colchester Castle
7pm

The exhibition would reveal many fascinating layers of history with archaeological collections including some of the most important Roman finds in Britain. Open only for those who attend the dinner.

Conference Dinner at Colchester Castle
8pm

Fig.3: Colchester Castle

The conference dinner will take place at Colchester Castle, the largest Norman Keep in Europe and one of England's most important heritage sites.

Please note that the conference dinner is only open to participants who booked the event during online registration. A voucher will be handed over to you upon registration on the first day of the conference. You will be asked to show this voucher before joining the event.

Buses depart from Colchester Castle at 10.50pm.

SCIENTIFIC PROGRAM

SCIENTIFIC PROGRAM
OVERVIEW

Tuesday, 01/September/2015	
2.30pm – 8.00pm	**Registration Conference** (*Foyer, EBS*)
6.00pm – 8.00pm	**Board Meeting – GfKl** (*EBS 2.68*)

Wednesday, 02/September/2015	
8.30am – 9.00am	**Registration Conference** (*Foyer, EBS*)
9.00am – 9.55am	**Opening Ceremony & Best Paper Award** (*EBS 2.2*)
10.00am – 10.45am	**Plenary 1: Claus Weihs** *"Efficient Global Optimization: Motivation, Variation, and Application"* Session Chair: Andreas Geyer-Schulz (*EBS 2.2*)
10.45am – 11.15am	**Refreshments**
11.15am – 12.55am	**Invited Sessions 1: Data Analysis in Finance** Chair: Krzysztof Jajuga (*EBS 2.2*)
	Contributed Sessions: Clustering I Chair: David Wishart (*EBS 2.34*)
	Contributed Sessions: Data Analysis I Chair: Karsten Lübke (*EBS 2.1*)
	Contributed Sessions: Engineering, Logistics and Optimisation Chair: Abdel Salhi (*EBS 2.65*)
1.00pm – 1.30pm	**AG-DANK Meeting** Chair: Hans-Joachim Mucha (*EBS 2.1*)
1.00pm – 2.30pm	**Lunch**
2.30pm – 3.15am	**Plenary 2: Christine Müller** *"Data depth"* Session Chair: Maurizio Vichi (*EBS 2.2*)
3.20pm – 4.35pm	**Invited Sessions 2: Outliers in Classification Procedures – Theory and Practice** Chair: *Józef Pociecha* (*EBS 2.2*)
	Contributed Sessions: Machine Learning and Knowledge Discovery I Chair: Cuevas Covarrubias Carlos (*EBS 2.34*)
	Contributed Sessions: Data Analysis II Chair: Willi Sauerbrei (*EBS 2.1*)
4.35pm – 5.00pm	**Refreshments**
5.00pm – 6.30pm	**GfKl annual general meeting** (*EBS 2.65*)
6.30pm – 7.30pm	**BCS annual general meeting** (*EBS 2.34*)

European Conference on Data Analysis 2015

Thursday, 03/September/2015	
9.00am – 9.45am	**Plenary 3: Iris Pigeot** *"Challenges in the statistical analysis of longitudinal data"* Session Chair: Berthold Lausen (*EBS 2.2*)
9.50am – 10.30am	**Semi-Plenary 1: Janette McQuillan** *"Crowdsourcing classifications to accelerate cancer research"* Session Chair: Hans A. Kestler (*EBS 2.2*)
9.50am – 10.30am	**Semi-Plenary 1: Hendrik Blockeel** *"Declarative Data Analysis"* Session Chair: Fionn Murtagh (*EBS 2.34*)
10.30am – 11.00am	**Refreshments**
11.00am – 12.15pm	**Invited Sessions 3: Data Science in Life Sciences** Chair: Hans A. Kestler (*EBS 2.2*)
	Contributed Sessions: Machine Learning and Knowledge Discovery II Chair: Andrzej Dudek (*EBS 2.34*)
	Contributed Sessions: Data Analysis III Chair: Christine Müller (*EBS 2.1*)
	Contributed Sessions: Education Chair: Andreas Geyer-Schulz (*EBS 2.65*)
12.15pm – 12.40pm	**AG-Biostatistics Meeting** Chair: Hans A. Kestler (*EBS 2.2*)
12.15pm – 1.45pm	**Lunch**
1.45pm – 2.30pm	**Plenary 4: Stefan Wrobel** *"Data Analytics in a Networked World"* Session Chair: Reinhold Decker (*EBS 2.2*)
2.35pm – 3.50pm	**Invited Sessions 4: Data Science** Chair: Adalbert Wilhelm (*EBS 2.2*)
	Contributed Sessions: Machine Learning and Knowledge Discovery III Chair: Alfred Ultsch (*EBS 2.34*)
	Contributed Sessions: Data Analysis IV Chair: Aris Perperoglou (*EBS 2.1*)
	Contributed Sessions: Marketing I Chair: Daniel Baier (*EBS 2.65*)

Thursday, 03/September/2015	
3.50pm – 4.15pm	**Refreshments**
4.15pm – 5.30pm	**Invited Sessions 5: Big Data** Chair: Berthold Lausen (*EBS 2.2*)
	Contributed Sessions: Machine Learning and Knowledge Discovery IV Chair: Ludwig Lausser (*EBS 2.34*)
	Contributed Sessions: Data Analysis V Chair: Hongsheng Dai (*EBS 2.1*)
	Contributed Sessions: Marketing II Chair: Reinhold Decker (*EBS 2.65*)

Friday, 04/September/2015	
09.00am – 10.15pm	**Contributed Sessions: Clustering II** Chair: Christian Hennig (*EBS 2.2*)
	Contributed Sessions: Digital Humanities and Social Sciences I Chair: Ali Ünlü (*EBS 2.34*)
	Contributed Sessions: Geosciences and Archeology Chair: Hans-Joachim Mucha (*EBS 2.1*)
	Contributed Sessions: Finance and Economics I Chair: Krzysztof Jajuga (*EBS 2.65*)
	Contributed Sessions: Musicology Chair: Claus Weihs (*EBS 2.66*)
10.15am – 10.45am	**Refreshments**
10.45am – 12.00pm	**Contributed Sessions: Classification** Chair: Axel Benner (*EBS 2.2*)
	Contributed Sessions: Digital Humanities and Social Sciences II Chair: Martin Behnisch (*EBS 2.34*)
	Contributed Sessions: Mathematical Foundations of Data Science Chair: Fionn Murtagh (*EBS 2.1*)
	Contributed Sessions: Finance and Economics II Chair: Adam Sagan (*EBS 2.65*)
12.05pm – 12.45pm	**Semi-Plenary 3: Arthur Gretton** *“Kernel nonparametric tests of homogeneity, independence and multi-variable interaction”* Session Chair: Claus Weihs (*EBS 2.2*)
12.05pm – 12.45pm	**Semi-Plenary 4: Ulf Brefeld** *“Capturing User Behaviour”* Session Chair: Alfred Ultsch (*EBS 2.34*)
12.50pm – 1.35pm	**Plenary 5: Andrzej Dudek** *“Cluster Analysis in the XXI century, new algorithms and tendencies”* Session Chair: Jozef Pociecha (*EBS 2.2*)
1.35pm – 3.00pm	**Farewell reception (including lunch)**

SCIENTIFIC PROGRAM
SESSIONS OVERVIEW

Wednesday, 2/September/2015 – AM Sessions

10.00am – 10.45am

Plenary 1
Chair: Andreas Geyer-Schulz (***EBS 2.2***)

11.15am – 12.55pm

Invited Sessions 1: Data Analysis in Finance
Chair: Krzysztof Jajuga (***EBS 2.2***)

Contributed Sessions: Clustering I
Chair: David Wishart (***EBS 2.34***)

Wednesday, 2/September/2015 – AM Sessions

11.15am – 12.55pm

Contributed Sessions: Data Analysis I
Chair: Karsten Lübke (*EBS 2.1*)

Contributed Sessions: Engineering, Logistics and Optimisation
Chair: Abdel Salhi (*EBS 2.65*)

……………………………………………………………………………………………...

Wednesday, 2/September/2015 – PM Sessions

Plenary 2
Chair: Maurizio Vichi (***EBS 2.2***)

2.30pm – 3.15pm

Data depth …………………………………………………………....... 4
Christine Müller

3.20pm – 4.35pm

Invited Sessions 2: Outliers in Classification Procedures – Theory and Practice
Chair: Józef Pociecha (***EBS 2.2***)

Problem of Outliers in Corporate Bankruptcy Prediction …………… 22
Barbara Pawełek, Józef Pociecha, Jadwiga Kostrzewska, Mateusz Baryła, Artur Lipieta

Outliers in Topic Modelling ……………………………………………... 23
Paweł Lula

Data Homogeneity in Classification of Objects on Local Housing Market ………………………………………………………………….. 24
Barbara Batóg, Iwona Foryś

Contributed Sessions: Machine Learning and Knowledge Discovery I
Chair: Cuevas-Covarrubias Carlos (***EBS 2.34***)

Partitioning high-dimensional multivariate self-affine time series into differentiated subseries …………………………………………………... 56
Christopher Michael Taylor

Ordinal signatures for prototype-based classifiers …………………….. 57
Andre Burkovski, Lyn-Rouven Schirra, Ludwig Lausser, Hans A. Kestler

Offline algorithm selection based on multicriteria optimization of contradicting performance measures ………………………………… 58
Daniel Horn, Aydin Demircioglu, Bernd Bischl, Tobias Glasmachers, Claus Weihs

Wednesday, 2/September/2015 – PM Sessions

3.20pm – 4.35pm

Contributed Sessions: Data Analysis II
Chair: Willi Sauerbrei (*EBS 2.1*)

Thursday, 3/September/2015 – AM Sessions

Plenary 3
Chair: Berthold Lausen (***EBS 2.2***)

9.00am – 9.45am

Challenges in the statistical analysis of longitudinal data 5
Iris Pigeot

Semi-Plenary 1
Chair: Hans A. Kestler (EBS 2.2)

9.50am – 10.30am

Crowdsourcing classifications to accelerate cancer research 6
Janette McQuillan

Semi-Plenary 2
Chair: Fionn Murtagh (EBS 2.34)

Declarative data analysis 7
Hendrik Blockeel

11.00am – 12.15pm

Invited Sessions 3: Data Science in Life Sciences
Chair: Hans A. Kestler (***EBS 2.2***)

Challenge of high- dimensional feature selection in survival analysis 26
Maral Saadati, Axel Benner

Understanding the Biological Functions of Gene Sets 28
Catharina Lippmann, Lotsch Joern, Alfred Ultsch

On variable selection and shrinkage strategies to derive multivariable regression models 30
Willi Sauerbrei, Hans van Houwelingen

Thursday, 3/September/2015 – AM Sessions

11.00am – 12.15pm

Contributed Sessions: Machine Learning and Knowledge Discovery II
Chair: Andrzej Dudek (***EBS 2.34***)

Contributed Sessions: Data Analysis III
Chair: Christine Müller (***EBS 2.1***)

Contributed Sessions: Education
Chair: Andreas Geyer- Schulz (***EBS 2.65***)

Thursday, 3/September/2015 – PM Sessions

Plenary 4
Chair: Reinhold Decker (***EBS 2.2***)

1.45pm – 2.30pm

Data Analytics in a Networked World ……………………….. 8
Stefan Wrobel

2.35pm – 3.50pm

Invited Sessions 4: Data Science
Chair: Adalbert Wilhelm (***EBS 2.2***)

Clustering through High Dimensional Data Scaling: Applications and Implementations ………………………………………………………….. 32
Fionn Murtagh, Pedro Contreras

An optimization approach to mining big data …………………………. 33
Abdel Salhi

Predicting hidden tourism: Exploratory data analysis and classification of regression coefficients …………………………………. 34
Claudio Conversano, Adalbert Wilhelm, Giula Contu, Francesco Mola

Contributed Sessions: Machine Learning and Knowledge Discovery III
Chair: Alfred Ultsch (***EBS 2.34***)

Feature selection algorithms for binarized data ……………………….. 76
Ludwig Lausser, Lyn-Rouven Schirra, Hans A. Kestler

A Comparison of Ensemble Methods for Motor Imagery Brain-Computer Interfaces …………………………………………………….... 77
Davide Valeriani, Ana Matran-Fernandez, Diego Perez-Liebana, Javier Asensio Cubero, Christian O'Connell, Andrei B. Iacob

Dynamic Process Modeling: Combining Control Charts to detect Concept drift in Nonstationary Environment ………………………….. 78
Dhouha Mejri, Mohamed Limam, Claus Weihs

Thursday, 3/September/2015 – PM Sessions

2.35pm – 3.50pm

Contributed Sessions: Data Analysis IV
Chair: Aris Perperoglou (*EBS 2.1*)

Contributed Sessions: Marketing I
Chair: Daniel Baier (*EBS 2.65*)

……………………………………………………………………………………………………...

Thursday, 3/September/2015 – PM Sessions

4.15pm – 5.30pm

Invited Sessions 5: Big Data
Chair: Berthold Lausen (***EBS 2.2***)

Contributed Sessions: Machine Learning and Knowledge Discovery IV
Chair: Ludwig Lausser (***EBS 2.34***)

Thursday, 3/September/2015 – PM Sessions

4.15pm – 5.30pm

Contributed Sessions: Data Analysis V
Chair: Hongsheng Dai (*EBS 2.1*)

Contributed Sessions: Marketing II
Chair: Reinhold Decker (*EBS 2.65*)

Friday, 4/September/2015 – AM Sessions

9.00am – 10.15am

Contributed Sessions: Clustering II
Chair: Christian Hennig (***EBS 2.2***)

Generalization, Combination and Extension of Functional Clustering Algorithms ………… 100
Christina Yassouridis, Friedrich Leisch

One dimensional Markov random field model for the analysis of ChIP- seq data with a non-parametric component ………… 101
Baba Bukar Alhaji, Hongsheng Dai, Andrew Harrison, Berthold Lausen

A Hierarchical Bayesian Model for joint Clustering of Clinical and Heterogenous Omics Data ………… 102
Ashar Ahmad, Holger Fröhlich

Contributed Sessions: Digital Humanities and Social Sciences I
Chair: Ali Ünlü (***EBS 2.34***)

Student Values Revisited - Measuring Dominant Interests in Personality ………… 104
Thomas Hummel, Victoria-Anne Schweigert, Andreas Geyer-Schulz

Movers and Stayers in a Religious Marketplace ………… 105
Barry William McDonald

A social sustainability model: An application to Mexican Small-Scale Dairy-Farming Households ………… 106
Monica Elizama Ruiz-Torres, Ana Lorga da Silva, Carlos Manuel Arriaga-Jordan, Francisco Ernesto Martinez-Castañeda

Contributed Sessions: **Geosciences and Archeology**
Chair: Hans-Joachim Mucha (***EBS 2.1***)

Data Mining in Atmospheric Gravity Waves ………… 108
Alfred Ultsch, Christopher Rogos, Christof Maul

Does Landscape Attractiveness affect Land Consumption in Germany? ………… 109
Martin Behnisch, Alfred Ultsch

Analysing past settlement size and location ………… 111
Irmela Herzog

Friday, 4/September/2015 – AM Sessions

9.00am – 10.15am

Contributed Sessions: Finance and Economics I

Chair: Krzysztof Jajuga (***EBS 2.65***)

Contributed Sessions: Musicology

Chair: Claus Weihs (***EBS 2.66***)

………………………………………………………………………………………………………...

Friday, 4/September/2015 – AM Sessions

10.45am – 12.00pm

Contributed Sessions: Classification
Chair: Axel Benner (***EBS 2.2***)

Contributed Sessions: Digital Humanities and Social Sciences II
Chair: Martin Behnisch (***EBS 2.34***)

Friday, 4/September/2015 – AM Sessions

10.45am – 12.00pm

Contributed Sessions: Mathematical Foundations of Data Science
Chair: Fionn Murtagh (***EBS 2.1***)

Contributed Sessions: Finance and Economics II
Chair: Adam Sagan (***EBS 2.65***)

Friday, 4/September/2015 – PM Sessions

	Semi-Plenary 3 Chair: Claus Weihs (***EBS 2.2***)	
12.05pm – 12.45pm	**Kernel nonparametric tests of homogeneity, independence and multi- variable interaction** ……………………………….. Arthur Gretton	9
	Semi-Plenary 4 Chair: Alfred Ultsch (***EBS 2.34***)	
	Capturing User Behaviour …………………………………….. Ulf Brefeld	10

12.50pm – 1.35pm	***Plenary 5*** Chair: Jozef Pociecha (***EBS 2.2***)	
	Cluster analysis in the XXI century, new algorithms and tendencies ……………………………………………………. Andrzej Dudek	11

SCIENTIFIC PROGRAM

LIBRARY and INFORMATION SCIENCE (LIS)

Library and Information Science (LIS)

September 2, 2015		
09:00am	**Opening of ECDA** (incl. Best Paper Award)	
10:00am	**Plenary 1**	
10:45am	*Coffee Break*	
	Chair: Frank Scholze (KIT Library) ***Room EBS 2.66***	
11:15am	**Building the Bridge: Mapping Different Knowledge Organization Systems in Economics** ……………………….. Author(s): Kempf, Andreas Oskar; Neubert, Joachim (ZBW - Leibniz Information Centre for Economics)	142
12:00pm	**Towards a Comprehensive Knowledge Organisation System for the Engineering Domain** ……………………….. Author(s): Bernauer, Elena (WTI Frankfurt,); Mehlberg, Martin (TIB Hannover); Runnwerth, Mila (TIB Hannover); Schmidt, Gudrun (WTI Frankfurt)	143
1:00pm	***Lunch***	
2:00pm	**Plenary 2**	
3:20pm	**Subject Cataloguing in an RDA Framework – Strategies and Practical Experience from Germany** ………………….. Author(s): Wiesenmüller, Heidrun (Stuttgart Media University)	144
4:00pm	**The Role of Classification Information in Open Access Repositories - current status and future directions** ………. Author(s): Summann, Friedrich; Dirk, Pieper (Bielefeld University Library)	145
4:40pm	*Coffee Break*	
5:00pm	**Annual Meeting GfKl**	
6:00pm	**Library Tour** (tbd)	
7:30pm	**Dinner** (The Old Siege House Bar & Brasserie at 75 East Street, Colchester, CO1 2TS) http://www.theoldsiegehousebarandbrasserie.co.uk/index.php/find-contact	

Library and Information Science (LIS)

September 3, 2015		
09:00am	**Plenary 3**	
09:50am	**Semi-Plenary**	
10:30am	*Coffee Break*	
	Chair: Magnus Pfeffer (Stuttgart Media University) ***Room EBS 2.66***	
11:00am	**Patent Claim Structure Recognition** ……………………….. Author(s): Hackl-Sommer, Rene; Schwantner, Michael (FIZ Kalrsruhe)	147
11:40am	**SMGlom - a Semantic Mathematical Glossary of the Next Generation** …………………………………………………. Author(s): Sperber, Wolfram (FIZ Karlsruhe, Germany); Kohlhase, Michael (Jacobs University Bremen)	148
12:20pm	*Lunch*	
1:45pm	**Plenary 4**	
	Chair: Heidrun Wiesenmüller (Stuttgart Media University) ***Room EBS 2.66***	
2:35pm	**The mapping tool "Cocoda"** ………………………………... Author(s): Balakrishnan, Uma (Common Library Network GBV)	150
3:15pm	**Automatic Identification of Synonym Relations in the Dutch Parliament Thesaurus** ……………………………….. Author(s): Aga, Rosa Tsegaye; Wartena, Christian (Hochschule Hannover, Germany); Lange, Otto (Next2Know); Aders, Nelleke (Dienst Informatievoorziening - Tweede Kamer der Staaten Generaal)	151

3:55pm	**Bibliographic Report 2014: A choice of relevant classification literature // Online Report 2014: A choice of nice web-features for subject cataloguing** …………………
	Author(s): Franke-Maier, Michael (Freie Universität Berlin); Peichl, Gerald (University St. Gallen)
		152

4:15pm	**Farewell**	

We kindly ask you to follow the instructions below to ensure your presentation goes smoothly.

- for *plenary talks*: you will have 40 minutes for your presentation and 5 minutes for discussion.
- for *semi-plenary talks*: you will have 35 minutes for your presentation and 5 minutes for discussion.
- for *invited* and *contributed sessions*: you will have 20 minutes for your presentation and 5 minutes for discussion.
- all presentations (in PowerPoint or as PDF.file) should be brought on a USB-Stick.
- if you wish to incorporate videos in your presentation, make sure you check that it works at the computers at the venue well beforehand (best the day before your presentation).
- all presentation computers have Internet access.

Chairs of the session will advise you when you have 3 minutes and 1 minute remaining.

At the day of your presentation:

- you should arrive at least 5 minutes before the session start to the assigned rooms to upload your presentation on the computer. AV assistance is available in each room.

ABSTRACTS

Part I

Plenary and Semi-Plenary Talks

Efficient Global Optimization: Motivation, Variation, and Application

Claus Weihs[1]

TU Dortmund University, Germany claus.weihs@tu-dortmund.de

Abstract. Modern statistical problem solvers confront needed optimization procedures with two major challenges that most classical and modern optimization techniques cannot master: many local optima and cost-intensive experiments. Classical deterministic optimization procedures often stay in local optima for non-convex problems, but global optimization is central for the adequate solution of most problems. Modern stochastic optimization may be able to manage many local optima, though at the price of using many function evaluations. Therefore, these techniques fail to cope with the second challenge: cost-intensive experiments. To overcome both challenges together, a class of efficient and global optimization (EGO) procedures is developed, first introduced by Jones et al. in 1998. EGO procedures are often built of the following steps: initial design, model for the relationship between target and influential factors, infill criterion combining assessment of target function value and uncertainty of prediction, optimization of the infill criterion to identify the next design point(s), experiment(s) in the new design point(s), and stopping criterion for finishing iterative optimization. All these steps can be varied. For example, kriging models are very popular, but classical regression models might also be appropriate for modeling the relationship between target and factors. A popular infill criterion is expected improvement, which is sometimes replaced by the so-called lcb-criterion. Applications are diverse. For example, the free parameters of an SVM might be tuned by EGO and compared to the results of a standard grid search. Other discussed applications will be the optimization of the cutting process with a diamond tipped drill core bit and model based optimization of music onset detection.

References

JONES, D.R., SCHONLAU, M., and WELCH, W.J. (1998): Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization 13, 455?492*.

Data depth

Christine H. Müller[1]

TU Dortmund University, Department of Statistics, Vogelpothsweg 87, D-44221 Dortmund, Germany, cmueller@statistik.tu-dortmund.de

Abstract. Starting with half-space depth and simplicial depth, an introduction of data depth for the multivariate location problem is given. These depth notions can be used for DD-plots and with this for classification. Another application of data depth is regression using the notion of a nonfit. This can be generalized to any model with residuals leading to residual depth and simplicial residual depth. It is shown that simplicial residual depth reduces to counting the number of alternating signs of residuals of specific subsets under some conditions. Although this is a nice property, the calculation of simplicial residual depth is time consuming and its asymptotic distribution is only known for models with one and two unknown paramters up to now. Therefore simplified simplicial depth notions based on alternating signs of residuals are proposed. An application for detecting a change point in a crack growth process is given.

References

LI, J., CUESTA-ALBERTOS, J.A. and LIU, R.Y. (2012): DD-classifier: nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association, 107:498, 737–753, 2012*.

KUSTOSZ, CH.P., MÜLLER, CH.H. and WENDLER, M. (2015). Simplified simplicial depth for regression and autoregressive growth processes. *Submitted.*

KUSTOSZ, CH.P. and MÜLLER, CH.H. (2014). Analysis of crack growth with robust, distributionfree estimators and tests for nonstationary autoregressive processes. *Statistical Papers 55, 125–140, 2014*.

MÜLLER, CH.H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis 95, 153-181, 2005*.

Keywords

HALF-SPACE DEPTH, SIMPLICIAL DEPTH, DD-PLOT, REGRESSION DEPTH, SIMPLIFIED SIMPLICIAL DEPTH

Challenges in the statistical analysis of longitudinal data

Iris Pigeot[1] and Claudia Börnhorst[1]

Leibniz Institute for Prevention Research and Epidemiology – BIPS

Abstract. During the last decades, numerous cohort studies have been established or newly initiated and provide large data resources that are often even further complemented by linkage to routine data. Moreover, the rapid increase of computing capacity enables the estimation of statistical models of almost any complexity. However, the selection of appropriate statistical models to answer research questions in the context of cohort studies, especially with respect to life-course questions, is still a challenge for epidemiologists. Life-course epidemiology mainly investigates risk accumulations, critical/sensitive periods as well as pathways of risks that may explain the development of diseases over the life-course. Based on examples from international, multi-centre cohort studies such as the IDEFICS study or the German National Cohort, problems and challenges in the statistical analysis of longitudinal data will be illustrated. An overview of statistical models typically applied in life-course research will be given where associations between early growth and the development of later diseases will serve as an example. Finally, these methods will be applied to the IDEFICS study which aimed to investigate the causes of diet- and lifestyle-induced health effects in 16.228 children aged 2 to 9 years from eight European countries. Main focus was on the aetiology of childhood obesity, which is known to track into adulthood. In addition, obesity itself has been noted as a risk factor for various diseases. But it is still unknown whether it is mainly the current weight status or (in addition) the trajectory of growth that affects current health. To answer this research question, a 2-step procedure was applied to assess the association between body mass index trajectories during infancy and childhood and later metabolic risk and hence to identify sensitive periods of growth during which the later metabolic risk may be affected.

Crowdsourcing Classifications to Accelerate Cancer Research

Janette McQuillan[1]

Cancer Research UK, Angel Building, 407 St John Street, London, EC1V 4AD
Janette.McQuillan@cancer.org.uk

Abstract. The past two decades have seen an exponential increase in the amount of medical data generated. This has sparked novel research questions across cancer disciplines as varied as epidemiology, genetics, and clinical practice. However, our ability to analyse these data to answer outstanding questions has not progressed at the same rate, leading to vast amounts of unexplored data within the field of cancer research. Without a way of analysing the backlog of data we are unable to unlock crucial information and increase our understanding of cancer. Visual recognition software is being developed to address this gap. However, these software solutions are expensive, require significant manual intervention and validation from researchers and are not accurate enough at identifying cancer cells. An analytical solution that could be used to address this problem is the use of a crowdsourcing approach. The Citizen Science programme at Cancer Research UK have utilised such an approach by developing a number of different products that allow the general public to participate in cancer research. Candido dos Reis et al (2015) demonstrate that Citizen Scientists are reasonably good at identifying the presence of cancer. This study will discuss some of the techniques used to further improve upon the levels of accuracy displayed such as optimisation of experimental design, classifier aggregation and user-weighting algorithms.

References

CANDIDO DOS REIS, F. J., LYNN, S., ALI, H. R., ECCLES, D., HANBY, A. and PROVENZANO, E., CALDAS, C., HOWAT, W. J.,MCDUFFUS, L., LIU, B. and others (2015): Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMedicine, 1-22.* (doi:10.1016/j.ebiom.2015.05.009)

Keywords

CROWDSOURCING, CLASSIFIER AGGREGATION, CANCER RESEARCH

Declarative data analysis

Hendrik Blockeel[1]

Katholieke Universiteit Leuven Hendrik.Blockeel@cs.kuleuven.be

Abstract. With increasing amounts of ever more complex forms of digital data becoming available, the methods for analyzing these data have also become more diverse and sophisticated. With this comes an increased risk of incorrect use of these methods, and a greater burden on the user to be knowledgeable about their assumptions. In addition, the user needs to know about a wide variety of methods to be able to apply the most suitable one to a particular problem. This combination of broad and deep knowledge is not sustainable.

The idea behind declarative data analysis is that the burden of choosing the right statistical methodology for answering a research question should no longer lie with the user, but with the system. The user should be able to simply describe the problem, formulate a question, and let the system take it from there. To achieve this, we need to find answers to questions such as: what languages are suitable for formulating these questions, and what execution mechanisms can we develop for them? In this talk, I will discuss recent and ongoing research in this direction. The talk will touch upon query languages for data mining and for statistical inference, declarative modeling for data mining, meta-learning, and constraint-based data mining. What connects these research threads is that they all strive to put intelligence about data analysis into the system, instead of assuming it resides in the user.

Data Analytics in a Networked World

Stefan Wrobel[1]

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Bonn, Germany

Abstract. The age of big data has given an enormous push to statistics, machine learning, and the neighboring analytics disciplines. Yet, big data is still widely perceived as referring primarily to the sheer volume of data. In reality, the true challenges of big data result from the characteristics of the data that we are dealing with. In this talk, we will focus in particular on the fact that today, data points are almost never measurements in isolation, but are part of a networked world, where objects and their relationships are to be taken into account. Consequently, analyzing data that have network or graph structure is becoming of enormous importance, as is the challenge of visualizing such data. In this talk, we will introduce the challenges brought about by network data, and illustrate why in a data-driven economy, linked data will become even more important in the future. We will then present a few examples of methods for analyzing graph-structured data, ranging from pattern enumeration via classification to the visual analysis of graph-structured trajectory data.

Kernel nonparametric tests of homogeneity, independence and multi-variable interaction

Arthur Gretton

University College London arthur.gretton@gmail.com

Abstract. We consider three nonparametric hypothesis testing problems: (1) Given samples from distributions p and q, a homogeneity test determines whether to accept or reject p=q; (2) Given a joint distribution p_{xy} over random variables x and y, an independence test investigates whether $p_{xy} = p_x p_y$, (3) Given a joint distribution over several variables, we may test for whether there exist a factorization (e.g., $P_{xyz} = P_{xy}P_z$, or for the case of total independence, $P_{xyz} = P_x P_y P_z$). The final test (3) is of particular interest in fitting directed graphical models, as it may be used in detecting cases where two independent causes individually have weak influence on a third dependent variable, but their combined effect has a strong influence, even when these variables have high dimension. We present nonparametric tests for the three cases described, based on distances between embeddings of probability measures to reproducing kernel Hilbert spaces (RKHS), which constitute the test statistics (eg for independence, the distance is between the embedding of the joint, and that of the product of the marginals). The tests benefit from decades of machine research on kernels for various domains, and thus apply to distributions on high dimensional vectors, images, strings, graphs, groups, and semigroups, among others. The energy distance and distance covariance statistics are particular instances of these RKHS statistics.

Capturing User Behaviour

Lutz Brefeld[1]

Technical University of Darmstadt brefeld@cs.tu-darmstadt.de

Abstract. The Web has become a major resource to serve user information and tangible needs. Websites usually rely on repeated user visits, so their success depends highly on how well they are able to anticipate a user's needs by providing the right content, at the right time, in the right places. Guessing the intention of users is thus not only fundamental for the overall user experience but is often directly linked to revenue. User navigation patterns, on the other hand, are driven by a mixture of long- and short term interests, spontaneous inspiration and external factors (e.g,. weather, location). Capturing user intent is therefore one of the most challenging problems in many information retrieval and recommendation tasks. I will introduce sequential models to capture and predict user behaviour in scenarios where only implicit feedback is given and no private data is available. I will report on empirical results for different domains and approaches.

Cluster analysis in the XXI century, new algorithms and tendencies

Andrzej Dudek[1]

Wrocław University of Economics,
Department of Econometrics and Computer Science,
Nowowowiejska 3, 58-500 Jelenia Góra, Poland
andrzej.dudek@ue.wroc.pl

Abstract. Cluster analysis is well – developed technique of data analysis with known families of algorithms like hierarchical clustering, model - based clustering, optimization methods and others dealing especially well with clusters given from normal distribution. Approximately from the end of previous and beginning of XXI century new techniques like spectral approach, ensemble approach and mean shift family arisen, which give also good results for untypical cluster shapes (such as well – known ”spirals” dataset).

The issues that should be addressed by modern methods of cluster analysis are additionally: the heterogeneity of the data, the size of data (rapidly growing) and computing capabilities of computers.

In the presentation an attempt will be made to present the challenges faced by cluster analysis as well as the results of simulation studies comparing classical and non-classical cluster analysis methods on some standard and ”fancy” datasets.

References

CHENG Y. (1995): Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 8, p. 790–799.*

JAIN A.K. (2010): Data clustering: 50 years beyond K-means. *Pattern Recognition Letters 31 p.651-–666*

NG, A., JORDAN, M., WIESS, Y. (2002): On spectral clustering: analysis and algorithm. [In:] T. Diettrich, S. Becker, Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*, MIT Press, p. 849–856.

STREHL A., GHOSH J. (2003): Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research, 3:583–617, ISSN 1533-7928*

Keywords

CLUSTERING, ENSEMBLE APPROACH, SPECTRAL CLUSTERING

Part II

Invited Sessions

Invited Session 1: Data Analysis in Finance

Chaired by Krzysztof Jajuga

Wednesday, September 2, 2015: 11:15am - 12:55pm Room: EBS 2.2

Estimation error and confidence intervals for AR-GARCH-based estimators of VaR and ES

Krzysztof Piontek

Department of Financial Investments and Risk Management
Wroclaw University of Economics, ul. Komandorska 118/120, Wroclaw, Poland
krzysztof.piontek@ue.wroc.pl

Abstract. Model risk is inevitable in many financial issues. The potential threat stems from ignorance of this risk existence, its magnitude and consequence. It arises mainly from incorrect assumptions and parameter uncertainty (e.g. short samples).

Due to the model risk, any quantile-based risk measure value is a random variable, and may be estimated using point or interval procedure. Interval estimation approach is extremely rare, however it should complement a point estimation.

Research presents issues related to the model risk in the process of Value-at-Risk (VaR) and Expected Shortfall (ES) estimation as a result of the incorrect parameter estimation one-dimensional AR-GARCH models. The aim of this study is to illustrate the width of VaR and ES confidence intervals for typical cases.

For chosen financial time series models, author calculated confidence intervals for VaR and ES based on the series with different numbers of observations. Intervals were determined on the basis of a two-stage bootstrap procedure taking into account both the errors in estimating the parameters of the AR-GARCH models, as well as the distribution of the standardized quantile residuals, without assumptions about the form of the conditional distribution. Simulated and market data were used.

The results of this study are also important for backtesting of the quantile-based risk measures. It is an extension of the previous studies of the author.

References

CHRISTOFFERSEN, P., GONÇALVES, S. (2005): Estimation risk in financial risk management, *Journal of Risk*, Vol. 7, No. 3, pp. 1–28

HANSEN, B. (2006): Interval forecasts and parameter uncertainty, *Journal of Econometrics* 135(1-2), pp. 377–398

LÖNNBARK, C. (2013): On the role of the estimation error in prediction of expected shortfall, *Journal of Banking & Finance*, Volume 37, 3, pp. 847–853

Keywords

MODEL RISK, ARMA-GARCH, INTERVAL ESTIMATION, VaR, ES

Variable Selection Methods for Forecasting Multiple Business Exits in Europe

Alessandra Amendola[1], Francesca Ametrano[2], Marialuisa Restaino[3], and Luca Sensini[4]

[1] Department of Economics and Statistics, University of Salerno, Italy
alamendola@unisa.it
[2] Department of Economics, University of Genoa, Italy
francesca.ametrano@gmail.com
[3] Department of Economics and Statistics, University of Salerno, Italy
mlrestaino@unisa.it
[4] Department of Management and Information Technology, University of Salerno, Italy
lsensini@unisa.it

Abstract. The difficulties experienced by firms and institutions during the Global Financial Crisis (GFC) demonstrated the importance of understanding the determinants of financial default risk and investigating the differences between causes of failure and between industries, regions and countries. The aim of the paper is to identify the main variables that drive the financial distress across the European countries paying attention to the different reasons that may cause the exit from the market. An approach that takes into account different causes of failure has been implemented at both national and European levels, allowing us to study the single-country specificities as well as the between-country interdependencies. The most significant variables have been selected by means of some variable selection methods (stepwise, lasso, adaptive lasso, and so on), and all methods have been compared in terms of predictive ability by means of some accuracy measures, widely used in the business failure literature, in order to assess which procedure outperforms. Then, the sign of the selected variables is compared for each country model, in order to evaluate the differences in the determinants of financial distress and in the predictive ability of the model set-ups and to give an economic evaluation and interpretation of the models.

References

AMENDOLA A., RESTAINO M. and SENSINI L. (2010): Variable selection in default risk models, *Journal of Risk Model Validation, 5(1), 1-20, 2010.*

CHANCHARAT N., DAVY P., MCCRAE M. and LODH S. (2010): Multiple States of Financially Distressed companies: Tests using a Competing-Risks Model, *Australasian Accounting Business and Finance Journal, 4(4), 27-49, 2010.*

DICKERSON A.P., GIBSON H.D. and TSAKALOTOS E. (2003): Is attack the best form of defence? A competing risks analysis of acquisition actitivyt in the UK, *Cambridge Journal of Economics, 27, 337-357, 2003.*

DU JARDIN, P. (2010): Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy, *Neurocomputing, 73, 2047–2060, 2010*.

HÄRDLE W., LEE Y., SCHAFER D. and YEH Y. (2009): Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for Predicting the Default Risk of Companies, *Journal of Forecasting, 28(6), 512–534, 2009*.

Keywords

VARIABLE SELECTION, BUSINESS FAILURE, PREDICTIVE ACCURACY, DEFAULT METHODS.

Vulnerability of CoVaR to wrong estimates of VaR

Katarzyna Kuziak

Department of Financial Investments and Risk Management
Wroclaw University of Economics, ul. Komandorska 118/120, Wroclaw, Poland
katarzyna.kuziak@ue.wroc.pl

Abstract. Amongst systemic risk measures that have been proposed like: CoVaR proposed by Adrian and Brunnermeier (2010), MES suggested by Acharya et al. (2010); SRISK proposed by Brownlees and Engle (2011) and the Shapley value (SV) approach of Tarashev et al. (2010) in this paper one have gained particular attention - CoVaR. CoVaR is defined as the VaR of the financial system given that the institution is under financial distress. Systemic Risk is measured by the Value at Risk (VaR) of the financial system (or a subset of it). This method is fundamentally based on using daily price data to forecast VaR as a first step in the calculation of CoVaR. The main goal of this paper is to analyse CoVaR from the point of view of model risk. Two sources of model risk will be analysed: model misspecification (e.g. misspecifying the underlying stochastic process) and incorrect model calibration (e.g. estimation errors; outliers; estimation intervals; calibration and revision of estimated parameters). The analysis will be based on simulation studies.

References

ACHARYA, V. V., PEDERSEN L.H., PHILIPPON T. and RICHARDSON M. (2010): Measuring Systemic Risk, NYU working paper.

ADRIAN, T., BRUNNERMEIER, M.K. (2011): CoVaR,. Technical Report, Federal Reserve Bank of New York, Staff Reports no. 348.

BROWNLEES, C. T. and ENGLE R. (2011): Volatility, Correlation and Tails for Systemic Risk Measurement, NYU working paper.

HANSEN, L.P. (2013): Challenges in Identifying and Measuring Systemic Risk available at http://www.nber.org/chapters/c12507.pdf.

TARASHEV, N., BORIO C. and TSATSARONIS K. (2010): Attributing systemic risk to individual institutions, Technical Report Working Papers No 308, BIS.

Keywords

MODEL RISK, SYSTEMIC RISK, VaR, CoVaR,

Using Mixture Models for Prediction from Time Series, with Application to Energy Use Data

Najla M. Qarmalah[1], Frank Coolen[1], and Jochen Einbeck[1]

Durham University, Durham, UK najla.qarmalah@durham.ac.uk

Abstract. Non–parametric smoothing is a technique for analysing the trends of data. Gijbels et al. (1999) show that exponential smoothing can be put into a non–parametric regression framework to minimise the average squared residual of previous one–step–ahead forecasts.

This research aims to use mixture models to improve predictions from time series data. Given data of the form $(t_i, y_i), i = 1, \dots, T$, we suppose a model for the k-th component as $y_i = m_k(t_i) + \varepsilon_{ik}$ with proportion $\pi_k(t_i)$ such that $0 < \pi_k(t_i) < 1$ and $\sum_{k=1}^{K} \pi_k(t_i) = 1$, K is the number of components, $m_k(t_i)$ are smooth unspecified regression functions, and the errors $\varepsilon_{ik} \sim N(0, \sigma^2)$ are independently distributed. Estimation of this model is achieved through a kernel–weighted version of the EM–algorithm, using exponential kernels with different bandwidths (neighbourhood sizes) h_k as weight functions. By modelling a mixture of local regressions centred at t_T but with different bandwidths h_k, the estimated mixture probabilities are informative for the amount of information available in the data set at the scale of resolution corresponding to each bandwidth. Nadaraya–Watson and local linear estimators are used to carry out the localized estimation step.

In addition, several approaches for prediction at time t_{T+1} from this model are investigated. The data under study give the energy use for 135 countries from 1971 to 2007. So far, the model used here enables prediction from a time series for a given country. Currently, we are looking at prediction for multiple countries in time series by borrowing strength from countries within the same cluster and also for multiple future time points.

References

GIJBELS, I., POPE, A. and WAND M. P. (1999): Understanding exponential smoothing via kernel regression. *Journal of Royal Statistical Society, 61, 39–50, 1999.*

International Energy Agency, Available at:http://www.iea.org/

Keywords

NON-PARAMETRIC SMOOTHING, EXPONENTIAL SMOOTHING, MIXTURE MODELS, TIME SERIES

Invited Session 2: Outliers in Classification Procedures - Theory and Practice

Chaired by Jozef Pociecha

Wednesday, September 2, 2015: 3:20pm - 4:55pm Room: EBS 2.2

Problem of Outliers in Corporate Bankruptcy Prediction

Barbara Pawełek, Józef Pociecha, Jadwiga Kostrzewska, Mateusz Baryła and Artur Lipieta

Cracow University of Economics, Department of Statistics, 27 Rakowicka Street, 31-510 Cracow, Poland {barbara.pawelek, jozef.pociecha, jadwiga.kostrzewska, mateusz.baryla, artur.lipieta}@uek.krakow.pl

Abstract. The results of financial condition analysis are used, among other things, in the research on bankruptcy prediction of companies. The assessment of financial data quality involves also the detection of outliers. In the literature on bankruptcy prediction one can find deliberations on the problem of outliers. The proposals for solving this problem range from not taking any actions, through replacing or removing the outliers, to applying robust methods. Therefore, in the empirical research, some doubts concerning the choice of an appropriate approach to the outliers appear.

The aim of the article is to present the outcomes of empirical research on the usefulness of selected techniques for identifying outliers in bankruptcy forecasting. In the study, both one-dimensional (e.g. based on centiles, Tukey's criterion) and multidimensional (e.g. depth functions) procedures of outliers detection will be considered. So as to assess the classification accuracy of chosen bankruptcy prediction methods for a test set, among other things, type I error, type II error, ROC curve and AUC measure will be used. The analysis will be based on data concerning manufacturing companies in Poland.

References

BELLOVARY, J.L., GIACOMINO, D.E. and AKERS, M.D. (2007): A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education, 33(4), 3-41*.

PAWEŁEK, B., KOSTRZEWSKA, J. and LIPIETA, A. (2015), The Problem of Outliers in the Research on the Financial Standing of Construction Enterprises in Poland. In: M. Papież and S. Śmiech (Eds.): *Proceedings of the 9th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-economic Phenomena*. Foundation of the Cracow University of Economics, Cracow.

Keywords

BANKRUPTCY, CLASSIFICATION, FORECASTING, OUTLIERS

Outliers in Topic Modelling

Paweł Lula[1]

Cracow University of Economics, Poland

Abstract. Methods of outliers' identification and studying the influence of outliers on the results of topic modelling performed by Latent Dirichlet Allocation are the main problems discussed in the presentation. Outliers can impact model estimation phase and can modify the results of model outcomes. The distribution of words within a topic definition and the distribution of topics within a document can be changed by untypical words.

The presentation will be divided into following parts:
a) short description of Latent Dirichlet Allocation,
b) studying the influence of outliers on the result of topic modelling,
c) the proposal of outliers' identification method,
d) evaluation of the proposed procedure,
e) conclusions.

During research simulation approach will be used.

Data Homogeneity in Classification of Objects on Local Housing Market

Barbara Batóg[1] and Iwona Foryś[2]

Institute of Econometrics and Statistics, Faculty of Economics and Management, University of Szczecin

Abstract. Real estate market is very dynamic market and is also strongly dependent on changes on social and economic environment. Especially housing market is very sensitive to economic situation of households, changes of demand caused by households and preferences concerning attributes of purchased apartments [Foryś 2011; Batóg, Foryś 2014]. Housing market participants are interested in dependency between price and attributes of apartments during every stage of business cycle [Batóg, Foryś 2011]. This knowledge enables each transaction party to estimate the value of apartment in case of information asymmetry and make a reasonable decision to buy or sell it. The fore mentioned dependency is also important for property appraisers in case of evaluation of apartments by means of comparative approach. In this approach the first step is choice of similar apartments in respect of attributes strongly influencing market value of apartment. Therefore the results of researches on housing market dealing with dependencies between prices and attributes and the results of classifications of purchased apartments taking into account attributes are very useful tool supporting decisions of housing market participants [Batóg, Foryś 2013].

The aims of current research are: comparison of influence of attributes on prices in different stages of business cycles, clustering of apartments according to their similarity and examination of influence of objects' similarity on the quality of clustering.

The presented research is based on information concerning all transactions on local housing market in Szczecin in 2006-2012 found in notary deeds collected by Authors. The transactions were conducted on one of housing estate in Szczecin named ''Zawadzkiego-Klonowica". The choice of this housing estate was caused by such characteristics as constant number of apartments and the same type and technology of buildings. The apartments differ in price, date of sale, area, number of rooms, floor the apartment is located on.

Keywords

HOMOGENEITY, CLASSIFICATION, HOUSING MARKET

Invited Session 3: Data Science in Life Sciences

Chaired by Hans Kestler

Thursday, September 3, 2015: 11:00am - 12:15pm Room: EBS 2.2

Challenge of high-dimensional feature selection for complex time to event endpoints

Maral Saadati, Axel Benner

Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany
benner@dkfz.de

Abstract. For the analysis of genomic information many researchers are faced with the challenges of high-dimensional data. Often feature selection is required to reduce dimensionality to manageable size.

While some methods, such as SCAD and the adaptive lasso, guarantee model consistency and asymptotic unbiasedness under certain conditions, an initial screening step is required to reduce the number of potential predictors to less than the sample size. Also other established methods, such as the lasso, which do not possess the aforementioned theoretical properties, but are able to handle a high-dimensional covariate space can benefit from an initial screening or dimension reduction step in terms of model fit (Paul et al., 2008).

Initially only basic ideas were pursued for dimension reduction, such as univariate score-based screening. However, the seminal paper by Fan and Lv (2008) on sure independent screening (SIS) addressed the need for investigating good screening methods and their properties. In particular for time-to-event endpoints there have been some recent developments. For example, Fan et al. (2010) investigated SIS for the Cox proportional hazards model; Gorst-Rasmussen and Scheike (2013) proposed independent screening based on a model-free statistic.

We present and discuss the impact of existing screening approaches on the analysis of survival data in simulations as well as real-life applications. Furthermore, we propose a novel dimension reduction approach in the context of competing risks models using cause-specific hazards. An added challenge is to incorporate the particular structure of competing risks, e.g. to combine variables with similar effect sizes on several competing hazards or to refrain from filtering variables with small, but opposing, effects on competing hazards.

Finally, all methods are compared with respect to variable selection performance and prediction accuracy of the final model.

References

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B*, 70(5), 849–911.

Fan J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, IMS Collections, 6, 70–86.

Gorst-Rasmussen, A., and Scheike T. (2013). Independent screening for single-index hazard rate models with ultra-high dimensional features. *J. R. Statist. Soc. B*, 75(2), 217–245.

Paul, D., Bair, E., Hastie, T., and Tibshirani, R. (2008). ”Preconditioning” for feature selection and regression in high-dimensional problems. *Annals of Statistics*, 36(4), 1595–1618.

Keywords

SURVIVAL ANALYSIS, HIGH DIMENSION, FEATURE SELECTION

Understanding the Biological Functions of Gene Sets

Catharina Lippmann[1], Jörn Loetsch[2], and Alfred Ultsch[3]

[1] Fraunhofer Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany
lippmanc@mathematik.uni-marburg.de

[2] Institute of Clinical Pharmacology, Goethe University, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany j.loetsch@em.uni-frankfurt.de

[3] DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany ultsch@mathematik.uni-marburg.de

Abstract. Next generation sequencing[1], microarray analysis[2] and genetical database searches, for example, for genes involved in pain[3], produce sets of genes which can contain several hundreds of genes. The central research question for these gene sets is, which biological roles/functions these genes in the organism perform[3]. To answer this question, an Over Representation Analysis (ORA)[4,5] can be applied using the Gene Ontology (GO) knowledge bases[6]. The answer given by an ORA is a directed acyclic graph (DAG)[7]. This DAG is a hierarchical representation of knowledge in form of a graph with nodes containing terms and connected by edges pointing to more detailed descriptions. Such a DAG represents the complete knowledge for the answer and may contain hundreds of GO terms. By its sheer size this obscures the understanding of the main biological functions of the genes[8]. Functional Abstraction derives for this rather unintelligible DAG a small number of topics that highlight different aspects of the functions of the gene set[8]. This allows the identification of new and so far overseen aspects[3]. The abstraction is achieved by first computing a numeric value for all nodes describing the remarkableness of the terms in the DAG. Remarkableness is a function of certainty and information value of a term. For each path from a leave to the root of the DAG the maximum remarkableness identifies an important term, called headline. Subsequently the number of headlines is reduced by subsumption or expanded by detailization using the structural features of the DAG[8]. In this work the division of the complete DAG into separate function specific DAGs which describe the Functional Areas is presented and compared to other approaches on several examples from current research on cancer and pain.

References

[1] METZKER, M.L. Sequencing Technologies — the next Generation, *Nature Reviews Genetics, 11, pp 31-46, 2010.*

[2] QUACKENBUSH, J. Computational Genetics: Computational Analysis of Microarray Data, Nature Reviews Genetics, 2, pp 418-427, 2001.

[3] Loetsch, J., Doehring, A., Mogil, J.S., Arndt, T., Geisslinger, G., Ultsch, A. Functional Genomics of Pain in analgesic Drug Development and Therapy, Pharmacol. Ther., 139, pp 60-70, 2013.

[4] Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Mueller, R., Meese, E., Lenhof, H.-P. GeneTrail - advanced Gene Set Enrichment Analysis, Nucleic Acids Res., 35: pp 186–192, 2007.

[5] Khatri, P., Draghici, S. Ontological Analysis of Gene Expression Data: current Tools, Limitations, and open Problems, Bioinformatics, 21, pp 3587–3595, 2005.

[6] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: Tool for the Unification of Biology, Nature Genet., 25, pp25-29, 2000.

[7] Zhang, S., Cao, J., Kong, Y.M., Scheuermann, R.H. GO-Bayes: Gene Ontology-based Overrepresentation Analysis using a Bayesian Approach, Bioinformatics, 26, pp 905-911, 2010.

[8] Ultsch, A., Loetsch, J. Functional Abstraction as a Method to Discover Knowledge in Gene Ontologies, PLoS One, 9(2), pp. 90-191, 2014.

On variable selection and shrinkage strategies to derive multivariable regression models

Willi Sauerbrei[1] and Hans van Houwelingen[2]

[1] Institute for Medical Biometry and Statistics, University Medical Center Freiburg, Freiburg, Germany wfs@imbi.uni-freiburg.de
[2] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands j.c.van_houwelingen@lumc.nl

Abstract. In many areas of science where empirical data are analyzed, a task is often to identify important variables with influence on an outcome. Most often this is done by using a variable selection strategy. Many have been proposed but each of them is criticized and none is generally accepted.

We will discuss that the aim of a model is the most important criteria to decide whether a specific variable selection strategy may be appropriate. We will stress the difference between models for prediction and for explanation and consider model sparsity as an important criteria. In the context of low-dimensional data we will consider stepwise approaches with and without post-selection shrinkage and the LASSO. Based on results in real examples and a simulation study we will argue that predictions are often very similar, irrespective of the strategy used to derive the model, but that the stop criteria has an important influence on the model selected for explanation, including its sparsity, interpretation and transportability. We conclude that backward elimination followed by post-selection parameterwise shrinkage (PWSF) is a suitable approach, provided that the amount of information in the data is not too small (van Houwelingen and Sauerbrei 2013).

To conduct such an analysis in practice the R package shrink has recently been provided (Dunkler et al 2015). For two or more variables which are associated it extends the present methodology by introducing ”’joint shrinkage factors”’.

References

VAN HOUWELINGEN, H.C. and SAUERBREI, W. (2013): Cross-validation, shrinkage and vaiable selection in linear regression revisited. *Open Journal of Statistics, 3, 79–102, 2013.*

DUNKLER, D., SAUERBREI, W. and HEINZE, G. (2015): Global, parameterwise and joint post–estimation shrinkage. *Journal of Statistical Software, to appear*.

Keywords

SHRINKAGE, VARIABLE SELECTION, LASSO

Invited Session 4: Data Science

Chaired by Adalbert Wilhelm

Thursday, September 3, 2015: 2:35pm - 3:50pm Room: EBS 2.2

Clustering through High Dimensional Data Scaling: Applications and Implementations

Fionn Murtagh[1] and Pedro Contreras[2]

[1] Dept. of Computing and Mathematics, University of Derby, and Dept. of Computing, Goldsmiths University of London fmurtagh@acm.org
[2] Thinking Safe Ltd., Egham pedro.contreras@acm.org

Abstract. Conventional random projection of high dimensional point clouds is to obtain a low dimensional projection of the data cloud. Our use of random projection is to obtain a consistent unidimensional scaling of the data in order to provide an efficient, scalable clustering of the data cloud. In the context of Correspondence Analysis of high dimensional point clouds, we show how data piling (or the concentration of the point cloud) facilitates finding a unidimensional scaling. A hierarchical clustering is then obtained directly from this scaling.

In a number of case studies, we illustrate the major computational benefits of this approach. We focus in particular on applications, in regard to analytics problems addressed, and software implementation environments.

References

CRITCHLEY, F., HEISER, W. (1988): Hierarchical trees can be perfectly scaled in one dimension. *Journal of Classification*, 5, 5—20.

MURTAGH, F. (2015): Big data scaling through metric mapping: exploiting the remarkable simplicity of very high dimensional spaces using Correspondence Analysis. In preparation.

MURTAGH, F., CONTRERAS, P. (2015): Random projection towards the Baire metric for high dimensional clustering. In: A. Gammerman, V. Vovk and H. Papadopoulos, Eds, *Statistical Learning and Data Sciences, Lecture Notes in Artificial Intelligence (LNAI)* Volume 9047, Springer, Berlin, 424-431.

NEWTON, G., CALLAHAN, A., DUMONTIER, M. (2009): Semantic journal mapping for search visualization in a large scale article digital library. In: *13th European Conference on Digital Libraries, ECDL 2009*, ACM.

Keywords

HIERARCHICAL CLUSTERING, BAIRE (ULTRA)METRIC, RANDOM PROJECTION, SEMANTIC VECTORS

An optimization approach to mining big data

Abdellah Salhi[1]

Department of Mathematical Sciences, University of Essex, UK as@essex.ac.uk

Abstract. Symbolic regression is regression with no assumed prior model. It boils down to searching in the space of all possible algebraic models for that which best fits the given dataset. Because it doesn't subsume a model to fit, it is, effectively, a data-mining tool that has the potential to uncover unperceived relationships, unlike traditional regression where we try to fit an assumed model. However, given the size of the space where a good model may reside, it is computationally challenging to search it. Coupled with the likely intractability of handling big data, [2], a new approach to dealing with this problem and that of using symbolic regression to mining it is needed. In this paper a representation of big data that facilitates its handling is suggested. An optimization approach to solving the symbolic regression problem is then described. This approach is a breakaway from the traditional genetic programming used in the past [1]. The approach we advocate is sleeker and requires few arbitrarily set parameters, [3]. Illustrations will be provided where necessary.

References

John Koza, "Genetic Programming: On the Programming of Computers by Means of Natural Selection v. 1", MIT Press, 1993.

Oliver Bracht, "Five ways to handle big data in R", in eoda und Datenanalysis", November 2013.

Abdellah Salhi and Eric S Fraga, "Nature-Inspired Optimisation Approaches and the New Plant Propagation Algorithm", Proceedings of the International Conference on Numerical Analysis and Optimisation (ICeMATH2011), 2011.

Keywords

BIG DATA, DATA-MINING, SYMBOLIC REGRESSION, OPTIMISATION, STRAWBERRY ALGORITHM, PPA

Predicting hidden tourism: Exploratory data analysis and classification of regression coefficients

Claudio Conversano[1], Adalbert F.X. Wilhelm[2], Giulia Contu[1], and Francesco Mola[1]

[1] Università degli Studi di Cagliari, Italy conversano@unica.it | giulia_contu@tiscali.it | mola@unica.it
[2] Jacobs University Bremen, Germany a.wilhelm@jacobs-university.de

Abstract. Tourism is for many regions an important economic factor and its proper planning requires reliable data. There exist forms of tourism that in some regions will not show up in the official registry of arrivals and bed nights spent in the area. In case these unregistered tourism flows do not remain below a threshold of irrelevance, estimation processes are needed that provide a more accurate forecast of tourism for planning purposes. Official data about tourism is collected for individual administrative communities in regular time intervals, for example monthly. Hence, panel data regression methods are a straightforward tool to build prediction models for tourism flows. This leads to quite a number of different model specifications that can be used to fit to the data at hand. Since, additionally, different proxies are available to identify the amount of hidden tourism we are faced with an experimental set-up defined by potential models and potential proxies even when using the same set of predictors in each model. Running all these different models yield a set of regression models that need to be structured to identify the most cogent prediction for the actual tourism flow. In order to do so, we apply numerical and visual cluster approaches. As a practical example, we use tourism data of municipalities in the Provincia di Oristano (District of Oristano), one of the four oldest districts in Sardinia. In order to estimate non-registered presences, various indicators for the two concepts urban waste and energy consumption have been used.

References

BALTAGI, B.H. (2013). *Econometric Analysis of Panel Data.* Chichester: Wiley.

CROISSANT, Y. , Millo, G. (2008). Panel Data Econometrics in R: The *plm* Package. *Journal of Statistical Software, 27(2), 2008*

Keywords

PANEL DATA, CLUSTERING, TOURISM, REGRESSION COEFFICIENTS

Invited Session 5: Big Data

Chaired by Berthold Lausen

Thursday, September 3, 2015: 4.15pm - 5:30pm Room: EBS 2.2

New Challenges in Analysing Big Data

Maurizio Vichi[1]

Sapienza University of Rome Maurizio.vichi@uniroma1.it

Abstract. Big Data frequently describe complex economic, social and demographic phenomena that manifest on individuals (units, objects, sites, with a spatial location), by means of a set of variables and showing both a diffusion over space and an evolution over time. These data show different relations between, objects (spatial correlation), between variables (cross-sectional correlation) and between times (time series correlation) that need to be analysed. Three or high dimensional arrays (data (iper)-cubes), are used to rearrange the huge number of statistical units (rows) with a spatial location, variables, (columns) and times (tubes). A modelling approach is proposed for different statistical analysis such as clustering and structural equation modelling.

Keywords

CLUSTERING; STRUCTURAL EQUATION MODELLING

Clustering and Classification of Infrared Hyperspectral Aerial images

Alfred Ultsch[1,3] and Andrew McGrath[2]

[1] Databionics Research Group, University of Marburg, Marburg, Germany
ultsch@informatik.uni-marburg.de
[2] Flinders University - Airborne Research Australia
andrew.mcgrath@airborneresearch.com.au
[3] Science Division, Academic Fyling (AKAlfieg), University of Frankfurt, Frankfurt, Germany

Abstract. High resolution aerial images of the hyperspectral infrared bands promise important advances in environmental and ecological research [1]. In 2012 and 2013, an experiment was conducted in Australia to determine the impact of continual sub-surface release of carbon dioxide on the growth of agricultural crops. Airborne hyperspectral imagery was one of the techniques employed to detect the effects of the CO2. However, standard hyperspectral analysis techniques and indices were unable to distinguish between plants affected by the CO2 and those stressed by other factors [1]. The hyperspectral data of the test plot are in 488 bands from 400 to 2500nm [2]. They are georeferenced and resampled onto a 0.5m resolution grid [1]. Here we report an application of Data Mining techniques [3] for a deeper analysis of the data to seek for a hyperspectral signature able to identify the CO2,-stressed vegetation specifically. A careful preprocessing allowed a data reduction by a factor of ca. 10. The remaining spectral bands were clustered using an emergent self-organizing map (ESOM) [4]. The clusters obtained could be related to the location of the healthy and stressed plants. This work shows the suitability and efficacy of modern Data Mining and Knowledge Discovery techniques for hyperspectral images.

References

[1] Guan, H., McGrath, A., Bennett, J., Zhu, C., Clay, R., Ewenz, C. (2011), Comparison of high-resolution aerial thermal imagery with air temperature in Adelaide, In: International Workshop on Urban Weather and Climate: Observation and Modeling, International Workshop on Urban Weather and Climate: Observation and Modeling, Beijing.

[2] Grahn, H, Geladi,P. (2007), Techniques and Applications of Hyperspectral Image Analysis, Wiley, Chichester.

[3] Hand, D., Mannila, H., Smyth, P. (2001), Principles of Data Mining, MIT Press, Cambridge MA.

[4] Ultsch, A. (2003), Maps for the Visualization of high-dimensional Data Spaces, In Proceedings Workshop on Self-Organizing Maps (WSOM 2003), Kyushu, Japan, pp, 225-230.

Mining for retail navigation and wearable data patterns

Rolando Medellin-Gasque[1,2], Henrik Nordmark[1], Anthony Mullen[1], Aris Perperoglou[2], Luca Citi[2], and Berthold Lausen[2]

[1] Profusion, London
[2] University of Essex, Wivenhoe Park, Colchester CO4 3SQ `rmedel@essex.ac.uk`

Abstract. E-commerce has revolutionized the way consumers browse and shop on-line. Personal browsing-logs can reveal on-line shopping-behaviour patterns that can be used to improve e-commerce in general. In this paper we look into a dataset comprised 171 variables including personal browsing-logs, wearables data, sentiment surveys, and demographic data from 28 employees in a controlled experiment over a period of 10 days. While the experiment focused on retrieving data from the wearable devices and measure the impact on employee well-being, a significant amount of data was obtained from people browsing and shopping on-line. Our aim is to mine browsing-logs across retail websites (document understanding) and look for correlations with the whole dataset (user understanding), especially wearables data [JIANG13]. More precisely, the objective is to identify purchase paths by defining a generic taxonomy for retail websites and cluster similar paths based on employee data. Purchase paths can be classified for example by the type and number of webpages accessed, the type of webpages opened at the same time, the type of user, the time to buy, etc., and can be further classified by incorporating wearables data (e.g., data on the sleeping behaviour the night before, heartbeat minute-readings, etc.). The resulting clusters can be analysed to identify critical paths to buy or send tailored adverts and, in general, understand on-line retail purchase behaviours. While the volume of the dataset id not significant, the uniqueness of this dataset allow us to build scalable models focused wearables and browse-logs data.

References

JIANG, D., PEI, J., LI, H. (2013). Mining search and browse logs for web search: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 4(4), 57.

Keywords

ECOMMERCE, NAVIGATION PATTERNS, WEARABLES DATA, BROWSER LOGS, HEARTBEAT DATA, RETAIL DATA

Part III

Contributed Sessions

Clustering I

Wednesday, September 2, 2015: 11:15am - 12:55pm Room: EBS 2.34

Validation of K-means Clustering: Why is Bootstrapping Better Than Subsampling?

Hans-Joachim Mucha

Weierstrass Institute for Applied Analysis and Stochastics (WIAS), 10117 Berlin,
Mohrenstraße 39, Germany, mucha@wias-berlin.de

In simulation studies based on many synthetic and real datasets, we found out that subsampling has a much weaker behavior in the finding of the true number of clusters K than bootstrapping (Mucha and Bartel 2013, Mucha 2015). But why? Based on further investigations, especially K-means clustering with the comparison of bootstrapping and a special version of subsampling named ”Boot2Sub”, we try to answer this question. Subsampling is resampling taken without replacement from the original data. Here a parameter, the cardinality of the drawn sample, is needed which causes usually a serious problem. The way out would be to take a bootstrap sample but discard multiple points. We call such a special subsampling scheme ”Boot2Sub”.

References

MUCHA, H.-J. and BARTEL H.-G. (2013): Soft Bootstrapping in Cluster Analysis and Its Comparison with Other Resampling Methods. In: M. Spiliopoulou, L. Schmidt-Thieme and R. Janning (Eds.): *Data Analysis, Machine Learning and Knowledge Discovery*. Springer, Berlin, forthcoming.

MUCHA, H.-J. (2015): Assessment of Stability in Partitional Clustering Using Resampling Techniques. Proceedings of the German-Polish Workshop in Dresden, KIT Karlsruhe, forthcoming.

Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters

Christian Hennig[1] and Chien-Ju Lin[2]

[1] Department of Statistical Science, University College London Gower St, London WC1E 6BT, United Kingdom c.hennig@ucl.ac.uk

[2] MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 0SR, United Kingdom chienju@mrc-bsu.cam.ac.uk

Abstract. Many cluster analysis methods deliver a clustering regardless of whether the dataset is indeed clustered or homogeneous, and need the number of clusters to be fixed in advance. Validation indexes such as the Average Silhouette Width are popular tools to measure the quality of a clustering and to estimate the number of clusters. Such indexes can be used for testing the homogeneity hypothesis against a clustering alternative by exploring their distribution, for a given number of clusters fitted by a given clustering method, under a null model formalising homogeneous data. This is done by parametric bootstrap here. The same approach can be used for assessing the number of clusters by comparing what is expected under the null model with what is observed under different numbers of clusters. Many datasets include some structure such as temporal or spatial autocorrelation that distinguishes them from a plain Gaussian or uniform model, but cannot be interpreted as clustering. Applications will be presented.

References

HENNIG, C. and LIN, C.-J. (2015) Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. http://arxiv.org/abs/1502.02574, to appear in *Statistics and Computing*.

Keywords

CLUSTER VALIDATION, MIXTURE MODEL, DISTANCE-BASED CLUSTERING

Tuning hierarchical clustering with domain knowledge

Johann M. Kraus[1] and Hans A. Kestler[2]

[1] Core Unit Medical Systems Biology, Ulm University, Ulm, Germany
johann.kraus@uni-ulm.de

[2] Leibniz Institute for Age Research, Fritz-Lipmann Institute, Jena, Germany
hkestler@fli-leibniz.de

Abstract. Cluster analysis subsumes a variety of methods from explorative data analysis that are used for detecting unknown structures hidden in a data set. These methods do not make use of domain knowledge about any possible grouping of the data. Many partitional cluster algorithms were adapted to make use of this kind of background information either by constraining the search process or by modifying the underlying metric. Limitations in the reproducibility of clustering results after small modifications of the data set, triggered the inclusion of domain knowledge into the hierarchical clustering process. Based on our previous work (Kestler et al. 2006, Kraus et al. 2007), we present a general framework for including domain knowledge into a hierarchical clustering process. Our new semi-supervised cluster strategy aims to assess the reliability of hierarchical clustering. Reliable clusters may be identified by searching for the most stable partitions under different clustering conditions, e.g. resampling data.

References

KESTLER, H.A., KRAUS, J.M., PALM, G., SCHWENKER, F. (2006), On the effects of constraints in semi-supervised hierarchical clustering. In: F. Schwenker, S. Marinai (Eds.): *Artificial neural networks in pattern recognition*. Springer, Berlin, 57–66.

KRAUS, J.M., PALM, G., KESTLER, H.A. (2007), On the robustness of semi-supervised hierarchical graph clustering in functional genomics. *5th International Workshop on Mining and Learning with Graphs*. Florenz, 147–150.

Keywords

SEMI-SUPERVISED CLUSTERING, ROBUSTNESS ANALYSIS, DOMAIN KNOWLEDGE

Data Analysis I

Wednesday, September 2, 2015: 11:15am - 12:55pm Room: EBS 2.1

Reduction of Dimensionality for Discrimination

Cuevas-Covarrubias C.[1] and Riccomagno E.[2]

[1] Anahuac University, Actuarial Sciences Faculty
[2] The University of Genova, Department of Mathematics

Abstract. This paper presents an algorithm for reduction of dimensionality useful in statistical classification problems where observations from two multivariate normal distributions are discriminated. It is based on Principal Components Analysis and consists of a simultaneous diagonalization of two covariance matrices. The criterion for reduction of dimensionality is given by the contribution of each principal component to the area under the ROC curve of a discriminant function. Linear and quadratic scores are considered, the focus being on the quadratic case. Similarities with analogous methods in the literature are discussed. Practical examples conclude the paper.

Keywords

MULTIVARIATE NORMAL, STATISTICAL CLASSIFICATION, AREA UNDER THE ROC CURVE, BINOMIAL MODEL.

Taxicab Correspondence Analysis of Sparse Contingency Tables

Vartan Choulakian

Université de Moncton,New Brunswick,Canada vartan.choulakian@umoncton.ca

Abstract. Visualization and interpretation of contingency tables by correspondence analysis (CA), as developed by Benzecri, has a rich structure based on Euclidean geometry. However, it is a well established fact that, often CA is very sensitive to sparse contingency tables, where we characterize sparsity as the existence of relatively high-valued cells. These include two widely discussed particular cases : Rare observations discussed by Rao (1995) and zero-block structure emphasized by Novak and Bar-Hen (2005) and Greenacre (2013). In this talk, we aim to emphasize the important roles played by L1 and L2 geometries (aka CA and TCA) in the visualization and interpretation of sparse contingency tables. Examples are provided.

References

CHOULAKIAN,V. (2006).Taxicab correspondence analysis. *Psychometrika, 71, 333-345.*

CHOULAKIAN,V. SIMONETTI, B. and GIA, T.P. (2014). Some further aspects of taxicab correspondence analysis. *Statistical Methods and Applications, 23, 401-416.*

GREENACRE, M. (2013).The contributions of rare objects in correspondence analysis. *Ecology, 94(1), 241-249.*

NOWAK, E. and BAR-HEN, A. (2005). Influence function and correspondence analysis. *Journal of Statistical Planning and Inference, 134, 26-35*

RAO, C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Questiió, 19, 23-63.*

Keywords

CORRESPONDENCE ANALYSIS, TAXICAB CORRESPONDENCE ANALYSIS, RELATIVELY HIGH-VALUED CELLS, RARE OBSERVATIONS, ZERO-BLOCK STRUCTURE

Unbiased estimation for linear regression when $n < v$

Saeed Aldahmani[1] and Hongsheng Dai[2]

[1] Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK
skaald@essex.ac.uk
[2] Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK
hdaia@essex.ac.uk

Abstract. A novel method is proposed for solving the linear regression problems when the number of observations n is smaller than the number of predictors v in an unbiased way. The proposed method is based on the idea of graphical models and provides unbiased parameter estimates under certain conditions, whereas the existing methods such as ridge regression, LASSO and least angle regression (LARS) give biased estimates. The new method is aimed to provide a detailed graphical correlation structure for the predictors, so that the real causal relationship between predictors and response could be identified. In contrast, existing methods often cannot identify the real important predictors which have possible causal effects on the response variable. To evaluate the proposed method, real and simulated data sets are used and the results are compared with ridge regression, LASSO and LARS. It is revealed by our experiments that the proposed method is better than all the other methods, especially, when $n < v$.

References

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004): Least angle regression. *The Annals of Statistics, 32, 407–499, 2004*.

Hoerl, A.E. Kennard R.W. (1970): Ridge regression: Biased estimation for nonorthogonal problems.*Technometrics, 12, 55–67, 1970* .

Lauritzen, S. (1996):*Graphical Models*. Oxford Univeristy, Press.

Tibshirani, R. (1996):Regression shrinkage and selection via the LASSO.*Journal of the Royal Statistical Society, 33, 267–288, 1996*.

Keywords

GRAPHICAL LEAST SQUARES ESTIMATE, LARS, LASSO, RIDGE REGRESSION.

Variable Selection in Multi-Label Classification using Probe Variables

Trudie Sandrock and Sarel Steel

Department of Statistics, University of Stellenbosch, South Africa sjst@sun.ac.za

Abstract. Multi-label classification problems arise in scenarios where every data instance can be associated simultaneously with more than one of several available labels. Application areas include music information retrieval, bioacoustics, text and image annotation. Variable selection in a multi-label context is even more challenging than in the single label case, and additional complexity is introduced by the fact that variables which may discriminate well between values of one of the responses will not necessarily do the same for the other responses. In this regard the concepts of local and global relevance of variables are defined in this paper. A multi-label variable selection procedure should take cognisance of the possibility that some variables may not be globally relevant, but could be locally relevant for one or more labels.

We propose a multi-label variable selection method, based on a binary relevance problem transformation. Different measures of variable importance are considered as filters. Probe variables are generated by randomly permuting variable values, and these probes are used to determine the number of variables to be selected. Empirical results obtained from applying our proposed technique as well as existing techniques (Spolaor et al, 2013) to benchmark datasets are reported. These results show that our technique performs marginally better, and simultaneously provides output that can be used to ascertain the local and global relevance of variables.

References

SPOLAOR, N., CHERMAN, E.A., MONARD, M.C. and LEE, H.D. (2013): A Comparison of Multi-Label Feature Selection Methods using the Problem Transformation Approach. *Electronic Notes in Theoretical Computer Science, 292, 135–151.*

TUV, E., BORISOV, A. and TORKKOLA, K. (2008), Ensemble-Based Variable Selection using Independent Probes. In: H. Liu and H. Motoda (Eds.): *Computational Methods of Feature Selection*. Chapman and Hall/CRC, 131–146.

Keywords

MULTI-LABEL, PROBE VARIABLES, SELECTION

Engineering, Logistics and Optimisation

Wednesday, September 2, 2015: 11:15am - 12:55pm Room: EBS 2.65

Electricity Load Forecasting Using Data Mining Techniques

Aziz Guergachi[1], Muhammad Shahbaz[2], Faqia Saeed[3], Sharmeen Zahra[3]

[1] Ryerson University, 350 Victoria Street, Toronto, Canada, a2guerga@ryerson.ca
[2] University of Engineering and Technology, Lahore, muhammad.shahbaz@gmail.com
[3] University of Engineering and Technology, Lahore

Abstract. Because of the lack of mature technologies for electricity storage, electric utilities find it challenging to build effective buffers for the smooth operation of power grids. Such a situation had previously put the Province of Ontario in the ironic position of having to pay neighbouring provinces and states as much as 22 cents a kilowatt hour to take its surplus power. On the other hand, in many developing countries, the infrastructures are so weak that the demand for electricity always exceeds the power supply and rolling blackouts are common or even normal daily events. In this research, we re-examine load forecasting techniques with the intention to not only address the needs of advanced economies to become more efficient, but to also extract patterns from the data sets that are readily available in these advanced economies to help with the planning of infrastructure development in developing nations.

References

Spears, J. (2011): Ontario pays others to take its surplus power. *Toronto Star, 2011.*

Farah, Y. and Sharif, M. (2014): Forecasting Electricity Consumption for Pakistan. *Journal of Emerging Technology and Advanced Engineering, Vol. 4 (4), 496-503, 2014.*

Corradi, O., et al. (2013): Controlling electricity consumption by forecasting its response to varying prices. *IEEE Transactions on Power Systems, 28(1), 421-429, 2013*

Keywords

ELECTRICITY LOAD FORECASTING, ELECTRICITY PRICES, WEATHER CONDITIONS, EXTRACTING PATTERNS FOR THE PLANNING OF INFRASTRUCTURE DEVELOPMENT IN DEVELOPING COUNTRIES.

Operating Context Estimation of Mobile Work Machines with Bayesian Inference and Machine Learning

Teemu Väyrynen, Suvi Peltokangas, Eero Anttila, and Matti Vilkko

Tampere University of Technology, Korkeakoulunkatu 10, 33720 Tampere, 33720, Tampere, Finland. firstname.lastname@tut.fi

Abstract. This paper presents a data-driven approach for estimating the operating conditions i.e. context of the mobile work machines with Bayesian inference and machine learning. Performance analysis and optimization of the mobile machines working in wide range of operating conditions has become an increasingly important trend during the recent years. One of the most interesting and potential approach for improving the performance of the mobile work machines is the utilization of large machine data bases and data-driven analysis methods. Machine learning algorithms can be used to organize data bases into operating context specific segments and search valuable information within each segment for the performance analysis of the machines.

In order to utilize this context specific reference information efficiently for an individual mobile work machine, an automated estimation method for operating context is required. In this work, first a machine learning model is created based on the large reference machine data base. Then this model is applied to the measurement signals of an individual mobile work machine to produce an estimate of an operating context. Finally, Bayesian inference is applied recursively to the context estimates to produce an estimate of the operating context with reliability information. The results of this paper demonstrate an industrial example of the approach being applied in the operating context estimation of a large fleet of mobile work machines.

References

GELMAN, A. et.al. *Bayesian Data Analysis*, United States of America, Chapman and Hall/CRC, 2000.

BISHOP, C. et.al. *Pattern Recognition and Machine Learning*, United States of America, Springer, 2009.

Keywords

MOBILE WORK MACHINE, BAYESIAN INFERENCE, MACHINE LEARNING, ESTIMATION

Modeling Mobile Work Machine Data with Machine Learning Methods

Teemu Väyrynen, Suvi Peltokangas, Eero Anttila, and Matti Vilkko

Tampere University of Technology, Korkeakoulunkatu 10, 33720 Tampere, 33720, Tampere, Finland. firstname.lastname@tut.fi

Abstract. This paper presents a comparison study of machine learning methods used in the performance modeling of the mobile work machines. Performance analysis and optimization of the mobile machines working in wide range of operating conditions has become an increasingly important trend during the recent years. One of the most interesting and potential approach for improving the performance of the mobile work machines is the utilization of large machine data bases and data-driven analysis methods. Machine learning algorithms can be used to organize data bases into operating context specific segments and search valuable information within each segment for machine analysis.

In this work, multiple machine learning methods, such as regression, decision trees, ensemble learning, and neural networks are applied to the mobile work machine data. The main goal of the work is to evaluate the modeling performance of the machine learning methods with the particular mobile work machine data. The methods are evaluated also based on application specific requirements including model structure complexity, execution time, etc. The complex nature of mobile work machines as well as the influence of the machine operator generate significant challenges for predicting the performances of machines. The results of this paper demonstrate an industrial example of the machine learning methods being applied in the data base of a large fleet of mobile work machines.

References

BISHOP, C. et.al., *Pattern Recognition and Machine Learning*, United States of America, Springer, 2009.

BREIMAN, L. et.al., *Classification and Regression Trees*, Monterey CA, Wadsworth and Brooks, 1984.

Keywords

MOBILE WORK MACHINE, MACHINE LEARNING, MODELING, PERFORMANCE

Selection of discriminant heterogeneous variables using multicriteria optimization

Hasna CHAMLAL[1], Tayeb OUADERHMAN[1], and Mehdi BAZZI[1]

LIAD Laboratory, fsac, Hassan II university, Casablanca, Morocco chamlal@yahoo.com

Abstract. . Given an heterogeneous explanatory variables, the problem is how to select the most pertinent. In this paper, we introduce an algorithm of variables selection in the problem of discrimination, which deal with heterogeneous data. The proposed algorithm calls for the criterion of discrimination introduced by Chamlal and Chah [1] (ψ_{cor}), and is based on a concordance measure between several heterogeneous variables (ψ_W), that we introduce here. We calculate the distribution of these coefficients under the assumption of independence.The algorithm , we propose here, is a version of forward stepwise selection without fixing a priori the number of variables to be selected by optimizing the two criterions, using the multicriteria optimization. To assess the performance of the proposed algorithm, an example of credit scoring medeling is provided.

References

CHAH, S. and CHAMLAL, H. (2000): Nouvelle approche pour la sélection des variables discriminantes.*Revue de la statistique Appliquée, tome 48,n4, 59–82*

KENDALL, M.G (1962):*Rank correlation methods*. Griffin,Londres

WIERZBICKI, A.P.(1980), The use of the reference Objectives in Multiobjective Optimization. In: G.Fandel and T.Gal(Eds), *Multiple Objective Making, Theory and Application*, Springer-Verlag, New York.

Keywords

PREORDONNANCE, CONCORDANCE, MULTICRITERIA OPTIMIZATION

Machine Learning and Knowledge Discovery I

Wednesday, September 2, 2015: 3:20pm - 4:35pm Room: EBS 2.34

Partitioning high-dimensional multivariate self-affine time series into differentiated subseries

Christopher M. Taylor[1]

University of Essex cmtayl@essex.ac.uk

Abstract. Given a multivariate time series, possibly of high dimension, with unknown and time-varying joint distribution, it is of interest to be able to completely partition the time series into disjoint, contiguous subseries, each of which has different distributional or pattern attributes from the proceeding and succeeding subseries. An additional feature of many time series is that they display self-affinity, so that subseries at one time scale are similar to subseries at another after application of an affine transformation. Such qualities are observed in time series from many disciplines, including biology, medicine, economics, finance and computer science. This paper defines the relevant multiobjective optimization problem with limited assumptions as a biobjective problem, maximizing difference between successive partitioned subseries and minimizing the difference between these coarse-grained subseries of the entire data set and fine-grained sub-subseries of some subseries of the multivariate time series. A specialized evolutionary algorithm is presented which finds optimal self-affine time series partitions with a minimum of choice parameters. The algorithm not only finds partitions for all possible numbers of subseries given data constraints, but also for self-affinities between different subseries and the entire data set. The resulting set of Pareto-efficient solution sets provides a rich representation of the self-affine properties of a multivariate time series at different locations and time scales.

References

MORALES,R. DI MATTEO,T. and ASTE,T. Dependency structure and scaling properties of financial time series are related. *Scientific reports, 4:4589, January 2014.*

PELLETIER J.D. and TURCOTTE D.L.. Self-affine time series: II. Applications and models. *Adv. Geophys 40, 91-166, 1999.*

WHITLEY, D., SORAYA R., and HECKENDORN R.B. The island model genetic algorithm: On separability, population size and convergence. *Journal of Computing and Information Technology 7, 33-48,1999.*

Keywords

SELF-AFFINE, TIME SERIES, MULTIVARIATE, HIGH-DIMENSIONAL, EVOLUTIONARY ALGORITHM, MULTIOBJECTIVE OPTIMIZATION

Ordinal signatures for prototype-based classifiers.

Andre Burkovski[1], Lyn Rouven Schirra[1], Ludwig Lausser[2], and Hans A. Kestler[2]

[1] Core Unit Medical Systems Biology, Ulm University, Ulm, Germany
{andre.burkovski, lyn-rouven.schirra}@uni-ulm.de
[2] Leibniz Institute for Age Research, Fritz-Lipmann Institute, Jena, Germany
{llausser, hkestler}@fli-leibniz.de

Abstract. The identification of prototypical patterns is one of the major goals in the classification of gene expression profiles. In this context feature selection methods are used to extract signatures of valuable markers. Prototype-based classifiers are of special interest, since they allow a direct biological interpretation. In this work we present prototype-based classifiers based on ordinal-scaled signatures.

The advantage of signatures on an ordinal scale is their invariance to a wide range of data transformations. Standard prototype-based classifiers may be adapted for this type of data. In order to process ordinal-scaled data, adapted instance-based and centroid-based classifiers can rely on rank-distances and rank-aggregation procedures, respectively.

Our experiments reveal that the proposed techniques result in the construction of signatures that improve the classification performance of prototype-based classifiers. The modified algorithms achieve classification results comparable to state-of-the-art classifiers while allowing for an easier interpretation.

Keywords

PROTOTYPE-BASED CLASSIFICATION, ORDNAL-SCALED DATA, RANK AGGREGATION

Offline algorithm selection based on multicriteria optimization of contradicting performance measures

Daniel Horn[1], Aydin Demircioglu[2], Bernd Bischl[3], Tobias Glasmachers[2], and Claus Weihs[1]

[1] Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany
{dhorn, weihs}@statistik.tu-dortmund.de
[2] Ruhr-Universität Bochum, 44780 Bochum, Germany {aydin.demircioglu, tobias.glasmachers}@ini.rub.de
[3] LMU München, 80539 München, Germany
bernd.bischl@stat.uni-muenchen.de

Abstract. In a world of ever-growing science new algorithms are frequently published. But how to decide which algorithm to use for solving a given problem? It is easy to say which algorithm to choose with respect to a single performance measure. Saying which algorithm to choose with respect to several, possibly contradicting measures is a lot harder, since the best algorithm can depend on a trade-off between the criteria. Moreover, most algorithms come along with a set of hyperparameters that can influence the trade-off. In general a single best algorithm does not exist, but a Pareto front of non-dominated algorithms.

In multicriteria optimization several methods for the estimation of a single Pareto front have been devised. Since both the optimization and single performance estimations may be stochastic, multiple optimization runs per algorithm are required, resulting in many Pareto fronts. Our goal is the estimation of the common Pareto front from these. In a first step all algorithm having a relevant influence on the Pareto front are computed, in a second step their order is identified. This results in a single reduced common Pareto front containing only relevant algorithms.

We apply our method in the setup of training Support Vector Machines (SVM). Though SVMs can be regarded as one of the best classifiers, especially in the big data domain their training is very time-consuming. Thus, several different algorithms for approximating the SVM problem have been proposed, which trade lower accuracy for faster training times. We estimate the Pareto front of several approximative SVM algorithms on benchmark datasets. By applying our analysis method we can estimate the reduced common Pareto front and give recommendations on which algorithms should be used.

Keywords

ALGORITHM SELECTION, MULTI-OBJECTIVE OPTIMIZATION, SUPPORT VECTOR MACHINE, LARGE SCALE, PARAMETER TUNING

Data Analysis II

Wednesday, September 2, 2015: 3:20pm - 4:35pm Room: EBS 2.1

Using Landmark Models and Reduced Rank Regression to identify Dynamic Effects on Gene Expression Data.

Aris Perperoglou[1] and Hans van Houwelingen[2]

[1] University of Essex, Colchester, UK, aperpe@essex.ac.uk
[2] Leiden University, Leiden, The Netherlands j.c.van_houwelingen@lumc.nl

Abstract. Consider the problem of identifying differentially expressed genes that have an influence on the survival function of a person. Biological reasoning indicates that effects of some genes might be time-dependent. Some genes may showcase a an early effect while there might be genes with a late effect. A Cox Proportional Hazards model will only be able to flag genes as significant when their effect remain constant throughout. However, we would like to also identify genes that have a dynamic behaviour over time, or genes that their effect becomes visible at some latter stage of the study.

We will address this problem by the use of univariable landmark models. The *Sliding Window Landmark Model* estimates hazard from a given landmark point t_{LM} up to a horizon time point t_{LM+w} for a window of width w. The time period is divided into a grid of landmark points, and a prediction from landmark point t_{LM} to t_{LM+w} is given by a simple Cox model within each window. We will illustrate this approach in a dataset of 295 breast cancer patients, diagnosed between 1984 and 1995 with information on 4919 genes. We will employ *Reduced Rank Regression* methods to gain a better understanding in the nature of the time dependent effects and how the estimation of parameters is influenced by the landmark time points. We will review theoretical aspects and show a practical implementation of our approach.

References

van Houwelingen HC (2007): Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics, 34, 70–85, 2007.*

Perperoglou A., Keramopoullos A., van Houwelingen HC (2007): Approaches in modelling long term-survival: an application to breast cancerq. *Statistics in Medicine, 26, 2666–2685, 2007.*

Keywords

NON-PROPORTIONALHAZARDS, GENE EXPRESSION, DYNAMIC SURVIVAL MODELLING

Developing a Case-mix Classification for Child and Adolescent Mental Health Services in England and Wales

Ms A. Macdougall[1], Dr. P. Martin[1] and Dr. A. Whale[1]

[1] Evidence Based Practice Unit (EBPU), University College London and Anna Freud Centre, 4-8 Rodney St, London amy.macdougall@annafreud.org

Abstract. Funding and payment systems are contentious issues within Child and Adolescent Mental Health Services (CAMHS), as they are in the NHS in general. There is evidence of severe underfunding of some CAMH services, as well as of wide regional variations in funding. The CAMHS Payment Systems project was commissioned to develop a casemix classification for CAMHS, with the aim of informing efforts to achieve a fairer distribution of resources. As part of this work, clinical records from eleven CAMH services were collected over 22 months. These records included clinician ratings of presenting problems, as well as contextual problems and complexity factors.

The aim was to create groups who were similar in terms of resource use and presenting problems. Three ways of classifying patients were compared. The first two had a statistical basis: unsupervised cluster analysis using a k-medoids algorithm on the presenting problems; supervised cluster analysis using recursive partitioning to find the combination of presenting problems that best predicted resource use. The third classification method was derived by a psychiatrist and psychologist on the basis of clinical judgement, and resulted in an algorithm that allocates patients to categories based on the treatment need implied by presenting information.

We used cross-validation on 10 stratified random test samples to compare the three methods in terms of their ability to predict resource use, using mixed negative binomial regression. The results suggest that unsupervised cluster analysis leads to the poorest prediction of resource use, while regression trees did not identify a reliable classification. The algorithmic classification based on clinical judgement provided the best reliable prediction of resource use, although between-service variation was large relative to the variance explained by case-mix. The resulting case-mix classification is clinically meaningful and has potential applications in CAMHS resource planning. Our presentation will also offer reflections on the reasons why statistical methods were outperformed by clinical judgement.

Keywords

CLUSTER ANALYSIS, RECURSIVE PARTITIONING, CAMHS, GLMM

Evaluation of Risk of Drug Consumption

E. Fehrman[2], A. K. Muhammad[1], E. M. Mirkes[1], V. Egan[3], and A. N. Gorban[1]

[1] University of Leicester, Leicester, LE1 7RH, UK ag153@le.ac.uk
[2] Rampton Hospital, Retford, Nottinghamshire, DN22 0PD, UK
[3] University of Nottingham, Nottingham, NG8 1BB, UK

Abstract. The study has two purposes: firstly, to identify the association of personality profiles (i.e. NEO-FFI-R [1], BIS, and ImpSS), demographics, and drug consumption; secondly, to predict the risk of drug consumption for each individual. The personality profiles for drug users are associated with a high score on neuroticism and a low score on conscientiousness [2]. The problem of risk evaluation for individuals has been approached recently [3]. They employed various machine learning algorithms for 'drug user/non user' classification problem. We evaluated the individual drug consumption risk for each drug. We used various methods: decision tree, random forest, *k*NN, discriminant analysis, Gaussian mix, probability density function estimation, logistic regression, and naïve Bayes and selected the most effective method for each drug. Sensitivity and specificity (evaluated by LOOCV) greater than 75% were obtained for VSA (volatile substance abuse) and methadone. Sensitivity and specificity greater than 70% were achieved for amphetamines, cannabis, cocaine, crack, ecstasy, heroin, ketamine, legal highs, and nicotine.

References

EGAN, V., DEARY, I., AUSTIN, E. (2000): The NEO-FFI: emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and A. *Personality and Individual Differences, 29, 907–920.*

TERRACCIANO, A., LÖCKENHOFF, C.E, CRUM, R.M., BIENVENU, O.J., COSTA, P.T. (2008): Five-Factor Model personality profiles of drug users. *BMC Psychiatry, 8 (1) (2008), 22.*

VALERO, S., DAIGRE, C., RODRÍGUEZ-CINTAS, L., BARRAL, C., GOMÀ-I-FREIXANET, M., FERRER, M., CASAS, M., RONCERO, C. (2014): Neuroticism and impulsivity: Their hierarchical organization in the personality characterization of drug-dependent patients from a decision tree learning perspective. *Comprehensive psychiatry, 55, (5), 1227–1233.*

Keywords

RISK, DRUGS, MACHINE LEARNING

Machine Learning and Knowledge Discovery II

Thursday, September 3, 2015: 11:00am - 12:15pm Room: EBS 2.34

Classification with the kernelized α-procedure

Tatjana Lange[1], Pavlo Mozharovskyi[2], and Oleksii Pokotylo[2]

[1] Merseburg University of Applied Sciences
[2] University of Cologne pokotylo@wiso.uni-koeln.de

Abstract. The idea of the kernel trick is to work methodologically in a rich extended space of features while performing computation applying kernel functions in the original usually low-dimensional space. By introducing kernels to the generalized portrait method (now known as Support Vector Machine, SVM), Vapnik (1998) has demonstrated the power of kernels, what gave rise to numerous applications of the methodology. Exploiting the idea of the generalized portrait of Vapnik and Lerner (1963), SVM maximizes the margin between separable parts of classes and balances between its width and amount of errors. Similarly, and being developed in parallel, the α-procedure (Vasil'ev, 1991; Vasil'ev and Lange, 1998) constructs a separating rule iteratively by minimizing the empirical risk in two-dimensional coordinate subspaces. In the current project, kernels are introduced to the α-procedure by employing it in a finite-dimensional explicit approximation of a reproducing kernel Hilbert space (RKHS). By its inductive character and natural robustness, the α-procedure selects a subspace of RKHS and does not need tuning a balancing parameter. This yields competitive classification performance and a high speed of training and classification. An experimental study compares the α-procedure and SVM employing a number of kernel functions.

References

VAPNIK, V.N. (1998): *Statistical learning theory*. Wiley, New York.

VAPNIK, V.N. and LERNER, A.Ya. (1963): Pattern recognition using generalized portrait method. *Automation and Remote Control, 24, 774-780.*

VASIL'EV, V.I. (1991): The reduction principle in pattern recognition learning (PRL) problem. *Pattern Recognition and Image Analysis, 1, 1.*

VASIL'EV, V.I. and LANGE, T.I. (1998): The duality principle in learning for pattern recognition. *Cybernetics and Computing Technique, 121, 7-16.*

Keywords

ALPHA-PROCEDURE, KERNEL TRICK, REPRODUCING KERNEL HILBERT SPACE, SUPPORT VECTOR MACHINE.

CNN-FM: Personalized Content-Aware Image Tag Recommendation

Hanh Thi Hong Nguyen, Martin Wistuba and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab, University of Hildesheim, Germany
{nthhanh,wistuba,schmidt-thieme}@ismll.de

Abstract. Social media services allow users to share and annotate their resources freely with keywords or tags that have valuable information to support organizing or searching uploaded images or videos. However, tagging is a time-consuming task; therefore, tag recommendation is used to suggest relevant tags to the users based on user profiling, contents of resources or collective knowledge. Recommending tags of images to users does not only depend on user preference but also strongly relies on the contents of images.

In this paper, we propose a method for image tag recommendation using both visual features of images and past tagging behavior of users. We combine convolutional neural networks (CNN), which are widely used and have achieved high performance in image classification and recognition, and factorization machines (FM), since factorization models are the state-of-art approach for tag recommendation. We examine two types of combination between CNN and FM: a sequential and a parallel process. In the former, CNN is used to extract features of an input image and FM is applied to calculate the ranking scores for tags. The recommendations in the latter are aggregated from candidate tags suggested by both CNN and FM. Preliminary experiments on publicly available image tag recommendation dataset demonstrate the effectiveness of our approach.

References

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012): ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25, 1097–1105, 2012.*

RENDLE, S. (2010): Factorization machines. In *Proceeding of the 10th IEEE International Conference on Data Mining. IEEE Computer Society, 2010.*

Keywords

CONTENT-AWARE TAG RECOMMENDATION, CONVOLUTIONAL NETS, FACTORIZATION MACHINE

Robust principal graphs for data approximation

A.N. Gorban[1], E.M. Mirkes[1], and A. Zinovyev[2]

[1] University of Leicester, Leicester, LE1 7RH, UK {ag153,em322}@le.ac.uk
[2] Institut Curie, 26 rue d'Ulm, F75248, Paris, France andrei.zinovyev@curie.fr

Abstract. Revealing hidden geometry and topology in data sets, where noise is present, is a very challenging task. To approach this problem, we define elastic graphs and develop algorithms of pluriharmonic embedding of the elastic graphs into the data spaces [1]. These pluriharmonic embeddings serve as ideal approximators, the energy functionals penalize the deviation from this ideal forms, and the truncated functionals estimate the robust approximation error. Instead of the quadratic energy functional we employ the trimmed energy functions [2] to provide robustness of the method. In the splitting algorithms of optimization they also produce systems of linear equations and make the construction of the approximators much more stable to noise and outliers. For the construction of objects, we use topological grammars [1]. An adaptation of graph rewriting technology makes the approach at the same time universal and efficient for the construction of geometric and topological objects of bounded complexity [3]. For the construction of the grammars of geometrical and topological objects we introduced some specific additional labels on the graph elements.

References

GORBAN, A.N., ZINOVYEV, A. (2010): Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems, 20 (3), 219–232*.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., MATRÁN, C., MAYO-ISCAR, A. (2008): A general trimming approach to robust cluster analysis. *Ann. Statist. 36 (3), 1324–1345*.

ZINOVYEV, A., MIRKES, E. (2013): Data complexity measured by principal graphs. *Computers & Mathematics with Applications, 65 (10), 1471–1482*.

Keywords

TOPOLOGICAL GRAMMAR, TRIMMED FUNCTIONALS

Data Analysis III

Thursday, September 3, 2015: 11:00am - 12:15pm Room: EBS 2.1

Longitudinal models for categorical data, are they useful for dynamic market segmentation?

Francesca Bassi[1]

[1] University of Padova, Italy francesca.bassi@unipd.it

Abstract. Dynamic market segmentation is a very important topic in many businesses where it is interesting to gain knowledge on the reference market and on its evolution over time. Various papers in the reference literature are devoted to the topic and different statistical models are proposed.

In this paper two statistical approaches to model categorical longitudinal data to perform dynamic market segmentation are compared. The latent class Markov model identifies a latent variable which classes represent market segments at an initial point in time, customers can switch to one segment to another between consecutive measurement occasions and a regression structure models the effects of covariates, describing customers' characteristics, on segments belonging and transition probabilities. The latent growth mixture approach models individual trajectories, describing a behaviour over time. Customers' characteristics may be inserted in the model to affect trajectories and trajectories may vary across latent groups, in our case, market segments.

An Attempt at Comparative Analysis of Higher Education Systems in European Union Member States

Marta Targaszewska

Wrocław University of Economics marta.targaszewska@ue.wroc.pl

Abstract. Europe 2020 is strategy proposed by the European Commission for the advancement of the economy of the European Union and Members States. The main priorities of the strategy are smart, sustainable and inclusive growth. According to the strategy, crucial role in achieving these goals, besides: employment, research and development, climate change and energy, fight against poverty and social exclusion plays tertiary education, which should be marked by high quality and excellence. It caused a need to measure quality and to monitor system of higher education. Due to this fact, there are dedicated programmes and projects (for example: Eurydice, European Tertiary Education Register – ETER, U-Map, POL-on, AHELO), which are helpful in controlling the work of European Union's universities.

This paper presents results of research on comparative analysis (such as principal component analysis, cluster analysis) of higher education systems in European Union member states. Empirical studies have been conducted using data from an ETER - database of higher education institutions in Europe.

References

EUROPEAN COMMISSION (2010): *Communication from THE COMMISSION, EUROPE 2020, A strategy for smart, sustainable and inclusive growth*. Brussels.

RENCHER, A.C. and CHRISTENSEN, W.F. (2012): *Methods of Multivariate Analysis*. Wiley&Sons, Hoboken, New Jersey.

Keywords

MULTIVARIATE ANALYSIS, HIGHER EDUCATION, PRINCIPAL COMPONENT ANALYSIS, CLUSTER ANALYSIS

Identification of spatial patterns of agricultural practices in
arid Tunisian: Statistical mapping of socio-economic survey data

Mohamed Jaouad[1]

Laboratory of Economics and rural societies Arid Regions Institute – Medenine, Tunisia
Mohamed.Jaouad@ira.rnrt.tn mjaouad63@gmail.com

Abstract. Spatial analysis is particularly relevant in the case of agricultural practices and/or when practices influence processes with a spatial dimension. In arid regions, agricultural practices are strongly influenced and related to the nature of the environment or space. Indeed, the important thing is not the identification or knowledge of the agricultural landscape or spatial structures but, when studying farmers' practices of farms in a region, we pay particular attention to the links between practices and space. We locate the practices in the regions and we are interested in the adaptability of these practices, both for the environment and sustainability of agriculture.

The identification of spatial structure, inducing a number of variables, is a recurring problem in many fields of application. This paper presents some aspects related to the inclusion of space in the statistical analysis of socio-economic data based on field surveys conducted in the region of Menzel Habib (Southern Tunisia) under the ROSELT program / OSS-IRA in 2004. The data includes demography, agro- economy, and environment and it was subject to principal components analysis.

Keywords

PRINCIPAL COMPONENT ANALYSIS, AGRICULTURAL PRACTICES, SPATIAL, SOCIO-ECONOMIC DATA, ARID, TUNISIA.

Education

Thursday, September 3, 2015: 11:00am - 12:15pm Room: EBS 2.65

Techniques for Sampling Quasi-orders

Ali Ünlü[1] and Martin Schrepp[2]

[1] Technische Universität München, Munich, Germany ali.uenlue@tum.de
[2] SAP AG, Walldorf, Germany martin.schrepp@sap.com

Abstract. In educational theories (e.g., learning spaces), mastery dependencies between test items are represented as quasi-orders (reflexive and transitive relations) on the item set of a knowledge domain. Item dependencies can be used for efficient adaptive knowledge assessment and derived through exploratory data analysis, for example by algorithms of Item Tree Analysis (ITA). To compare ITA-type methods, typically large-scale simulation studies are employed, with samples of quasi-orders at their basis randomly generated and assumed to underlie the data. In this context, a serious problem is the fact that all of the algorithms are sensitive to the underlying quasi-order structure. Thus, it is crucial to base any simulation study that aims at comparing the algorithms in a reliable manner on representative samples, in the sense that each quasi-order is contained in a sample with the same probability. Suboptimal sampling strategies were considered in previous studies, thereby leading to biased conclusions. In this paper, we introduce sampling techniques that allow us to generate representative, or very close to representative, random quasi-orders.

References

FALMAGNE, J.-CL. and DOIGNON, J.-P. (2011): *Learning Spaces*. Springer, Berlin.

SARGIN, A. and ÜNLÜ, A. (2009): Inductive Item Tree Analysis: Corrections, Improvements, and Comparisons. *Mathematical Social Sciences, 58, 376–392*.

SCHREPP, M. (1999): On the Empirical Construction of Implications between Bi-valued Test Items. *Mathematical Social Sciences, 38, 361–375*.

ÜNLÜ, A. and SCHREPP, M. (2015): Biasing Effects of Non-representative Samples of Quasi-orders in the Assessment of Recovery Quality of IITA-type Item Hierarchy Mining. Manuscript in press in the Proceedings of ECDA 2014.

Keywords

ITEM TREE ANALYSIS, LEARNING SPACE, RANDOM QUASI-ORDER, REPRESENTATIVE SAMPLING

Predictive Modelling of Evidence Informed Teaching

Dell Zhang[1] and Chris Brown[2]

[1] DCSIS, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK
`dell.z@ieee.org`
[2] LCLL, UCL Institute of Education, 20 Bedford Way, London WC1H 0AL, UK
`chris.brown@ioe.ac.uk`

Abstract. Recently *evidence informed teaching* has been receiving more and more attention from policy makers etc. It has been shown that there are substantial benefits associated with school teachers using information/evidence from research to enhance their classroom practice.

In this paper, we analyse the questionnaire survey data collected from 79 English primary schools about the situation of evidence informed teaching. Specifically, we build a predictive model to see what external factors could help to close the gap between teachers' belief and behaviour in evidence informed teaching, which is the first of its kind to our knowledge.

The major challenge, from the data mining perspective, is that the five-level *Likert scale* responses ("strongly disagree", "disagree", "neither disagree nor agree", "agree", and "strongly agree") are neither categorical nor interval, but actually ordinal, which requires special consideration when we apply statistical analysis or machine learning algorithms. Adapting *Gradient Boosted Trees (GBT)*, we achieve a decent prediction accuracy (MAE=0.36) and gain new insights into possible interventions for promoting evidence informed teaching.

References

BROWN, C. (2014): *Evidence-Informed Policy and Practice in Education: A Sociological Grounding*. Bloomsbury Academic, London.

GOLDACRE, B. (2013): Building evidence into education. Technical report, Department for Education, UK.

LIKERT, R. (1932): A technique for the measurement of attitudes. *Archives of Psychology, 5, 228–238, 1932*.

Keywords

EVIDENCE INFORMED TEACHING, ORDINAL DATA ANALYSIS, PREDICTIVE MODELLING

Application of unfolding techniques to information retrieval sequences in teachers' judgment formation processes

Thomas Hörstermann[1], Sabine Krolak-Schwerdt[2], and Matthias Böhmer[3]

[1] University of Luxembourg, ECCS, thomas.hoerstermann@uni.lu
[2] University of Luxembourg, ECCS, sabine.krolak@uni.lu
[3] University of Luxembourg, ECCS, matthias.boehmer@uni.lu

Abstract. Assessment of student performance is inherent to educational systems, and teachers play a key role in most educational assessments. In this study, we investigated the information retrieval process in teachers' judgment formation, namely the kind and sequence of information considered prior to a secondary school track recommendation. German teachers got information about a fictitious student's performance (e.g. math grade), behavior (e.g. autonomy) and social background (e.g. language spoken at home) in the form of labeled information fields on a computer screen. The fields could be uncovered via mouse click and were re-covered once another field was uncovered. A two-dimensional unfolding model was fit to the data of information retrieval sequences. Results show three clusters of information fields in the model, in line with the information categories of student performance, behavior, and social background. The degree of inconsistency of student information influences the length of the information retrieval, but doesn't lead to a shift in teachers' preference orders. The study shows that unfolding techniques might be a promising approach in modeling educational judgment formation processes.

References

BÖHMER, I., HÖRSTERMANN, T., GRÄSEL, C., KROLAK-SCHWERDT, S. and GLOCK, S. (in press): Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung: Welcher Urteilsregel folgen Lehrkräfte? *Journal of Educational Research Online*.

CARROLL, J. D. (1980): Models and methods for multidimensional analysis of preferential choice data (or other dominance data). In: E. D. Lautermann and H. Feger (Eds.): *Similarity and Choice.* Hans Huber, Wien, 234-289.

Keywords

TEACHER JUDGMENTS, JUDGMENT FORMATION, UNFOLDING

Machine Learning and Knowledge Discovery III

Thursday, September 3, 2015: 2.35pm - 3.50pm Room: EBS 2.34

Feature selection algorithms for binarized data

Ludwig Lausser[1], Lyn Rouven Schirra[2,3], and Hans A. Kestler[1,2]

[1] Leibniz Institute for Age Research, Fritz-Lipmann Institute, Jena, Germany {llausser, hkestler}@fli-leibniz.de
[2] Institue of Neural Information Processing, Ulm University, Ulm, Germany
[3] Institute of Number Theory and Probability Theory, Ulm University, Ulm, Germany lyn-rouven.schirra@uni-ulm.de

Abstract. A major challenge in analyzing high-dimensional bio-molecular data is the construction of low-dimensional and interpretable classification models. A common technique is to incorporate a feature selection process into the training of the classification model. This generates models that will hopefully exclude a large amount of noisy and irrelevant signals and will only rely on a small subset of informative features.

A further attempt to increase the interpretability of a decision rule is to replace a quantitative model by a qualitative one. A common preprocessing step for these models is binarization, which is the task of discretizing a real-valued signal into a binary one. The discretized features can then be interpreted as indicators of high/low concentration levels.

In this study we examine the interaction of purely data-driven feature selection methods and supervised as well as unsupervised binarization algorithms. The benefit of these feature selection/binarization interactions will be tested in different classification scenarios. As quality criteria for these preprocessing cascades, the classifiers' performance and the features' selection stabilities will be analyzed.

Keywords

BINARIZATION, FEATURE SELECTION, CLASSIFICATION, BIOINFORMATICS

A Comparison of Ensemble Methods for Motor Imagery Brain-Computer Interfaces

Davide Valeriani[1], Ana Matran-Fernandez[1], Diego Perez-Liebana[1], Javier Asensio-Cubero[2], Christian O'Connell[3], and Andrei Iacob[1]

[1] School of Computer Science and Electronic Engineering, University of Essex, UK
{dvaler, amatra, dperez, abiaco}@essex.ac.uk
[2] Not applicable capitan.cambio@gmail.com
[3] The Computer Laboratory, University of Cambridge, UK co362@cam.ac.uk

Abstract. A Brain-Computer Interface (BCI) provides an alternative means of communication for people who are locked-in. For a BCI to work, the user will perform a specific mental task whilst wearing an Electroencephalography (EEG) cap that contains several electrodes. In particular, in a Motor Imagery (MI) BCI, users imagine themselves performing specific movements, e.g., rotating the right hand or moving his/her feet. The signals recorded by these electrodes are then preprocessed and fed to a classifier that will decide which of the possible actions is being performed. The output of the classifier is then sent to a device (e.g., a computer or wheelchair) for its execution.

In this paper, we will compare the performance of different systems (several ensembles using various voting algorithms and multiclass classifiers) on a 4-class MI task (left/right hand and feet movement imagery, plus an "idle" state). These methods will be ranked using a combination of different evaluation metrics. The best system will be applied to a real-time BCI used in an international competition.

References

LEE, F., et al. (2005): A comparative analysis of multi-class EEG classification for brain computer interface. *Proceedings of the 10th Computer Vision Winter Workshop, 195–204, 2005.*

LOTTE, F., et al. (2007): A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering, 4, R1–R13, 2007.*

Keywords

BRAIN-COMPUTER INTERFACE, ENSEMBLE, MULTICLASS, MOTOR IMAGERY

Dynamic Process Modeling: Combining Control Charts to detect Concept drift in Nonstationary Environment

Dhouha Mejri[1], Mohamed Limam[2], and Claus Weihs[3]

[1] Technische Universität Dortmund and University of Tunis, dhouha.mejri@tu-dortmund.de
[2] Dhofar University, Oman and University of Tunis, mohamedmtlimam@gmail.com
[3] Technische Universität Dortmund, claus.weihs@t-online.de

Abstract. Dynamics are fundamental properties of batch learning processes. Recently, dynamic process monitoring has interested many researchers due to the importance of dealing with time-changing data stream process in real world applications. A Dynamic Ensemble Control Model is presented in this paper to control such processes. It combines individual charts based on ensemble methods with a batch learning process to both monitor small and large shift simultaneously. It consists of three steps: first, transforming the task of determining the state of the process into a classification problem by treating control charts as attributes of the data where the drift has to be predicted. Second, Dynamic Weighted Majority- Winnow (DWM-WIN), as an ensemble method, is applied to combine different control charts. Third, misclassification error rates of DWM-WIN are monitored based on the time adjusting control chart for concept drift detection. The proposed control chart does not only exhibit superior robustness to individual charts but also presents a new heuristic for shift learning and batch monitoring in nonstationary environment.

References

MEJRI, D., KHANCHEL, R., and LIMAM, M., (2013): Ensemble method for concept drift in nonstationary environment, *Journal of Statistical computation and Simulation, 83, 1115-1128.*

MEJRI, D., LIMAM, M., and WEIHS, C., (2015): Monitoring a dynamic weighted majority method based on real datasets with concept drift, *Europeen Conference on Data Analysis*, Springer, Bremen.

Keywords

CONCEPT DRIFT, ADAPTIVE CONTROL CHARTS, DYNAMIC WEIGHTED MAJORITY ALGORITHM, ENSEMBLE METHODS

Data Analysis IV

Thursday, September 3, 2015: 2.35pm - 3.50pm Room: EBS 2.1

Distance Analysis of Football Player Performance Data

Serhat Emre Akhanli[1] and Christian Hennig[2]

[1] Department of Statistical Science, University College London Gower St, London WC1E 6BT, United Kingdom serhat.akhanli.14@ucl.ac.uk
[2] Department of Statistical Science, University College London Gower St, London WC1E 6BT, United Kingdom c.hennig@ucl.ac.uk

Abstract. We present a new idea to map football player's information by using multidimensional scaling, and to cluster football players. The actual goal is to define a proper distance measure between players. We believe that this type of information can be very useful for football scouts when assessing players; also journalists and football fans will be interested in this information. The data was assembled from whoscored.com. Variables are of mixed type, containing nominal, ordinal, count and continuous information. In the data pre-processing stage, four different steps are followed through for continuous and count variables: 1) representation (i.e., considerations regarding how the relevant information is most appropriately represented, e.g., relative to minutes played), 2) transformation (football knowledge as well as the skewness of the distribution of some count variables indicates that transformation should be used to decrease the effective distance between higher values compared to the distances between lower values), 3) standardisation (in order to make within-variable variations comparable), and 4) variable weighting including variable selection.

In a final phase, all the different types of distance measures are combined by using the principle of the Gower dissimilarity. We show outcomes of multidimensional scaling and potentially clustering.

References

GOWER, J. C. and LEGENDRE, P. (1986), Metric and Euclidean properties of dissimilarity coefficients *Journal of classification, 3.1: 5-48.*

HENNIG, C. and HAUSDORF, B. (2006), Design of Dissimilarity Measures: A New Dissimilarity Between Species Distribution Areas.*Data Science and Classification, 29-38.*

Keywords

FOOTBALL, MIXED TYPE DATA, DATA PRE-PROCESSING, DISTANCE MEASURE, MULTIDIMENSIONAL SCALING

Reduction of the Dimensionality of the Data in Hedonic Modelling Using Principal Component Analysis and Partial Least Squares

Anna Król

Wrocław University of Economics anna.krol@ue.wroc.pl

Abstract. The theory of hedonic models states that it is possible to precisely describe the price of heterogeneous commodity by a set of its characteristics. However, the quality of the model depends on the completeness of the set of significant attributes of commodity used for estimation. In cases where the number of explanatory variables (attributes) is similar (or greater than) the number of observations in the dataset, or when the explanatory variables are strongly correlated with each other, it is not possible to use standard methods of estimation (OLS).

The aim of this article is to examine the usefulness of principal component analysis and partial least squares method in reducing the number of dimensions of the data in case of problems with collinearity. It is assumed that the use of these methods can yield better results than alternative solution - the removal of problematic variables from the data set. Empirical research was carried out for selected groups of commodities which are commonly used in hedonic modelling (durable goods, real estate). The databases were created using tool for collecting data from web pages developed by the author.

References

DIEWERT, W.E. (2003): Hedonic Regressions. A Consumer Theory Approach. In: R.C. Feenstra, M.D. Shapiro (Eds.): *Scanner Data and Price Indexes*. University of Chicago Press, 317–348.

HASTE, T., TIBISHIRANI, R. and FRIEDMAN, J. (2009): *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, New York.

HELLAND, I.S. (1990): Partial Least Squares Regression and Statistical Models. *Scandinavian Journal of Statistics, 17(2), 97–114.*

Keywords

HEDONIC METHODS, PRINCIPAL COMPONENT ANALYSIS, PARTIAL LEAST SQUARES

Robust Bayesian Linear Regression Model for Preprocessing Large-Scale Genomic Data

Yoshiko Hayashi, Baba Bukar Alhaji, Hongsheng Dai, Berthold Lausen

Department of Mathematical Science, University of Essex, UK yhayasa@essex.ac.uk bbalha@essex.ac.uk hdaia@essex.ac.uk blausen@essex.ac.uk

Abstract. We have developed a robust linear regression model using heavy-tailed distribution for large scale multiple hypothesis testing. We start with an investigation of the limitation of the robust linear regression model. The use of a heavy-tailed distribution is common for the error term of a robust linear regression model. However, the model is not robust against the outlier, when the independent variable is approaching infinity. We explain this using heavy-tailed modelling, and provide the minimum controllable range, where the robust model works well. We also propose total and individual tests for outlier detection using the posterior predictive distribution.

Secondly, we apply the robust regression model to a large scale multiple testing hypothesis and diagnose the model using the predictive posterior distribution. Finally we suggest how we should apply Bayesian False discovery rate on the model. We illustrate the method and our results using gene expression data of colorectal cancer.

References

ANDRADE, J.A.A. and O'HAGAN, A. (2011), "Bayesian robustness modell of location and scale parameters," *Scandinavian Journal of Statistics, 38, 691-711, 2011*.

LEWIN, A., RICHARDSON, S., MARSHALL, C., GLAZIER, A., and AITMAN,T. (2006) "Bayesian modeling of differential gene expression,"*Biometrics, 62, 1-9, 2006*.

MARSHALL, E. C. and SPIEGELHALTER, D. J. (2003), "Approximate cross-validatory predictive checks in disease mapping models," *Statistics in Medicine, 22, 1649-1660, 2003*.

Keywords

ROBUST MODELLING, BAYESIAN ANALYSIS, FALSE DISCOVERY RATE

Marketing I

Thursday, September 3, 2015: 2.35pm - 3.50pm Room: EBS 2.65

Nominal Response - Based Conjoint Analysis in Co-Creation of Product Value

Adam Sagan[1] Aneta Rybicka[2] and Justyna Brzezińska-Grabowska[3]

[1] Cracow University of Economics sagana@uek.krakow.pl
[2] Wroclaw University of Economics aneta.rybicka@ue.wroc.pl
[3] University of Economics in Katowice justyna.brzezinska-grabowska@ue.katowice.pl

Abstract. Value co-creation process is, according to Vargo and Lush the core concept in marketing orientation within the framework of the Service Dominant Logic (Vargo and Lush 2013). It combines the analysis of product offer utilities and value ordering and preference of the customers.

The aim of the paper is adoption of Nominal Response IRT (NR-IRT) polytomous measurement model (Bock 1972) within the framework of conjoint analysis of banking product on Polish market and to compare the findings with the choice based model (conditional logit model with the specific alternatives) Proposed approach links the tradition of conjoint measurement of dominance structures in attitude measurement with the preference analysis.

Incorporating the explicit measurement model within the conjoint analysis allows for error - in variables effect and controlling for downward bias in estimates of the parameters. Additionally, it combines the fundamental conjoint measurement of Rasch tradition with the preference - oriented conjoint analysis (Neubauer 2001).

References

BOCK, R.D. (1972): Estimating Item Parameters and Latent Ability when Responses are Scored in Two or More Nominal Categories. *Psychometrika, 37, 29–1*.

NEUBAUER, G. (2003). An IRT-Approach for Conjoint Analysis, In: A. Ferligoj and A. Mrvar (Eds.): *Developments in Applied Statistics*, Metodoloski zvezki, 19, Ljubljana, 35-47.

VARGO, L., LUSH, R. (2013). Service Co-Creation in SDL Approach, *Journal of Marketing. 12(2), 33-33*.

Keywords

CONJOINT MEASUREMENT, NOMINAL RESPONSE MODEL, VALUE CO-CREATION

Accounting for the IIA Property in Market Simulations: A Comparison of Choice Rules Using Simulated Choice-Based Conjoint Data

Maren Hein[1], Peter Kurz[2] and Winfried J. Steiner[1]

[1] Department of Marketing, Clausthal University of Technology,
38678 Clausthal-Zellerfeld maren.hein@tu-clausthal.de,
winfried.steiner@tu-clausthal.de

[2] TNS Deutschland GmbH, Landsberger Str. 284,
80687 München peter.kurz@tns-global.com

Abstract. Market simulations enable managers to explore how respondents react to new product entries or product design modifications (”what-if” scenarios) and to predict shares of preference under hypothetical market scenarios. In order to predict which products respondents would choose, choice rules are used that translate respondents’ part-worths into expected individual choice probabilities. However, the selection of the choice rule should be well-considered, because different choice rules do not necessarily yield the same share predictions. Using the CBC-HB model to estimate individual part-worth utilities we systematically compare market share predictions based on the first choice rule, logit choice rule, randomized first choice rule and in particular HB random draws. We propose a simulation study in order to examine the conditions under which one of these choice rules recovers market shares better than the other. Further, we investigate the ability of the different choice rules to handle nearly similar alternatives included in holdout choice scenarios to assess how well the choice rules account for the IIA property (cf. Orme and Huber 2000; Orme and Baker 2000). The performance of the four choice rules is evaluated under experimentally varying conditions using statistical criteria for predictive accuracy.

References

ORME, B. and HUBER, J. (2000): Improving the value of conjoint simulations. *Marketing Research, 12 (4), 12–20, 2000.*

ORME, B. and BAKER, G. (2000), Comparing hierarchical Bayes draws and randomized first choice for conjoint simulations. In: Sawtooth Software Research Paper Series, Sequim, WA.

Keywords

CHOICE RULES, IIA, SIMULATION

Using Cluster Analysis for the Identification of Heterogeneous Brand Images

Daniel Böger[1], Pascal Kottemann[2], Reinhold Decker[3], and Martin Meißner[4]

[1] Bielefeld University, Department of Business Administration and Economics, P.O. Box 10 01 31, 33501 Bielefeld, Germany dboeger@wiwi.uni-bielefeld.de

[2] Bielefeld University, Department of Business Administration and Economics, P.O. Box 10 01 31, 33501 Bielefeld, Germany pkottemann@wiwi.uni-bielefeld.de

[3] Bielefeld University, Department of Business Administration and Economics, P.O. Box 10 01 31, 33501 Bielefeld, Germany rdecker@uni-bielefeld.de

[4] University of Southern Denmark, Department of Environmental and Business Economics, Niels Bohrs Vej 9, 6700 Esbjerg, Denmark meissner@sam.sdu.dk

Abstract. In recent years, the Brand Concept Maps (BCM) approach has received considerable attention as a promising tool for measuring consumers' brand images. A key advantage of the BCM approach is that it provides a clear set of rules marketing researchers can apply to aggregate individual brand concept map information into a so-called consensus map. The latter describes the sample's most relevant brand associations and their interconnections in the consumers' minds. Assuming that many brands are perceived differently by consumer segments, it is surprising that the original BCM approach does not provide specific rules how to cope with heterogeneous brand image perceptions.

Against this background, the aim of this paper is to develop a set of improved aggregation rules which enable marketing researchers to identify consumer segments with heterogeneous brand perceptions. In doing so, we first discuss similarity measures for individual brand association networks and compare several cluster solutions. Second, we examine the resulting images of well-known brands. Our findings indicate that brand images can be very heterogeneous and marketing researchers should be careful with respect to which conclusions they draw when applying the original BCM approach. Additionally, we provide recommendations on how to best identify brand image segments from this kind of data.

Keywords

BRAND CONCEPT MAPS, CLUSTER ANALYSIS, HETEROGENEOUS BRAND IMAGES

Machine Learning and Knowledge Discovery IV

Thursday, September 3, 2015: 4.15pm - 5.30pm Room: EBS 2.34

Using Multi-Word Concepts for Classification and Exploration of Medical Texts

Van Hyfte, Dirk[1], De Boe, Benjamin[1] and Bouzinier, Michael[2]

[1] InterSystems Benelux, Medialaan 32/1 B-1800 Vilvoorde Belgium
Dirk.VanHyfte@intersystems.com
[2] InterSystems Corp. One Memorial Drive Cambridge, MA 02142, USA
Misha.Bouzinier@intersystems.com

Abstract. In the modern world 80% of healthcare information still resides in free text notes and reports. In most cases they are analyzed using various bag-of-words models. This requires often expensive mapping to ontologies and that the same word might have different meaning in different contexts. An alternative approach used by InterSystems iKnow identifies meaningful word groups in a domain agnostic way using just linguistics language model. However this leads to a problem of low frequency for many concepts insufficient for classification or clustering purposes.

We will discuss selecting the most appropriate word groups for classification using information theory approach, using dominance for a word group calculated as a TF/IDF type of measure. To analyze application of this technique we calculate domain coverage profile: percentage of the documents in the corpus as a function of the number of individual concepts used for coverage and the minimum number of selected concepts in an individual document. We compare domain coverage profiles for EPRs (Electronic Patient Records), PubMed abstracts and Social Media.

References

Turney, P.D. and Pantel P. (2010) From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Articial Intelligence Research*, 37 141-188, 2010.

Friedman, C., Pauline K., and Rzhetsky A. (2002) Two Biomedical Sublanguages. *Journal of Biomedical Informatics* 35.4: 222-35, 2012

Bronselaer, A, Debergh, S, Van Hyfte, D, De Tre, G (2010) Text clustering based on concept-relational decomposition *ICL Proceedings* 357-359, 2010

Baayen, H. *Word Frequency Distributions* Springer Science & Business Media, 2001

De Boe, B., Bouzinier, M, Van Hyfte D. (2013) Extending the PMML Text Model for Text Categorization *SIGKDD Conference (KDD-2013)* 2013

Keywords

LINGUISTICS, HEALTHCARE, MEDICINE, NLP, FREQUENCY

textMF: Text-based Matrix Factorization for Geolocation Prediction in Twitter Stream

Nghia Duong-Trung, Nicolas Schilling, Lucas Rego Drumond and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab
University of Hildesheim, Germany
{duongn,schilling,ldrumond,schmidt-thieme}@ismll.uni-hildesheim.de

Abstract. Micro-blogging services, such as Twitter, provide an indispensable channel to communicate, access, and exchange current affairs, as they have reached a wide range of individuals, organizations, companies along with others. Understanding the user's geographical location can enable us to provide information and recommend businesses in a location-aware recommender system. The geographical location prediction problem we address is to predict the user's geolocation in general and at a particular posting time.

In this paper, we propose `textMF` - Text-based Matrix Factorization which associates latent features to users and tweets by factorizing the content matrix. The learned features are used in a linear regression function for the task of geolocation prediction. Additionally, we extend our model to deal with the task of predicting user's geolocation at a particular posting time. Preliminary experiments on the publicly available real world data sets demonstrate the efficacy of our approach on different prediction granularity levels.

References

HAN, B., COOK, P. and BALDWIN, T. (2014): Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research, 451-500, 2014.*

EISENSTEIN, J. et al. (2010): A latent variable model for geographic lexical variation. *Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 1277-1287, 2010.*

Keywords

MATRIX FACTORIZATION, TWITTER, GEOLOCATION, TEXT REGRESSION

Exploring the Relationship of Citation and Article Level Metrics with Linguistic Patterns and Quantities

Arlene Casey[1], Samad Ahmadi[2], and Fionn Murtagh[3]

[1] De Montfort University, U.K. arlene.casey@myemail.dmu.ac.uk
[2] De Montfort University, U.K. sahmadi@dmu.ac.uk
[3] University of Derby, and Goldsmiths University of London, U.K. fmurtagh@acm.org

Abstract. A journal impact factor or number of citations an article has received is generally accepted as an indication of the level of scholarly output of an article. The growth of online access to scholarly articles and the availability of social media platforms, which publicly discuss articles, is giving rise to more information being available in terms of article level metrics, increasingly known as altmetrics. Many academic publishers are providing such altmetrics on their websites such as Mendeley readership statistics or an overall altmetric score. It is difficult to judge what level of importance to put on these altmetrics as it is possible an article of high scholarly output may not have any valid altmetrics. In this research we study available article metrics on scholarly papers and explore their relationship to each other and with linguistic patterns. The linguistic patterns are items such as citation occurrences in sentences, sentence ratios per sections. Our research helps to provide an understanding of article level metrics and their relationship to quantifiable aspects of content, in addition it assists in validating measures for altmetrics.

Keywords

CITATION ANALYSIS, ALTMETRICS,JOURNAL RANKING

Data Analysis V

Thursday, September 3, 2015: 4.15pm - 5.30pm Room: EBS 2.1

Two Dimensional Smoothing via an Optimised Whittaker Smoother.

Sri Utami Zuliana[1] and Aris Perperoglou[2]

[1] University of Essex, Colchester, UK, `sutami@essex.ac.uk`
[2] University of Essex, Colchester, UK, `aperpe@essex.ac.uk`

Abstract. The histogram is one of the most powerful tools of data visualization. A problem might rise however, when trying to plot many points onto one simple graph. As the number of observations becomes larger and larger many scatter-plots end up being to busy for the eye to understand. Often, in moderate to large datasets, a collection of many observations on one plane will end up revealing a cloud of points where all structure remains obscured by the superposition of one point onto another. Depending on what is the medium where such a graph will be illustrated, it becomes a waste of ink or space.

To address this problem Eilers and Goeman illustrated a way of smoothing scatter-plots in two directions using penalized b-splines or p-splines. This approach has been implemented in package `gamlss.util` via command `scattersmooth`. In this work we are focusing on the paper by Eilers and Goeman where a scatter-plot is enhanced using smoothed densities. We will start off with the same approach, where penalized splines are applied on the *x* and *y* directions, respectively. However, we will go a step further and show how the optimal smoothed scatter-plot can be obtained by estimating the amount of penalty needed for each graph. We view penalized splines as random effects that their variance depends on the penalty weight. We will revise the algorithm and extend it to apply to two dimensional smoothing.

References

Eilers PH, Goeman JJ. (2004): Enhancing scatterplots with smoothed densities.*Bioinformatics, 20(5), 623–628, 2004.*

Keywords

PENALIZED SPLINES, H-LIKELIHOOD, EFFICIENT COMPUTATIONS

Robust approach to life expectancy projection

Grażyna Trzpiot[1] and Justyna Majewska[1]

University of Economics in Katowice, Poland

Abstract. Future trends in life expectancy are highly relevant for public policy, fiscal and health care system planning. To project life expectancy into the future, the Lee-Carter model is often referred as a standard. However, this model is sensitive to the choice of the historical period if the reduction in mortality does not follow a linear trend. We will consider robust models. Robust versions of well-known selection criteria will be discussed, but also criteria based on robust bootstrap. We will also discuss robust model building/selection methods for high dimensional data.

On a Comprehensive Metadata Framework for Artificial Cluster Data Generation

Rainer Dangl and Friedrich Leisch

Institute for Applied Statistics and Computing
University of Natural Resources and Life Sciences, Vienna
Peter-Jordan-Strasse 82, 1190 Vienna, Austria

rainer.dangl@boku.ac.at, friedrich.leisch@boku.ac.at

Abstract. Solid results in unsupervised model validation require thoroghly tested validation methods and algorithms. We intend to optimize their development by proposing a framework that streamlines the way artificial test data is constructed. This improves comparability between existing and new algorithms and offers a more transparent way of assessing performance.

In essence, the framework, developed with R, consists of metadata object definitons for various kinds of data types (e.g. metric, functional or ordinal data). These objects impose a certain structure on the metadata information that enables R to assemble the actual data sets in a way that the user can generate all desired data sets (custom functions for random number generation, location of group means, etc.) while at the same time providing a reliable structure for the metadata that is the same for all data sets.

The necessary functions for (meta)data generation have been implemented in R package `bdlp`, which is under development and thus at the moment hosted on R-forge.

References

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The Design and Analysis of Benchmark Experiments. *Journal of Computational and Graphical Statistics, 14*(3), 675-699.

Keywords

BENCHMARKING, MODEL VALIDATION, CLUSTERING, UNSUPERVISED LEARNING, ARTIFICIAL DATA

Marketing II

Thursday, September 3, 2015: 4.15pm - 5.30pm Room: EBS 2.65

Measuring the Acceptance of New Technologies in Marketing: Surveys vs. Online Reviews

Daniel Baier[1], Alexandra Rese[2], and Stefanie Schreiber[2]

[1] University of Bayreuth, Chair of Innovation and Dialogue Marketing, Universitätsstraße 30, 95447 Bayreuth, Germany, daniel.baier@uni-bayreuth.de

[2] BTU Cottbus-Senftenberg, Chair of Marketing and Innovation Management, Erich-Weinert-Straße 1, 03046 Cottbus, Germany, rese|stefanie.schreiber@tu-cottbus.de

Abstract. Measuring the acceptance of new technologies has a long tradition in marketing. The main goal is to understand how consumers come to accept and use a new technology (e.g., recommendation systems in online shopping, Baier, Stüber 2010). For this purpose, several models have been proposed and tested (e.g. basing on the technology acceptance model by Davis). But, over the years, besides extensions of the underlying modeling assumptions, few progress has been made with respect to the data collection and analysis approach: Still, surveys and structural equation modeling is the dominant measurement approach.

The question arises whether newer measurement approaches could be applied for this purpose. First attempts (e.g. Rese et al. 2014) basing on online reviews will be extended in this paper: Experiments are conducted where respondents participate in a traditional survey and – additionally – are encouraged to score the new technology on various scales (e.g. star ratings, bipolar semantic scales, nonformatted comments). The results are analyzed and compared across the measurement approaches as well as results from online reviews from user fora. The results are promising: It seems that the data collection via surveys can be replaced – with some reservations – by the analysis of online reviews.

References

BAIER, D., STÜBER, E. (2010): Acceptance of Recommendations to Buy in Online Retailing. *Journal of Retailing and Consumer Services, 17 (3), 3–180.*

RESE, A., SCHREIBER, S., BAIER, D. (2014): Technology Acceptance Modeling of Augmented Reality at the POS: Can Surveys be Replaced by an Analysis of Online Reviews? *Journal of Retailing and Consumer Services, 21 (5), 869–876.*

Keywords

ONLINE REVIEW, SURVEY, TECHNOLOGY ACCEPTANCE MODEL

The Role of Cultural Dimensions in the Acceptance of Augmented Reality in Retailing

Alexandra Rese[1], Eleonora Pantano[2], and Daniel Baier[3]

[1] BTU Cottbus-Senftenberg, Chair of Marketing and Innovation Management, Erich-Weinert-Straße 1, 03046 Cottbus, Germany, rese@tu-cottbus.de

[2] Middlesex University London, Department of Marketing, Branding, and Tourism, The Burroughs, London, NW4 4BT, UK, e.pantano@mdx.ac.uk

[3] University of Bayreuth, Chair of Innovation and Dialogue Marketing, Universitätsstraße 30, 95447 Bayreuth, Germany, daniel.baier@uni-bayreuth.de

Abstract. Information technology and in particular Augmented Reality (AR) is increasingly used in retailing across national boundaries. AR is an interactive technology to combine ”real and computer-generated digital information into the user's view of the physical real world in such a way that they appear as one environment” (Olsson et al. 2012, p. 29). For example the Italian Luxottica group trading Ray-Ban, offers an internationally available virtual mirror application to try-on sunglasses and eyeglasses. Since research has proposed and shown that national culture has an influence on behavioural models (see e.g. Srite and Karahanna 2006) this paper examines the acceptance of AR technology across two cultures, e.g. Italy and Germany, relying on the Technology Acceptance Model (TAM) (Davis 1986) and taking two different cultural measurement approaches into account. On the one hand, the Hofstede dimensions surveyed across participants in Germany and Italy were used. On the other hand, Hofstede's cultural distance measure is used.

References

DAVIS, F.D. (1986): A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. PhD thesis, Massachusetts Institute of Technology, Sloan School of Management.

OLSSON, T., KÄRKKÄINEN, T., LAGERSTAM, E., VENTÄ-OLKKONEN, L. (2012): User Evaluation of Mobile Augmented Reality Scenarios. *Journal of Ambient Intelligence and Smart Environments, 4 (1), 29–47.*

SRITE, M., KARAHANNA, E. (2006): The Role of Espoused National Cultural Values in Technology Acceptance. *MIS Quarterly, 30 (3), 679–704.*

Keywords

AUGMENTED REALITY, TECHNOLOGY ACCEPTANCE MODEL

Measuring Reciprocal Effects of New Product Information on Brand Image

Anja Hörmeyer, Pascal Kottemann and Reinhold Decker

Bielefeld University, Department of Business Administration and Economics, P.O. Box 10 01 31, 33501 Bielefeld, Germany (ahoermeyer@wiwi.uni-bielefeld.de, pkottemann@wiwi.uni-bielefeld.de, rdecker@uni-bielefeld.de)

Abstract. Today's marketplaces are increasingly characterized as highly competitive and companies' offerings are perceived as being interchangeable to a high degree. Accordingly, it is necessary for companies to continuously introduce new products into the market to achieve competitive advantages and to support strategic growth. In this regard, brands are important cues for consumers to evaluate the value of new products with respect to quality and risk perception as well as trust.

In this paper, we investigate reciprocal effects resulting from the announcement of new products that vary in terms of their degree of consistency with prior products of the same brand on the brand's image. We therefore build on the associative network memory model as well as on Keller's conceptualization of customer-based brand equity (Keller 1993) and apply the advanced Brand Concept Maps approach (Schnittka et al. 2012). An empirical study with a total of 300 respondents was conducted to investigate changes in the associative network structure of the well-known brand *Nike*. Our findings show the stability of the brand image, i.e., the brand is able to introduce new products without harmful effects, independent from the level of consistency. Thus, managers of strong brands do not need to fear temporal image deterioration resulting from the introduction of a new product and should be encouraged to extend their offerings in a wider range.

References

KELLER, K. L. (1993): Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing, 57(1), 1–22*.

SCHNITTKA, O., SATTLER, H. and ZENKER, S. (2012): Advanced brand concept maps: A new approach for evaluating the favorability of brand association networks. *International Journal of Research in Marketing, 29(3), 265–274*.

Keywords

BRAND CONCEPT MAPS, BRAND IMAGE, RECIPROCAL EFFECTS, BRAND ASSOCIATION NETWORKS

Clustering II

Friday, September 4, 2015: 9.00am - 10.15am Room: EBS 2.2

Generalization, Combination and Extension of Functional Clustering Algorithms

Christina Yassouridis and Friedrich Leisch

University of Resources and Life Sciences, Austria

Abstract. Clustering functional data is mostly based on the projection of the curves onto an adequate basis and building random effects models of the basis coefficients. The parameters can be fitted by an EM-algorithm. Alternatively, distance based models are used in the literature. Such as in the multidimensional case, a variety of derivations of these models has been published. Although their calculation procedure is similar, their implementations are very different including distinct hyper parameters and data formats as input. This makes it difficult for the user to apply and particularly to compare them. Furthermore they are mostly limited to specific basis functions. The presentation aims to show the common elements between existing models in highly cited articles, first on a theoretical basis. Later their implementation is analyzed and it is illustrated how common code chunks could be extracted and how the algorithms could be improved and extended to a more general level. A special consideration is given to those models including the possibility of sparse measurements. Finally, they were compared on simulated datasets. An `R`-package is in the process of being designed, including the modified algorithms and integrated into a unique framework.

Keywords

FUNCTIONAL MIXED EFFECTS, FUNCTIONAL CLUSTERING, GENERALIZATION, SPARSE MODELS

One dimensional Markov random field model for the analysis of ChIP-seq data with a non-parametric component

Baba B Alhaji[1], Hongsheng Dai[1], Andrew Harrison[1], and Berthold Lausen[1]

Department of Mathematical Sciences, University of Essex, Colchester, UK;
bbalha@essex.ac.uk

Abstract. In classification problems, many application areas require distinguishing a *signal* from a *noise* component. In ChIP-seq data for example, the structure of the *signal* does not follow a standard distribution, therefore the *signal* distribution is usually further modelled as a mixture of component distributions (Spyrou *et al*(2009)).

However, modelling the *signal* as a mixture of distributions is computationally challenging due to the difficulties in justifying the exact number of components to be used and due to the label switching problem (Stephens (2000)). We proposed one dimensional Bayesian hidden Markov random field mixture model with parametric and non-parametric distributions. The non-parametric distribution is used to model the *signal* component. The proposed model accounts for spatial dependencies in the data. We consider the case of discrete data and show how this new methodology leads to more accurate parameter estimation and smaller classification rate. We show an application of the method to data generated by ChIP-sequencing experiments for 200 base pairs window size.

References

SPYROU, C. and STARK, R. and LYNCH, A. G. and TAVARE, S. (2009): Bayespeak: Bayesian analysis of chip-seq data. *BMC Bioinformatics, 10, 299, 2009*.

STEPHENS, M. (2000): Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, 62, 795–809, 2000*.

Keywords

BAYESIAN, MIXTURE MODEL, LABEL SWITCHING, MARKOV MODEL, CHIP-SEQ

A Hierarchical Bayesian Model for joint Clustering of Clinical and Heterogenous Omics Data

Ashar Ahmad[1] and Holger Fröhlich[1]

University of Bonn, Institute for Computer Science, Algorithmic Bioinformatics, c/o
Bonn-Aachen International Center for IT, Dahlmannstr. 2, 53113 Bonn
ashar@bit.uni-bonn.de
frohlich@bit.uni-bonn.de

Abstract. Discovery of clinically relevant disease sub-types is of prime importance in personalized medicine. Disease subtype identification has traditionally been explored in an unsupervised machine learning paradigm which involves clustering of patients based on available omics data, such as gene expression. The clinical significance of obtained clusters can then be established in a post-hoc analyses step focusing, for example, on differences in the survival of patients falling in different clusters. The major limitation of such an approach is the failure to guarantee the clinical relevance of the obtained clusters, because patients clustering together based on omics data may not show significant differences with respect to their clinical outcomes.

In our work we propose a new algorithm which simultaneously clusters heterogenous omics and clinical data in order to find clinically relevant disease subtypes. For this purpose we formulate a novel hierarchical Bayesian graphical model which combines a Dirichlet Process Mixture Model (DPMM) with an Accelerated Failure Time (AFT) model. In this way we make sure that patients are grouped in the same cluster only when they show similar characteristics with respect to molecular features across data types (e.g. gene expression, DNA methylation) as well as survival times.

Our model contains an automatic feature selection (Bayesian LASSO) to find out the most relevant molecular features in a cluster and data type specific manner, thus effectively identifying biomarker signatures for each disease subtype. We develop a Gibbs sampling algorithm for Bayesian inference, which allows us to estimate posterior probabilities for all model parameteris, including number of clusters, cluster memberships of patients and cluster specific influences of specific molecular features.

We extensively test our model in simulation studies and apply it to cancer patient data from the TCGA repository. Notably, our method is able to find sub-groups, which differ from previously published ones.

Keywords

HIERARCHICAL BAYESIAN MODELLING, PERSONALIZED MEDICINE, HIGH DIMENSIONAL DATA, DATA INTEGRATION

Digital Humanities and Social Sciences I

Friday, September 4, 2015: 9.00am - 10.15am Room: EBS 2.34

Student Values Revisited
- Measuring Dominant Interests in Personality

Thomas Hummel[1], Victoria-Anne Schweigert[1], and Andreas Geyer-Schulz[1]

Institute of Information Systems and Marketing (IISM), Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe
{thomas.hummel, victoria-anne.schweigert,
andreas.geyer-schulz}@kit.edu

Abstract. In this contribution we replicate a survey of students' attitudes towards values derived from Allport et al. (English, 1960) and Roth (German, 1972). These authors used questions on preferences to identify the underlying attitudes towards values based on the personality typology of Spranger (1921). In addition, the survey has been augmented with Kahle's (1983) list of values (LOV) instrument based on a ranking scale. This opens the door for a comparison of different instruments for the investigation of value systems based on different cognitive theories and to explore the existence of associations between the different constructs. Regarding usability we study different types of interaction elements in the web-survey (numeric field entries, radio-buttons, and drag-and-drop rankings). These are randomly assigned to the test persons in order to study whether interaction style has a systematic bias on answers. We present and critically discuss the results in a historic context.

References

ALLPORT, G. W., VERNON, P. E. and LINDZEY, G. (1960): *Study of Values. A Scale for Measuring Dominant Interests in Personality* (Manual & Test Booklet), (3rd ed.). Houghton Mifflin Company, Boston.

KAHLE, L. R. (1983). *Social Values and Social Change: Adaptation to Life in America*. Praeger Scientific, New York.

ROTH, E. (1972). *Der Werteinstellungstest. Eine Skala zur Messung dominanter Interessen der Persönlichkeit* (Handanweisung & Testheft). Hans Huber, Bern.

SPRANGER, E. (1921). *Lebensformen: Geisteswissenschaftliche Psychologie und Ethik der Persönlichkeit*. Max Niemeyer Verlag, Halle (Saale).

Keywords

VALUE SYSTEMS, LIST OF VALUES, REPLICATION, INTERACTION STYLE BIAS

Movers and Stayers in a Religious Marketplace

Barry McDonald[1]

Institute of Natural and Mathematical Sciences, Massey University at Albany, Auckland, New Zealand b.mcdonald@massey.ac.nz

Abstract. New Zealand is a developed western country whose citizens are mostly of British, Irish or other European descent. Religious affiliation in New Zealand is remarkably fluid: a recent panel study showed about 15% of respondents changed their religious designation in just three years.

This talk will examine factors associated with people who start/stop identifying with particular religious affiliations or with no religion. Sources include panel data from a random sample of the electoral roll, and a large attitudinal survey of Protestant and Catholic church attenders. Multivariate analysis and graphical methods will be used to explore what differentiates the religious movers and stayers.

The talk may suggest useful methods and factors for exploring religious change in other western societies.

Keywords

RELIGIOUS CHANGE, MULTIVARIATE ANALYSIS, NEW ZEALAND

A social sustainability model: An application to Mexican Small-Scale Dairy-Farming Households

Mónica Elizama Ruiz-Torres[1] Ana Lorga da Silva[2], Carlos Manuel Arriaga-Jordán[1] and Francisco Ernesto Martínez-Castañeda[1]

[1] Instituto de Ciencias Agropecuarias y Rurales (ICAR). Universidad Autónoma del Estado de México. Instituto Literario 100, Col. Centro, 50000. Toluca, Edo. Méx., México
monica.ruiz.torres24@gmail.com, cmarriagaj@uaemex.mx, femartinezc@uaemex.mx

[2] Escola de Ciências Económicas e das Organizações, ULHT, Lisboa, Portugal
ana.lorga@ulusofona.pt

Abstract. Although sustainability became a research area since Brundtland Report (1987), actually there is no single definition of social sustainability. It has been described as a multidimensional concept; with a several themes that include the satisfaction of basic needs (KARAMI and MANSOORABADI, 2008), or tangible (drinking water, safe food, medication and home) and intangible needs (education, culture, equity and justice (VALLANCE *et all*, 2011). In this study, we combine ethnographic and statistical methods in order to analyze social sustainability through two variables: Economic Relations (ER) and Social Relations (SR), in Mexican small-scale Dairy-farming households. A Multivariate regression models estimated by the Ordinary Least Squares method, were estimated for both variables.

References

KARAMI, E. and MANSOORABADI, A.(2008): Sustainable agricultural attitudes and behaviours: a gender analysis of Iranian farmers. *Environment Development and Sustainability, 10:6, 883-898, 2008*

VALLANCE, S., PERKINS, H. C., and DIXON, J. E. (2011). What is social sustainability? A clarification of concepts. *Geoforum, 42, 342-348, 2011*

Keywords

LIVESTOCK, LIVELIHOODS, ETHNOGRAPHY, ORDINARY LEAST SQUARES

Geosciences and Archeology

Friday, September 4, 2015: 9.00am - 10.15am Room: EBS 2.1

Data Mining in Atmospheric Gravity Waves

Alfred Ultsch1[1,2] Christopher Rogos[2] and Christof Maul[3]

[1] Databionics Research Group, University of Marburg, Marburg, Germany
ultsch@informatik.uni-marburg.de

[2] Science Division, Academical Fyling (AKAfly), University of Frankfurt, Frankfurt, Germany

[3] Institute for Physical and Theoretical Chemisty, Technical University of Braunschweig, Braunschweig, Germany

Abstract. Gravity waves can emerge as a result of the perturbation of atmospheric ciculatory systems [1]. They encompass periodic, yet geographically stationary, changes in temperature, pressure and vertical wind components. Occurrence of such waves is frequent if strong winds hit high mountains [2]. Secondary effect of such waves may also be encountered as clear air turbulence (CAT) in commercial flights. Atmospheric gravity waves strongly influence weather phenomena and on a larger time scale climatic processes. They are responsible for the vertical mixing of the atmosphere from the mesosphere up to the stratosphere [2]. First results from research flights in the Pyrenees during the spring 2015 measuring campaign are reported. Several flights with a sensor equipped unpowered glider in altitudes between 2000 and 7000m were undertaken. Data Mining and Knowledge Discovery methods from the Databionics Lab in Marburg were applied [3]. The results point to so far new and interesting patterns in the structure and formation of lee waves. These findings are compared with results from other measuring campaigns [4,5].

References

[1] Achatz, U., R. Klein, F. Senf: Gravity waves, scale asymptotics and the pseudo-incompressible equations. J. Fluid Mech. 663: 120 - 147, 2010

[2] Placke, M., P. Hoffmann, M. Gerding, E. Becker. M. Rapp: Testing linear gravity wave theory with simultaneous wind and temperature data from the mesosphere. J. Atmos. Sol.- Terr. Phys. 93: 57 - 69, 2013.

[3] Ultsch, A.: Swarm Data Mining for the Fine Structure of Thermals, Technical Soaring, Vol. 36, Nr. 4, pp. 37 - 44, 2013.

[4] Stromberg,I. M., Mill, C. S., Choularton, T. W. and Gallagher, M. W.: A case study of stably stratified airflow over the pennines using an instrumented glider, Boundary-Layer Meteorology, Volume 46, Numbers 1-2, pp 153-168, Springer, Netherlands, 1989

[5] Millane,R., Brown,P., Richard,G., Enevoldson,E., Murray,J.E.: Estimating mountain wave windspeeds from sailplane flight data, Proceedings of SPIE - The International Society for Optical Engineering, E01, pp104-109, 2004.

Does Landscape Attractiveness affect Land Consumption in Germany?

Martin Behnisch[1] and Alfred Ultsch[2]

[1] Leibniz Institute of Ecological and Regional Development, Weberplatz 1, 01217 Dresden
m.behnisch@ioer.de

[2] Databionics Research Group, University of Marburg, Marburg, Germany
ultsch@informatik.uni-marburg.de

Abstract. Land consumption means the transformation of open space into settlement and transportation area (Thomas 2011). Excessive land consumption is a contradiction to sustainable land management (Helming et al. 2007). A fine grained data base on 11441 municipalities in Germany allows the empirical investigation of determinants of land consumption (Behnisch et al 2014, Krueger et al. 2013). Do people build their houses and infrastructure where the landscape is beautiful? Or do they settle where large spaces are available? Here a new indicator for Landscape Attractiveness is analysed on the administrative level of municipalities (Stein/Walz 2015). The aim is to quantify the influence of factors that are associated with land consumption on the variable Landscape Attractiveness. The influence on land consumption is multifactorial. Furthermore, it is likely that there are subgroups of municipalities, for example, coastal municipalities which are characterized by many touristic facilities and suburban communities located in attractive natural environments (Hersperger/Buergi, 2009). Data Mining techniques for the identification of subsets and in particular density modeling as, for example, Gauss mixture models, are required for modeling these types of influences (Hand et al. 2001). Machine Learning and Knowledge Discovery techniques, which have been developed and used for spatial research (Behnisch/Ultsch 2015, Ultsch et al. 2015) are applied in order to produce understandable descriptions of land consumption patterns with regard to attractiveness. The results are both relevant in support of decision-making processes as well as in the observation and monitoring of spatially referenced development and planning processes.

References

Behnisch, M., Ultsch, A.: Urban data mining: spatiotemporal exploration of multidimensional data, Building Research & Information, 37(5-6), 520-532, 2009.

Behnisch, M., Ultsch, A.: Knowledge Discovery in Spatial Planning Data - A Concept for Cluster Understanding, in: Helbich, M., Arsanjani, J.J., Leitner, M. (Eds.): Computational Approaches for Urban Environments, Geotechnologies and the Environment Series, Volume 13, Springer, Berlin, 49-75, 2015.

Behnisch, M., Kretschmer, O., Ultsch, A.: Towards an Understanding of Land Consumption Outline of Influential Factors as a Basis for Multidimensional Analyses, Erdkunde - Archive for Scientific Geography, University of Bonn (forthcoming).

Dempster, A. P., Laird, N. M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B (Methodological), 39(1), 1-38, 1977.

Hand, M., Mannila, H., Smyth, P.: Principles of Data Mining, MIT Press, Cambridge, MA, 2001. Helming, K., Perez-Soba, Tabbush, P.: Sustainability Impact Assessment of Land Use Changes, Springer, Berlin, 2007.

Hersperber, A. M., Buergi, M.: Going beyond landscape change description: Quantifying the importance of driving forces of landscape change in a Central Europe case study, Land Use Policy, 26(3), 640-648, 2009.

Krueger, T., Meinel, G., Schumacher, U.: Land-use monitoring by topographic data analysis, Cartography and Geographic Information Science, 40(3), 220-228, 2013.

Stein, C., Walz, U.: Indikator zur landschaftlichen Attraktivität Deutschlands, in: M. Behnisch and Alfred Ultsch (Eds.): Beiträge zur Erforschung von Einflussgrößen und Regelhaftigkeiten der Flächeninanspruchnahme. Springer Spektrum, Heidelberg, (forthcoming).

Thomas, J. (2011): Uncontrolled land consumption versus resource saving land use in Germany, Land Tenure Journal, 1, 79-100.

Ultsch, A., Kretschmer, O., Behnisch, M.: Systematic Data Mining into Land Consumption in Germany, Proc. of the 14th International Conference on Computers in Urban Planning and Urban Management (July 7–0, 2015, Cambridge, MA USA), 2015 (forthcoming).

Analysing past settlement size and location

Irmela Herzog

The Rhineland Commission for Archaeological Monuments and Sites
The Rhineland Regional Council i.herzog@lvr.de

Abstract. In archaeology, many studies analyse the settlement patterns of some period in the past. These studies are mainly based on biased samples of find spot data that hardly allow estimating the settlement size, its function and the exact time period of use. A historical map finished in 1715 indicating not only settlement locations but also their sizes forms a more reliable data basis. Kolmogorov-Smirnov tests are applied to identify relevant variables for settlement locations namely slope, soil quality, local prominence and least-cost distance to creeks. The tests show that the settlement pattern of single farmsteads differs from that of settlements formed by two or more farmsteads. Moreover, least-cost Thiessen polygons and least-cost triangulations show that settlement size has an impact on the territory allocated to the settlement and on the distance to the neighbouring settlements.

References

HERZOG, I. (2014), Testing Models for Medieval Settlement Location. In: M. Spiliopoulou, L. Schmidt-Thieme, and R. Janning (Eds.): *Data Analysis, Machine Learning and Knowledge Discovery*. Springer, Cham Heidelberg New York Dordrecht London, 351–358.

Keywords

KOLMOGOROV-SMIRNOV TEST, LEAST-COST PATHS, THIESSEN POLYGONS, TRIANGULATION

Finance and Economics I

Friday, September 4, 2015: 9.00am - 10.15am Room: EBS 2.65

Integrated measures of household risk in financial plans optimized under a consumption rate constraint

Radoslaw Pietrzyk[1] and Pawel Rokita[2]

[1] Wroclaw University of Economics radoslaw.pietrzyk@ue.wroc.pl
[2] Wroclaw University of Economics pawel.rokita@ue.wroc.pl

Abstract. Measures of risk suited to life-long financial plan of a household differ from other popular risk measures that allow for integration of a number of risk types in a one measure, like for instance Value at Risk, Expected Shortfall (financial institutions and investors) or Earnings at Risk and Cash Flow at Risk (enterprises). Household measure of risk should address threats to accomplishment of life objectives of the household and must take into account its life cycle. In this research, there are presented some propositions of new downside risk measures that fulfill these conditions. The focus is here on analyzing how risk measured in this way reacts when financial plan is optimized under some enforced levels of consumption. A particular question of interest is if optimization of a financial plan for a given level of consumption leads to reduction of risk. A more general research issue is whether it is possible to identify any tradeoff between expected consumption and the level of risk of the whole financial plan.

References

BODIE, Z., DETEMPLE, J. B., OTRUBA, S., and WALTER, S. (2004). Optimal consumption portfolio choices and retirement planning. *Journal of Economic Dynamics and Control, 28(6), 1115–1148.*

COCCO, J. F., GOMES, F. J. (2012). Longevity risk, retirement savings, and financial innovation. *Journal of Financial Economics, 103(3), 507–529.*

SIEGMANN, A., LUCAS, A. (2005). Discrete-Time Financial Planning Models Under Loss-Averse Preferences. *Operations Research, 53(3), 403–414.*

Keywords

INTEGRATED MEASURE OF RISK, HOUSEHOLD, FINANCIAL PLANNING, CUMULATED SURPLUS MODEL, LIFE CYCLE

The Analysis of Consumers' Preferences with the Application of Multivariate Models – Hedonic Regression and Multidimensional Scaling

Marta Dziechciarz-Duda[1] and Anna Król[2]

[1] Wrocław University of Economics marta.dziechciarz@ue.wroc.pl
[2] Wrocław University of Economics anna.krol@ue.wroc.pl

Abstract. The study attempts to confront the results obtained from multidimensional scaling and hedonic modelling to assess consumers' preferences with respect to attributes of chosen durable good. The research was performed using two sets of data concerning smartphones. Assessment of consumers' preferences was obtained by analyzing data from on–line survey study with the application of multidimensional scaling. Simultaneously the estimated hedonic model (basing on the dataset consisting in price lists from Polish Internet shops) provided the prices of goods' characteristics.

On durables markets declarative behaviours of buyers transpose rarely (or not at all) to the actual purchasing decisions. In addition, an important factor in changing consumers' preferences and habits is technological progress, which occurs in almost all sectors involved in the production of durable consumer goods. The combined use of multidimensional scaling and hedonic regression allowed for broader insight into the issue of consumers' preferences, particularly in relation to the existing market offer.

References

BORG, I., GROENEN, P.J.F. (2005): *Modern Multidimensional Scaling. Theory and Applications*, Springer-Verlag, New York.

DIEWERT, W.E. (2003), Hedonic Regressions. A Consumer Theory Approach. In: R.C. Feenstra and M.D. Shapiro (Eds.): *Scanner Data and Price Indexes*. University of Chicago Press, 317–348.

EVANS, M., JAMAL, A. and FOXALL, G. (2006): *Consumer Behaviour*, John Wiley & Sons, Hoboken.

WALESIAK, M., GATNAR, E. (Eds.) (2009): *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa.

Keywords

MULTIDIMENSIONAL SCALING, HEDONIC MODELLING, CONSUMERS' PREFERENCES, SMARTPHONES

Immigrants' Access to Healthcare System in Eastern Macedonia and Thrace Between 2005 and 2011

Theodosios Theodosiou[1], Persefoni Polychronidou[1], and Anastasios G. Karasavvoglou[1]

Department of Accounting, Eastern Macedonia and Thrace Institute of Technology, Ag. Loukas, Kavala, 654 04, Greece
theodosiou@statnous.com, polychr@teikav.edu.gr,
akarasa@teikav.edu.gr

Abstract. The main goal of the present research is to assess the immigrants access to the Greek National Healthcare System for the region of Eastern Macedonia and Thrace (EMT) between years 2005 and 2011. The data are from five different hospitals belonging to the Greek National Healthcare System. Four characteristics were analyzed, namely the duration of hospitalisation, the diagnosis, the cost of hospitalisation and the nationality, using statistical data mining methods.

The results indicate that the cost of hospitalisation for the Greek patients is higher for all the years compared to the immigrants for the hospitals of Kavala and Komotini, whereas for the hospitals of Didimoticho and Drama there is no significant difference. The cost of hospitalisation for the Greeks at the hospital of Xanthi is lower compared to the one for the immigrants. The immigrants have higher duration of hospitalisation compared to Greek patients for the hospitals of Kavala, Xanthi and Didimoticho, whereas for the hospitals of Drama and Komotini the duration does not differ between immigrants and Greek patients.

ACKNOWLEDGEMENTS: This research was supported by the Project ”Immigrants and Health Services - The case of Eastern Macedonia and Thrace region” that is co-funded by the European Union (European Social Fund) and National Resources - ARCHIMEDES III.

References

LLOP-GIRONES, A., VARGAS LORENZO, I., GARCIA-SUBIRATS, I., AILLER, MB. and VAZQUEZ NAVARRETE, ML. (2014): Immigrants' access to health care in Spain: a review. *Rev Esp Salud Publica, 88:6, 715–734, 2014.*

FIELD, A., MILES, J., FIELD, Z. (2012): *Discovering Statistics Using R.* SAGE, London.

Keywords

IMMIGRANTS, HEALTHCARE SYSTEM, HOSPITALISATION, COST

Musicology

Friday, September 4, 2015: 9.00am - 10.15am Room: EBS 2.66

Multivariate Supervised Classification for Tone Onset Detection

Nadja Bauer, Klaus Friedrichs and Claus Weihs

TU Dortmund, Chair of Computational Statistics
{bauer, friedrichs, weihs}@statistik.tu-dortmund.de

Abstract. The classical onset detection approach consists of the following steps: splitting the ongoing signal in possibly overlapping windows, computing an onset detection function (ODF) in each window and picking the local maxima of ODF exceeding a certain threshold value. Some ODFs use the increase of the signal amplitude as an indicator of a tone onset while other consider signal spectral or phase information. The drawback of the classical approach is its limitation to only one detection function. Since a binary decision (onset or no onset) has to be done in each window, applying classification methods appears to be very self-evident. In this case, not only one but many ODFs can be used to train an appropriate classification model.

However, there are many differences to the usual classification tasks. Firstly, the goodness of the classification rule cannot be assessed by the confusion matrix since not the direct matches between the estimated and the true (0,1)-onset vectors are essential, but whether the estimated onset times lie within a tolerance interval around the true onset times. Secondly, due to possible window overlapping an onset can be marked in many neighboring windows while only one mark would be correct. Lastly, the classification rule depends on the window size and the overlap so that, if these both parameters are optimized, a new rule has to be fitted for every combination.

Only in very few papers classification has been applied to tone onset detection. Here, we propose a novel strategy for considering many ODFs and applying the classification model for estimating onset detection times. The main idea is handling the tone onset probability predicted by the classification model as an ODF. We compare the multivariate approach with the classical one and discuss aspects of its further developing.

Keywords

TONE ONSET DETECTION, MULTIVARIATE CLASSIFICATION

Learning Statistical Regularities of a Simple Musical ‘Genre’: The Case of Radio Station Jingles

Daniel Müllensiefen[1], Hauke Egermann[2], and Sean Burrows[1]

[1] Department of Psychology, Goldsmiths, University of London
d.mullensiefen@gold.ac.uk
[2] Fachgebiet Audiokommunikation, Technische Universität Berlin
hauke.egermann@tu-berlin.de

Abstract. Radio Station Jingles are the fruit flies of systematic music research. They are simple enough to be modelled fully in terms of the features of their musical structure and yet they are complete musical objects that exist in their own right, serve a cognitive function for listeners and are of commercial value to radio stations. This paper aims to identify the statistical regularities of a corpus of radio station jingles by describing the boundaries of this ‘musical genre’ with a probabilistic model that is based on automatically extracted melodic features (Müllensiefen & Halpern, 2014). Subsequently we assess the perceptual validity of the probabilistic model with data from two listening experiments that test whether ordinary radio listeners internalise the frequency distributions of melodic features of the radio jingles genre.

A corpus of 92 radio station jingles was compiled. Continuous melodic features with a Gaussian distribution (such as mean pitch interval size) were modelled using kernel density estimation; count features (such as jingle length) were modelled with a negative binomial model; and for categorical features (such as tonality) probability estimates were derived from relative class frequencies. Feature probabilities were then used as predictor variables to model the listener response data from two listening experiments showing that ordinary radio listeners are indeed able to distinguish between jingles that adhere to the stylistic boundaries and those that do not. Backwards model selection was used to arrive at a model that makes use of only 3 features reflecting major v. minor tonality, melodic contour and pitch interval size that are sufficient to describe listener behaviour.

References

MÜLLENSIEFEN, D. and HALPERN, A. (2014): The role of features and context in recognition of novel melodies. *Music Perception, 31, 418-435*.

Keywords

MUSICAL CORPUS ANALYSIS, STATISTICAL PERCEPTUAL LEARNING, RADIO STATION JINGLES, MELODIC FEATURE ANALYSIS

Near-neighbour search in acoustic feature spaces: a case study in contrafactum and parody

Christophe Rhodes[1]

Goldsmiths, University of London, New Cross, SE14 6NW, United Kingdom
c.rhodes@gold.ac.uk

Abstract. One problem often encountered in the study of historical creative works is that of attribution: works may have survived only in anthologies, collections, or transcriptions; they may have been misattributed or mis-catalogued when first published or at later stages in their history. Attribution of newly-discovered artworks is an ongoing process; as new information comes to light, previous decisions are revised and refined by the scholarly community.

We present our work on applying near-neighbour search in a high-dimensional acoustic feature space (Casey *et al.*, 2008), using state-of-the-art acoustic chroma features (Mauch and Dixon, 2010) to investigate a corpus of recordings of sixteenth-century polyphony. Specifically, we report on the capability of the tools to reproduce musicological judgments such as that of Picker (2001), where some works previously attributed to Josquin des Prez were found to be by Nicolas Gombert. We provide the data and publish the workflow involved in order to make this investigation reproducible by other researchers.

References

CASEY, M., RHODES, C. and SLANEY, M (2008): Analysis of Minimum Distances in High-Dimensional Musical Spaces. *IEEE Transactions on Audio, Speech and Language Processing, 16:5, 1015–1028*

MAUCH, M. and DIXON, S. (2010): Approximate Note Transcription for the Improved Identification of Difficult Chords. In: Proc. International Society for Music Information Retrieval Conference, Utrecht, Netherlands, 135–140

PICKER, M. (2001): A spurious motet of Josquin, a chanson by Gombert, and some related works: A case study in contrafactum and parody. In: Quellenstudium und musikalische Analyse: Festschrift Martin Just zum 70. Geburtstag, 33–45, Ergon–Verlag

Keywords

MUSIC CORPUS ANALYSIS, NEAREST-NEIGHBOUR SEARCH, ATTRIBUTION, SIMILARITY

Classification

Friday, September 4, 2015: 10.45am - 12.00pm Room: EBS 2.2

New Methods for the Classification of inequally distributed Data: ABC-plots and computed ABC-analysis

J. Lotsch[1] and A. Ultsch[2]

[1] Institute of Clinical Pharmacology, Goethe - University, Frankfurt am Main, Germany
j.loetsch@em.uni-frankfurt.de
[2] Databionics Research Group, University of Marburg, Marburg, Germany
ultsch@informatik.uni-marburg.de

Abstract. The assessment of inequal distributions aiming at the selection of only the relevant items is an important step in data mining of high dimensional data [Foster/Stine 04]. Typical examples are the selection of relevant components for a principal component analysis (PCA) [Jolliffe 02] repectively independent component analyses (ICA) [Hendrikse et al 07], or the selection of variables used in symbolic classifiers in Machine Learning (e.g. CART, ID3 etc.) [Guyon/Elisseeff 03]. Procedures to identify the important few items (e.g. eigenvalues, variables, components) as opposed to the "trivial many" [Pareto 09, Juran 75] rely often on "cookbook recipes". This means the selection is based on heuristics with subjectively chosen and often unreported criteria. Recently ABC-plots and the computed ABC-analysis have been introduced [Ultsch/Lotsch 2015] and published in form of a R library on CRAN. ABC-plots display Lorenz-curves in a way that was already used by Lorenz himself in 1905 [Kleiber 05]. ABC-plots allow a sensible comparison criterion of inequal distributions with a suitable Uniform distribution, rather than with the unrealistic Identity distribution [Coulter 89]. The computed ABC-analysis is an algorithmic parameter-free classification of a distribution into distinct sets of the important versus the unimportant variables. In this work the properties of ABC-curves, the ABC-plot and the computed ABC-analysis will be presented. Applications of these methods to typical distributions found in data mining and knowledge discovery, in particular in "Big Data" from life sciences are given.

References

Coulter, P.B. (1989) Measuring Inequality, Westview Press.

Foster, D.P., Stine, R.A. (2004) Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy, Journal of the American Statistical Association, 99 pp. 303-313.

Guyon, I., Elisseeff,A. (2003, Eds.) Special Issue on Variable and Feature Selection, Journal of Machine Learning Research.

Hendrikse, A.J., Veldhuis, R.N.J., Spreeuwers, L.J. (2007) Component ordering in independent component analysis based on data power, 28th Symp.Information Theory in the Benelux, pp. 211-218, Enschede, The Netherlands.

Jolliffe, I.T. (2002) Principal Component Analysis, Springer Series in Statistics, 2nd ed., Springer, NY.

Juran, J.M. (1975) The Non-Pareto Principle - Mea Culpa, Quality Progress, 8, pp 8-9.

Kleiber, C. (2005) The Lorenz Curve in Economics and Econometrics, Technical Report, University of Dortmund, No.3.

Pareto, V. (1909) Manuale di economia politica, Milan: Societa editrice libraria, revised and translated into French as Manuel d'economie politique. Paris: Giard et Briere.

Ultsch, A., Lotsch, J. (2015) Computed ABC analysis for rational selection of most informative variables in multivariate data, PlosOne, in press.

Experimental evaluation of business document classifiers

Giorgi Bubashvili[1] and Adalbert F.X. Wilhelm[2]

[1] Jacobs University Bremen g.bubashvili@jacobs-university.de
[2] Jacobs University Bremen a.wilhelm@jacobs-university.de

Abstract. Enterprises are confronted with a rapidly growing amount of digital documents which need to be stored and processed further. Successful handling often requires transforming raw data generated by enterprise resource planning (ERP) systems into structured datasets based on classifying the document types. Our research aims at experimentally evaluating the performance of different tree-based and regression-based classifications methods under varying parameter settings. The ultimate application goal is to develop a process flow-chart enabling experts to specify optimal trade-off points between the size of the training-dataset and the classification accuracy rate. This will be used to partition the data in a pre-processing step to increase prediction accuracy and decrease the misclassification risk. To keep the overall process time as low as possible the classifier performance needs to be augmented with modelling complexity proxies adding some penalising component. Based on the classifier performance evaluation on four typical but heterogeneous data sets, different penalising approaches are discussed regarding various complexity configurations.

References

SHARMA, A. K., SAHNI, S. (2011). A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering, 3(5), 1890–1895, 2011*

ZHANG, C., MA, Y. (2012). *Ensemble machine learning methods and applications.* New York: Springer.

HOTHORN, T., LAUSEN, B. (2002), Bagging combined classifiers. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Analysis*. Springer, Berlin, 203–210.

Keywords

DECISION TREES, ENSEMBLE METHODS, PENALTY TERM, PREDICTION ACCURACY

Decision Trees for the Imputation of Categorical Data

Tobias Rockel[1], Dieter William Joenssen[1], and Udo Bankhofer[1]

Ilmenau University of Technology, Helmholtzplatz 3, 98693 Ilmenau, Germany
Tobias.Rockel@tu-ilmenau.de
Dieter.Joenssen@tu-ilmenau.de
Udo.Bankhofer@tu-ilmenau.de

Abstract. Resolving the problem of missing data via imputation can be reduced to two steps. First, the missing values are predicted. Second, the predicted values are copied over the missing values. While the second step is straightforward, any prediction method may fundamentally be used for the first step. One type of predictive model that may be employed for the missing data is the classification tree. However, literature on the predictive accuracy of these machine learning algorithms in the missing data context is scant to date. Therefore, the aim of this paper is to analyze the imputation quality offered by classification trees. To this end, real data are used in a simulation study to gauge which factors influence the performance of classification trees in different missing data settings. The study design includes various stochastic and deterministic missing data methods and other factors, such as the missingness mechanism. An evaluation of different quality measures indicates how classification trees fare in predicting missing values.

References

BREIMAN, L., FRIEDMAN, J.H., STONE, C. and OLSHEN, R.A. (1998): *Classification and Regression Trees*. CRC Press, Boca Raton.

van BUUREN, S. (2012): *Flexible Imputation of Missing Data*. CRC Press, Boca Raton.

LITTLE, R.J.A. and RUBIN, D.B. (2002): *Statistical Analysis with Missing Data*. Wiley, Hoboken.

QUINLAN, J.R. (1993): *C4.5. Programs for Machine Learning*. Morgan Kaufmann, San Mateo.

Keywords

DECISION TREES, IMPUTATION, PREDICTIVE ACCURACY, MISING DATA

Digital Humanities and Social Sciences II

Friday, September 4, 2015: 10.45am - 12.00pm Room: EBS 2.34

Performance of Text Mining of open question vs. Likert-Scale questions in predicting learning outcomes

Karsten Lübke, Bianca Krol, Stefan Ebener, and Rüdiger Buchkremer

FOM Hochschule für Oekonomie & Management, Leimkugelstraße 6, 45141 Essen, Germany karsten.luebke@fom.de

Abstract. There is a long history of scientific research on learning skills, learning strategies and learning aids of students and the link of these to academic success (Diseth, 2011).

Measurement of the application of learning strategies on student success has often been performed via some variation of the Likert-Scale questionnaire (Carey et al., 2014). However, in the context of our learning strategy survey we hereby report an accompanying text analytics study of unstructured text comments (text retrieval, text mining, see Dhillon, 2001). The empirical results are based on an online-survey on learning strategies of students studying while working where 245 out of 297 also filled in the open-question: Which learning strategy was successful for you?

The Likert-Scale items on the one hand are compared to the concepts extracted by text mining techniques on the other hand. Both are benchmarked (Hothorn et al., 2005) in terms of their linking to reported learning outcome based on the residuals after controlling for covariates like e.g. gender, subject or prior-knowledge.

References

CAREY, J., et al. (2014): Development of an Instrument to Measure Student Use of Academic Success Skills - An Exploratory Factor Analysis. *Measurement and Evaluation in Counseling and Development, 47, 171–180, 2014.*

DHILLON, I.S. and MODHA, D.S. (2001): Concept decompositions for large sparse text data using clustering. *Machine learning, 42, 143–175, 2001.*

DISETH, Å. (2011): Self-efficacy, goal orientations and learning strategies as mediators between preceding and subsequent academic achievement. *Learning and Individual Differences, 21, 191–195, 2011.*

HOTHORN, T., et al. (2005): The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14, 675–699, 2005.*

Keywords

LEARNING OUTCOMES, TEXT MINING, LIKERT-SCALE ITEMS, BENCHMARKING

The comparison and influence of different estimation methods on the parameter estimates and fit indices in SEM models under 7-point Likert scale

Piotr Tarka[1]

Poznan University of Economics, Department of Market Research
piotr.tarka@ue.poznan.pl

Abstract. The article discusses the issues and problems resulting from the impact of different methods of estimation on the level of the obtained parameters and the goodness-of-fit of the data in context of construction of Structural Equation Models (SEM), pertaining to the data collected on the 7-point Likert scale. The objective of the conducted experimental analysis was to compare selected data estimation methods such as ML, MLM, MLV, MLMV, WLS, WLSM and WLSMV, for which the parameter statistics were calculated, and for which, the quality of fit of the respective SEM model was evaluated. Eventually, from the list of all presented methods in the article, the author selected the most optimal approach in accordance with the scale of measurement which was under consideration. The area of the empirical study and the subject of investigation was related to the market research methodology effectiveness and application of structural equation data modeling in the sample (N=391) of firms.

References

RAYKOV, T. and WIDAMAN, K.F. (1995): Issues in applied structural equation modeling research. *Structural Equation Modeling, 2, 4, 289–318.*

FLORA, B.D, and CURRAN, P.J. (2004): An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9, 4, 466–491.*

BOLLEN, K.A. (2010): A comment on model evaluation and modification. *Multivariate Behavioral Research, 25, 2, 181–185.*

Keywords

ESTIMATION METHODS, FIT INDICES, 7-POINT LIKERT SCALE, STRUCTURAL EQUATION MODELS

A tool for Measuring Behavioral health and Forecasting healthy life years

Artur Parreira[1] and Ana Lorga da Silva[2]

[1] Fundação CESGRANRIO, Brasil; CPES, ULHT, Lisboa, Portugal
arturmparreira@gmail.com
[2] ECEO, CPES - ULHT, Campo Grande 376, 1749-024 Lisboa, Portugal and
CEDRIC-CNAM, Paris, France ana.lorga@ulusofona.pt

Abstract. This work is based on the concept of behavioral health, as a set of healthy behaviors in the different areas of everyday life This set of health factors that can serve as a basis for assessing the health status of a population and forecasting healthy life years. To build a practical and easily usable instrument for measuring these patterns of behavior and life skills is a way for the adoption of better practices and a better state of health. The goal of this study is to build an easily usable tool to define the level of behavioral health . To assure the validity and consistency of the instrument, the authors chose behavioral patterns strongly based on literature (Report PHEIAC-EU, 2013) This instrument has the form of a questionnaire, covering the main areas of the construct of behavioral health: dietary patterns, (DRISKELL *et al*, 2008.): sleep and circadian rest; physical exercise and body maintenance; relationship experiences in the organization of everyday life, including behaviors at work (WHO, 2010), and financial and environmental safety. In this study, it undergoes a process of validation in order to ensure the empirical validity, internal consistency and the definition of criteria and standards for interpretation and evaluation of results.

References

DRISKELL, M.M., DYMENT, S., MAURIELLO, L., CASTLE, P. and SHERMAN, K. (2008): Relationships among multiple behaviors for childhood and adolescent obesity prevention. *Prev Med. 46(3):209–15, 2008*

EVERITT, B. S. AND DUNN, G. (2010):*Applied Multivariate Data Analysis, Wiley*

OMS (WHO) (2010):*Ambientes de trabalho Saudáveis. Brasília: Edição SESI*

PHEIAC-EU Report (2013):*Evaluation of the use and impacto f the Community Heakth Indicators. EU.*

Keywords

BEHAVIORAL HEALTH, FORECASTING, MEASURE OF PATTERNS, MULTIVARIATE DATA ANALYSIS

Mathematical Foundations of Data Science

Friday, September 4, 2015: 10.45am - 12.00pm Room: EBS 2.1

Graph-Based Lattice Similarity Mapping

F. Domenach[1] and Z. Rajabi[1]

Computer Science Department, University of Nicosia, 46 Makedonitissas Ave., PO Box 24005, 1700 Nicosia, Cyprus `domenach.f@unic.ac.cy, zr.1995@yahoo.com`

Abstract. This paper is in the formal concept analysis framework, an algebraic hierarchisation method of data based on the notion of extent / intent, i.e. of maximally shared attributes and objects. We present here a heuristic for graph-based similarity measures between two formal concept lattices, and compare it to results of a previous paper which introduced a structured-based dissimilarity between concept lattices defined on the same set of attributes. We define an expressive model using a mapping between objects of the two lattices. A key point of our approach is that the correspondence in this mapping may not be unique and may associate each object of the first lattice to several objects of the other one.

References

CHAMPIN, P.-A. and SOLNON, C. (2003): Measuring the Similarity of Labeled Graphs, in Ashley, K. and Bridge, D. (ed.), *Case-Based Reasoning Research and Development*, Springer Berlin Heidelberg, 2689, 80-95

DOMENACH, F. (2015): Similarity Measures of Concept Lattices, in B. Lausen et al. (eds.), *Data Science, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-662-44983-7_8, pp. 89-99

GANTER, B. and WILLE, R. (1999): *Formal Concept Analysis: Mathematical Foundations*. Springer.

Keywords

FORMAL CONCEPT ANALYSIS, LATTICE, SIMILARITY MEASURE

Weak Invariant Measures of an Action of the Automorphism Group of a Graph

Fabian Ball[1] and Andreas Geyer-Schulz[2]

[1] Karlsruhe Institute of Technology fabian.ball@kit.edu
[2] Karlsruhe Institute of Technology andreas.geyer-schulz@kit.edu

Abstract. Graph automorphisms cause the following problem with partition comparison measures used for the evaluation of cluster solutions: They violate the axiom that $d(x,y) = 0$ implies that $x \equiv y$, (d is a measure between two points x and y of some space S). In measure theory, e.g. Doob (1994) handles this problem formally by replacing this axiom of a metric space by the axiom of pseudometric space which states that $d(x,y) = 0$ does not imply $x \equiv y$. In applications, the transformation from a metric space to a pseudometric space requires the identification of the equivalence classes defined by the automorphism groups and the replacement of members of an equivalence class by the “canonical” representative of the class before evaluating the solution of the cluster problem. Finding the automorphism group of a graph is equivalent to solving the graph-isomorphism problem.

In this contribution we identify three invariant measures of group actions (the Kolmogorov-Sinai entropy, modularity and the partition type) and investigate how a combination of these measures can be used for extracting graph automorphisms from a sample of dendrograms which are the result of efficient ensemble learning algorithms for modularity clustering (see Ovelgönne and Geyer-Schulz (2012)).

References

DOOB, J. L. (1994): Measure Theory. Springer, New York.

OVELÖNNE, M. and GEYER-SCHULZ, A. (2012): An Ensemble Learning Strategy for Graph-Clustering. In: D. A. Bader, H. Meyerhenke, P. Sanders, D. Wagner (Eds.): *10th DIMACS Implementation Challenge – Graph Partitioning and Graph Clustering*. DIMACS, Rutgers University, Piscataway.

Keywords

ENTROPY, GRAPH AUTOMORPHISMS, INVARIANCE, CLUSTER EVALUATION

The Constructive Implicit Function Theorem and Proof in Logistic Mixtures

Xiao Liu and Ali Ünlü

Chair for Methods in Empirical Educational Research, TUM School of Education and Centre for International Student Assessment (ZIB), TU München, Arcisstr. 21, 80333 Munich, Germany {x.liu, ali.uenlue}@tum.de

Abstract. There is the work by Bridges et al. (1999) on the key features of a constructive proof of the implicit function theorem, including some applications to physics and mechanics. For mixtures of logistic distributions such information is lacking, although a special instance of the implicit function theorem prevails therein. The theorem is needed to see that the ridgeline function, which carries information about the topography and critical points of a general logistic mixture problem, is well-defined (Liu and Ünlü (2014)). In this paper, we express the implicit function theorem and related constructive techniques in their multivariate extension and propose analogs of Bridges and colleagues' results for the multivariate logistic mixture setting. In particular, the techniques such as the inverse of Lagrange's mean value theorem (Sahoo and Riedel (1998)) allow to prove that the key concept of a logistic ridgeline function is well-defined in proper vicinities of its arguments.

References

BRIDGES, D., CALUDE, C., PAVLOV, B. and STEFANESCU, D. (1999): The Constructive Implicit Function Theorem and Applications in Mechanics. *Chaos Solitons and Fractals, 10, 927–934*.

LIU, X. and ÜNLÜ, A. (2014): Multivariate Logistic Mixtures. Talk at the *European Conference on Data Analysis* (*ECDA*) *2014*. Bremen, Germany.

SAHOO, P.K. and RIEDEL, T. (1998): *Mean Value Theorems and Functional Equations*. World Scientific, New Jersey.

Keywords

CONSTRUCTIVE IMPLICIT FUNCTION THEOREM, LOGISTIC MIXTURE, LAGRANGE MEAN VALUE THEOREM, RIDGELINE

Finance and Economics II

Friday, September 4, 2015: 10.45am - 12.00pm Room: EBS 2.65

The Application of the GlueVaR Measure in Risk Assessment on the Metal Market

Grażyna Trzpiot[1] and Dominik Krezolek[1]

University of Economics in Katowice, Department of Demography and Economic Statistics

Abstract. The purpose of the study is the application of a new risk measure called GlueVaR. This measure is closely related to Value-at-Risk and Conditional VaR, and its main advantage is the property of subadditivity. The property of subadditivity has a particular application in risk measurement, especially in extreme risk measurement. The research area chosen for this study is the metal market.

Keywords

GlueVaR, RISK MEASUREMENT, METAL MARKET

Hedonic Price Indices on the Apartments' Secondary Markets in Poland

Anna Król[1] and Marta Targaszewska[2]

[1] Wrocław University of Economics anna.krol@ue.wroc.pl
[2] Wrocław University of Economics marta.targaszewska@ue.wroc.pl

Abstract. The presented research addresses the problem of efficient methods of measuring actual price change on the real estate markets. In the class of strictly heterogeneous goods the price of a good observed in the period t usually may only be compared with the price in period $t+1$ of „similar" good. Because of the above mentioned fact the problem of quality difference is immanent feature of price measurement process on real estate markets (for example, there do not exist two identical properties on the apartments' market due to the fact that the prices on this market are highly influenced by neighbourhood-related and location-related attributes, such as the city district or the distance from city centre to the apartment).

In response to this problem hedonic methods of adjusting price indices to disparities in the goods quality have been applied. Hedonic regression describes the relationship between price of good (P) and the set of its characteristics X by certain function f: $P = f(X;\beta;\varepsilon)$. The estimate of the vector of parameters β, obtained by estimation of correctly specified hedonic regression model, allows to calculate theoretical price of a given good with specified set of significant characteristics which allows to measure „true" price change. In presented research direct methods for calculating hedonic price indices have been applied for data from secondary apartments' markets in five biggest Polish cities: Warszawa, Kraków, Łódź, Wrocław and Poznań.

References

AIZCORBE, A.M. (2014): *A Practical Guide to Price Index and Hedonic Techniques*. Oxford University Press, Oxford.

TRIPLETT, J. (2006): *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes*. OECD Directorate for Science, Technology and Industry, OECD Publishing, Paris.

VON DER LIPPE, P. (2007): *Index Theory and Price Statistics*. Peter Lang Verlag, Bern.

Keywords

HEDONIC PRICE INDEX, APARTMENTS, PRICE STATISTICS

Models of Income Distributions for Knowledge Discovery

Alfred Ultsch[1], Stefan Schnabel[1] and Michael C. Thrun[1]

Databionics Research Group, University of Marburg, Germany

Abstract. Descriptions of income distributions using a single distribution, like Lognormal or Gamma are often quite poor in describing the tails of the distribution [1]. This led to separate models for the upper vs. lower parts of income distributions [2]. For example [3-5] describe the high-income region with the Pareto power laws. Other authors model the low to medium income region using Exponential [6], Lognormal [7] or Gamma distri- butions [8, 9]. The high income range is often modeled using the cumulative distribution function (cdf) [10], whereas the low to medium income regions are modeled using the probability density function (pdf) [11]. Usually no systematic limits between low, medium and high income are defined [3]. A goal for a valid and suitable model for income distributions is to derive a theory of the mechanisms which operate in a society (Computational Social Science) and explains the observed distribution [12]. Here a model for income distributions as a mixture of components is proposed. The model is derived using the Pareto Density Estimation (PDE) [14] for an estimation of the pdf. PDE has been designed in particular to identify groups/classes in a dataset [13]. Precise limits for the classes can be calculated using the theorem of Bayes. Our model suggests that there are different groups/classes in a society, which contribute to the total distribution of income in their own way. The approach is demonstrated on several real world data sets including actual income data from Germany.

References

1. Dagum, C., New Model of personal Income-Distribution-Specification and Estimation, Economie appliquee, 30(3), pp 413-437, 1977.
2. Richmond, P., et al., A Review of empirical Studies and Models of Income Distributions in Society. Wiley, Berlin, 2006.
3. Chatterjee, A., B.K. Chakrabarti, and R.B. Stinchcombe, Master Equation for a kinetic Model of a Trading Market and its analytic Solution. Phys Rev E Stat Nonlin Soft Matter Phys. 72(2 Pt 2, p. 026126, 2005.
4. Levy, M. and S. Solomon, New Evidence for the power-law Distribution of Wealth, Physica A: Statistical Mechanics and its Applications, 242(1), p. 90-94, 1997.
5. Dragulescu, A. and V.M. Yakovenko, Exponential and power-law Probability Distributions of Wealth and Income in the United Kingdom and the United States. Physica A: Statistical Mechanics and its Applications, 299(1), pp 213-221, 2001.

6. Chakrabarti, A.S. and B.K. Chakrabarti, Statistical theories of income and wealth distribution. Economics: The Open-Access, Open-Assessment E-Journal. 4, p. 4, 2010.

7. Clementi, F. and M. Gallegati, Pareto's law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States, in Econophysics of wealth distributions, Springer, New York, pp 3-14, 2005.

8. Ferrero, J.C., The statistical Distribution of Money and the Rate of Money Transference. Physica A: Statistical Mechanics and its Applications, 341, p. 575-585, 2004.

9. Scafetta, N., S. Picozzi, and B.J. West, An out-of-equilibrium Model of the Distributions of Wealth. Quantitative Finance. 4(3): pp 353-364, 2004.

10.Di Matteo, T., T. Aste, and S. Hyde, Exchanges in complex Networks: Income and Wealth Distributions, arXiv preprint, 2003.

11.Dragulescu, A. and V.M. Yakovenko, Evidence for the exponential Distribution of Income in the USA. The European Physical Journal B-Condensed Matter and Complex Systems, 20(4), pp 585-589, 2001.

12.Cioffi-Revilla, C., Introduction to Computational Social Science: Principles and Applications. Springer, Berlin, 2013.

13.Ultsch, A., Pareto Density Estimation: A Density Estimation for Knowledge Discovery, in Innovations in classification, data science, and information systems, Springer, New York, pp 91-100, 2005.

Part IV

LIS'2015 Workshop

LIS Workshop - day 1

Wednesday, September 2, 2015: 11:15am - 4:40pm Room: EBS 2.66

Building the Bridge: Mapping Different Knowledge Organization Systems in Economics

Andreas Oskar Kempf and Joachim Neubert

ZBW – Leibniz Information Centre for Economics, 20354 Hamburg, Germany
a.kempf@zbw.eu j.neubert@zbw.eu

Abstract. In economics researchers are used to classify their research papers according to the JEL classification codes[1], a classification system quarterly published by the American Economic Association and assigned by economists in addition to author keywords. Aside from this exists the STW Thesaurus for Economics[2] a domain-specific controlled vocabulary maintained by the German National Library, and mainly used for indexing by information specialists.

Here we deal with the question in how far both knowledge organization systems could be mapped onto each other? Building on economists' knowledge of the JEL codes we, this way, would like to encourage economists to use a standardized vocabulary for indexing their own research papers. Starting with a discussion on methodological issues we present a mapping between the JEL classification codes and the STW on the level of its category system. In doing so we are testing an automated vocabulary alignment tool. We will conclude by a discussion on whether a useful mapping is possible which could motivate the development of a web service that on the basis of the JEL codes offers economists STW categories they could select subject headings from to index their own research papers.

References

DEXTRE CLARKE, S.G. (2010): In Pursuit of Interoperability: Can We Standardize Mapping Types?. In: F. Boteram et al. (Eds.): *Concepts in Context*. Ergon, Wuerzburg, 91-110.

JACOBS, J.-H. et al. (2010): Benefits of the CrissCross project for conceptual interoperability and retrieval. In: C. Gnoli and F. Mazzocchi (Eds.): *Paradigms and conceptual systems in knowledge organization*. Ergon, Wuerzburg, 236-241.

Keywords

KNOWLEDGE ORGANIZATION SYSTEMS, MAPPING, ECONOMICS

[1] https://www.aeaweb.org/econlit/jelCodes.php

[2] http://zbw.eu/stw

Towards a Comprehensive Knowledge Organisation System for the Engineering Domain

Elena Bernauer[1], Martin Mehlberg[2], Mila Runnwerth[2], and Gudrun Schmidt[1]

[1] WTI-Frankfurt eG, Ferdinand-Happ-Straße 32, 60314 Frankfurt am Main, Germany `firstname.familiyname@wti-frankfurt.de`

[2] German National Library of Science and Technology, Welfengarten 1B, 30167 Hannover, Germany `firstname.familyname@tib.uni-hannover.de`

Abstract. To improve information seeking services for the economically and scientifically significant engineering domain, the German National Library of Science and Technology and the information service provider WTI have launched the development of an expert knowledge organisation system (KOS) aiming at a domain ontology for engineering . We present the step-by-step approach of our joint venture by outlining what has been implemented so far and give an outlook on what lies ahead: providing a software environment for administrating and curating thesauri; semi-automated merging, linking, and cross-referencing of several domain-specific vocabularies (distributed approach to KOS development); semi-automated vocabulary enrichment by term-extraction from recent scientific literature; multi-lingual access. As a selected use case scenario we also present a new visual exploration concept for a KOS-based literature research within an information provider's stock.

References

FROSTERUS, M. and TUOMINEN, J. and PESSALA, S. and HYVÖNEN, E. (2015): Linked Open Ontology Cloud – Managing a System of Interlinked Cross-domain Light-weight Ontologies. `http://www.seco.tkk.fi/publications/submitted/frosterus-et-al-loo-cloud-model.pdf`

Keywords

KNOWLEDGE ORGANISATION SYSTEM (KOS), ENGINEERING

Subject Cataloguing in an RDA Framework – Strategies and Practical Experience from Germany

Heidrun Wiesenmüller

Stuttgart Media University wiesenmueller@hdm-stuttgart.de

Abstract. Based on FRBR, the new international cataloguing standard "Resource Description and Access" (RDA) does not only cover descriptive cataloguing, but also the relationships between works and their subject(s). When RDA was first published, the relevant chapters had not been worked out, but a new general chapter on subject relationships was added in April 2015. After describing the current framework which RDA offers for subject cataloguing, strategies and practical experience from Germany are explained. Since 2012, there is only one integrated authority file which is used for both descriptive and subject cataloguing. With the completion of the move to RDA in early 2016, the cooperation will be further intensified – although it proves difficult in some areas (e.g. records for manuscripts). The German subject indexing language "Regeln für den Schlagwortkatalog" (RSWK) will change considerably: Form aspects will no longer be part of the subject headings string, but will be transferred to the RDA element "Nature of content" (RDA 7.2). However, it seems obvious that RDA as a general framework can never replace specific subject indexing standards like LCSH and RSWK.

References

WIESENMÜLLER, H. and HORNY, S. (2015): *Basiswissen RDA. Eine Einführung für deutschsprachige Anwender*. De Gruyter Saur, Berlin

Keywords

RESOURCE DESCRIPTION AND ACCESS (RDA), SUBJECT CATALOGUING, REGELN FÜR DEN SCHLAGWORTKATALOG (RSWK)

The Role of Classification Information in Open Access Repositories
current status and future directions

Friedrich Summann[1] and Dirk Pieper[2]

[1] Bielefeld University Library friedrich.summann@uni-bielefeld.de
[2] Bielefeld University Library dirk.pieper@uni-bielefeld.de

Abstract. Repositories have reached a strong role in the scholarly information landscape. Today thousands of implementations all around the world contain more than 100 million publications and related objects accompanied with high-level metadata. These metadata contain a broad variety of classification information. Bielefeld University Library runs the academic search engine BASE and harvests all these metadata via OAI-PMH. This is an excellent groundwork to explore and analyze the current situation and coverage of classification information in repositories. This presentation will give a survey of the current situation based on a detailed analysis of the data basis. It will show the used classifications and the quality level related to aspects as the repository type, geographical coverage and discipline relations.

Several activities in supporting automatic classification have been undertaken in the repositories context. Bielefeld UL has implemented an automatic classification tool aligned with BASE based on computer linguistic technology. As a result currently more than 10 Mill. documents could be enriched with automatic processed DDC codes stored in the BASE index.

For the future we expect a growing delivery of classification information and in combination with the increasing provision of open data interfaces (including classification-related information) new perspectives of adding and processing such type of data. In combination with efforts as metadata guidelines and vocabulary standards this will support the expansion of frequency and quality of classification information in repositories.

References

LOESCH, M., WALTINGER, U., HORSTMANN, W., and MEHLER, A. (2011). Building a DDC-annotated Corpus from OAI Metadata. Journal of Digital Information 12.

Keywords

REPOSITORIES, OAI-PMH, DDC, AUTOMATIC CLASSIFICATION

LIS Workshop - day 2

Thursday, September 3, 2015: 11:00am - 4:15pm Room: EBS 2.66

Patent Claim Structure Recognition

Rene Hackl-Sommer and Michael Schwantner

FIZ Karlsruhe {Rene.Hackl-Sommer,
Michael.Schwantner}@FIZ-Karlsruhe.de

Abstract. In many fields, information professionals are under unrelenting pressure to efficiently perform their tasks in the face of ever increasing amounts of data. This holds true all the more in the realm of patent search, notorious for being complex and difficult. In patents, the claims section is the judicially most relevant part. It is written in a very idiosyncratic style, containing independent and dependent claims, thus forming a hierarchy. We present our work aimed at identifying that hierarchy from the full text. Beginning with a short introduction into the subject matter of patent claims and typical use cases for searching in claims, we then proceed to show results from a preliminary context analysis that support developing the subsequent strategy. All English claims from the European Patents Fulltext (EPFULL) database were utilised in that analysis. We then give an impression of the myriad of possibilities with which claim dependency can be indicated in the text and show a way of how to capture them. Additionally, we describe several of the problem areas that were encountered, in particular pertaining to noisy data and representation and indexing challenges. Before concluding, we show results from our internal evaluations, in which an f-score greater than 0.95 was achieved. Lastly, some areas of further research are indicated.

Keywords

TEXT MINING, NATURAL LANGUAGE PROCESSING, INFORMATION EXTRACTION, PATENT CLAIMS, PATENTS

SMGloM a Semantic Mathematical Glossary of the Next Generation

Michael Kohlhase[1] and Wolfram Sperber[2]

[1] School of Engineering & Science, Jacobs University Bremen, Campus Ring 1, D-28759 Bremen, Germany m.kohlhase@jacobs-university.de
[2] Zentralblatt MATH, FIZ Karlsruhe, Franklinstr. 11, D-10587 Berlin, Germany wolfram@zentralblatt-math.org

Abstract. Glossaries are indispensable tools for the scholarly communication for a long time. In mathematics, glossaries appeared originally as alphabetical lists of used terms and symbols at the cover of books. Today, digital glossaries are a collective name for encyclopedias, thesauri, dictionaries, etc. containing a broad spectrum of ontological relations. Glossaries are in combination with controlled vocabularies and classification schemes an essential part of knowledge management in mathematics. In the talk, a concept and prototype of a new mathematics glossary, the Semantics Multilingual Glossary of Mathematics (SMGloM, http://mathhub.info/smglom/smglom) will be presented. Mathematics language is a natural language but it is different from common languages as English or German in some sense. Besides textual notations, symbols and formulae are often used for the presentation of mathematical objects or concepts allowing to present mathematical content in a highly condensed form. In other words, a duality between textual notations and symbols and formulas is characteristic for mathematics. The use of mathematical symbols and formulas is not self-evident, symbols and formulas are - as textual notations - not unique, the same symbol or formula can be used in different meanings, different symbols or formulas can be describe the same mathematical object. So, we have to differ between presentation and content. Wikipedia; Encyclopedia of Mathematics, Planet Math, MathWorld, and others have created worthy and large glossaries for mathematics. But, new concepts and languages for the encoding of mathematics content are necessary to develop more powerful tools for an enhanced semantic presentation and analysis of mathematics. New languages, as Semantic TEX the W3C standard MathML, or Open Mathematical Documents (OMDoc) and further theories for formalization of mathematical knowledge are used for the design of SMGloM.

The talk is focused to the design, the data model and possible use cases of SMGloM. A SMGloM entry describes a mathematical object or concept. Each entry consists of a definition, symbols (formal identifiers of the object or concept), and a set of verbalizations and notations. The verbalizations relate it to lexical entries. The definitions in SMGloM will be written in a common language, mathematical symbols are TEX encoded. Semantic annotations are realized in Seemantic TEXS. MGloM entries are embedded in their mathematical context. Other mathematical objects or concepts which are used in the definition of a SMGloM entry have to be explicitly declared in the SMGloM entry. SMGloM entries are realized by modules, the module signature contains the language independent parts as identifier, symbols and relations to other SMGlom entries. Language bindings contain multilingual definitions of the mathematical object or concept. SMGloM makes important ontological and linguistic realizations of mathematical object or concepts explicit. Symbols get a semantification by their direct linking to an mathematical object or concept. This allows new powerful applications of the glossary. A possible use case is an enhanced (machine-based) content analysis and classification: Identified symbols and nota- tions allow a semantic interpretation. Also a linking to the Mathematics Subject Classification (MSC) scheme is possible Currently, the SMGloM prototype covers more than 500 SMGloM entries of mathematical objects and concepts with more than 1,500 language bindings in English, German, Turkish, Romanian, and Chinese. The technical framwork of SMGloM is complex. Currently, tools for input, quality control (formal correctness and content), and the transformation of content, and access and user interfaces to SMGloM are under development. SMGloM is planned as a grassroot activity basing on volunteer contributions of the mathematical community. SMGloM is designed as a distributed system. SMGloM will be available as an Open Access service. The SMGloM prototype was developed by Jacobs University Bremen and FIZ Karlsruhe. The work was partially funded within a project of the Leibniz association.

Keywords

KNOWLEDGE MANAGEMENT, SEMANTIC WEG, ONTOLOGIES, MATHEMATICS ON THE WEB

The mapping tool "Cocoda"

Umamaheswari Balakrishnan[1] and Andreas Krausz[1]

Common Library Network GBV balakrishnan@gbv.de krausz@gbv.de

Abstract. The endeavor to develop a nation-wide classification scheme had led to the adoption of the DDC in the year 2000 by the national library of Germany. Since then number of mapping projects have been undertaken to map DDC with several local schemes. However, exhaustive concordances between different library KOS and the DDC are still rather rare. The reason for this lies in the lack of infrastructure that would accelerate the intellectual mapping process.

The project ”coli-conc” aims to address this issue and develop a framework that would

- facilitate semi-automatic generation of mappings
- ease their management
- allow an easy use and comfortable exchange of the same on a single platform and
- concurrently provide quality monitoring that would aid quality management.

To achieve the above objectives, a set of reusable software modules would be created so as to enable uniform access to knowledge organization systems, concordances and concordance assessments. These modules will be provided as a web application to support effective processing of concordances. In addition, existing software (KOS software, mapping algorithms, cataloguing software ...) will be evaluated and enhanced with new components for storage, access to and analysis of different concordances. These components will also be linked with each other through uniform and open APIs, so that a shared infrastructure is developed for the management, exchange and building of concordances.

The presentation will illustrate the initial works undertaken towards this aim. It will give an overview of the project ”coli-conc”, the works carried out so far within the scope of this project; demonstrate the features of the web-interface of the prototype concordance tool ”Cocoda”, briefly highlight the procedure and the architecture of the tool along with its technical details.

Keywords

CPLI-CONC, MACOING TOOL, COCODA, CONPORDANCES

Automatic Identification of Synonym Relations in the Dutch Parliament Thesaurus

Rosa Tsegaye Aga[1], Christian Wartena[1], Otto Lange[2], and Nelleke Aders[3]

[1] Hochschule Hannover, Expo Plaza 12, D-30539 Hannover
rosa-tsegaye.aga@hs-hannover.de@hs-hannover.de
[2] Next2Know, Grebbeweg 19b, NL-3902 HG Veenendaal otto@next2know.nl
[3] Dienst Informatievoorziening – Tweede Kamer der Staten Generaal, Postbus 20018,
NL-2500 EA Den Haag N.aders@tweedekamer.nl

Abstract. The Information Service of the Second Chamber of the Dutch parliament (*Tweede Kamer der Staten Generaal*) archives and indexes documents produced in the parliamentary process and other documents that are possibly relevant to the parliament. For indexing a special thesaurus covering all topics relevant to the society and the parliament is used. The Dutch Parliament is investigating additional alternatives to make information available, by full text search, automatic indexing and automatic classification. For good results for full text retrieval and automatic classification it turns out to be important to add more synonyms to the existing thesaurus terms.

In the present work we investigate the possibilities to find synonyms for terms of the parliaments thesaurus automatically. We propose to use distributional similarity (DS) for this. To test the potentials of DS we extracted over 6000 pairs of terms from the parliament thesaurus. Half of the pairs consists of two words that refer to the same concept, i.e. one of the terms is a non-preferred label for the other or both are non-preferred labels for the same third term. The remaining 3000 pairs consist of words that are equally distributed over related but not synonymous and completely unrelated pairs.

For each word a co-occurrence profile was constructed using a corpus of 47 Million words from texts from bestanden.officielebekendmakingen.nl, the site with official texts from the Dutch government. We used a Support Vector Machine to learn a classifier that distinguishes between pairs of synonymous and non-synonymous words. Using ten-fold cross validation we were able to classify 69% correct. When we include string similarity, the fraction of correctly classified pairs goes up to 75%. However, the positive effect of string similarity seems not to carry over to data sets, in which much more pairs of words have to be considered.

Keywords

CONTROLLED VOCABULARIES, DISTRIBUTIONAL SEMANTICS

Bibliographic Report 2014: A choice of relevant classification literature
Online Report 2014: A choice of nice web-features for subject cataloguing

Gerald Peichl[1] and Michael Franke[2]

[1] University Library, University of St. Gallen
[2] University Library, Freie Universität Berlin

Abstract. For several years LIS included a bibliographic report, presented by Bernd Lorenz, Member of the Scientific Program Committee of the Workshop on Classification and Subject Indexing in Library and Information Science. Essentially the bibliographic Report reviews library journals articles about case studies and classification theory.
Last year for the first time the bibliographic report implemented additionally an online report related to the digital world of subject cataloguing and presented interesting web-features. This should be continued in 2015 at the University of Essex, Colchester, UK.

profusion

Profusion is a data science consultancy specialising in helping organisations understand and improve their relationship with people

Founded in London in 2011, Profusion boasts a large data science team that has particular expertise in statistics and machine learning, with more than 60 staff spread between its offices in London and Dubai. The team's experience includes making sense of data for online and offline marketing, connected IoT devices (wearable technology and smart city related), marketing problems, logistics, customer service and employee wellbeing programmes

Profusion also works closely with academic institutions and is proud of its Knowledge Transfer Partnership with The University of Essex

Clients include
Sony Consumer Entertainment,
Dixons Carphone, The Kingfisher Group,
Luxottica, Hawes & Curtis and HSBC

http://www.profusion.com

profusion

Institute for Analytics
and Data Science