Book of Abstracts

Multidisciplinary Facets of Data Science

EC DA 2019

EUROPEAN CONFERENCE ON DATA ANALYSIS

Bayreuth|Germany

18 - 20 March

A Joint Data Science Conference of

Gesellschaft für Klassifikation (GfKl)
Data Science Society,

Section on Classification and Data Analysis (SKAD)
of the Polish Statistical Association,

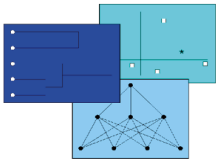
Classification and Data Analysis Group (CLADAG)
of the Italian Statistical Society,

Japanese Classification Society (JCS),

British Classification Society (BCS),

European Association for Data Science (EuADS)

organized by the **University of Bayreuth**



General Conference Chair and Program Chairs

Daniel Baier	University of Bayreuth, Germany
Berthold Lausen	University of Essex, United Kingdom
Angela Montanari	Università di Bologna, Italy

Program Committee

Bernd Bischl	LMU Munich, Germany
Ricarda Bouncken	University of Bayreuth, Germany
Paula Brito	University of Porto, Portugal
Ines Brusch	BTU Cottbus-Senftenberg, Germany
Michael Brusch	Anhalt University of Applied Science, Germany
Andrea Cerioli	Universita degli Studi di Padova, Italy
Theodoros Chatzipantelis	Aristotle University of Thessaloniki, Greece
Andreas Christmann	University of Bayreuth, Germany
Cristina Davino	University of Naples Federico II, Italy
Reinhold Decker	University of Bielefeld, Germany
Jose Dias	Instituto Universitario di Lisboa, Portugal
Peter A. Flach	University of Bristol, United Kingdom
Sugnet Gardner-Lubbe	University of Cape Town, South Africa
Claas Christian Germelmann	University of Bayreuth, Germany
Andreas Geyer-Schulz	KIT Karlsruhe, Germany
Henner Gimpel	University of Augsburg, Germany
Dominik Heider	University of Marburg, Germany
Christian Hennig	University College London, United Kingdom
Eyke Hüllermeier	University of Paderborn, Germany
Tadashi Imaizumi	Tama University, Japan
Stefan Jablonski	University of Bayreuth, Germany
Krzysztof Jajuga	Wroclaw University of Economics, Poland
Hans Kestler	University of Ulm, Germany
Friedrich Leisch	Universität für Bodenkultur Wien, Austria
Hermann Locarek-Junge	TU Dresden, Germany
Karsten Lübke	FOM University of Applied Sciences Dortmund, Germany
Fionn Murtagh	University of Huddersfield, United Kingdom
Atsuho Nakayama	Tokyo Metropolitan University, Japan
Akinori Okada	Tama University, Japan
Francesco Palumbo	University of Naples Federico II, Italy
Józef Pociecha	Cracow University of Economics, Poland
Alexandra Rese	University of Bayreuth, Germany
Roberto Rocci	Universita di Roma Tor Vergata, Italy
Adam Sagan	Cracow University of Economics, Poland
Ute Schmid	University of Bamberg, Germany
Lars Schmidt-Thieme	University of Hildesheim, Germany
Frank Scholze	KIT Karlsruhe, Germany
Winfried Steiner	TU Clausthal-Zellerfeld, Germany
Alfred Ultsch	University of Marburg, Germany
Nils Urbach	University of Bayreuth, Germany
Katrijn van Deun	Tilburg University, The Netherlands
Maurizio Vichi	Sapienza Universita di Roma, Italy
Claus Weihs	University of Dortmund, Germany
Adalbert Wilhelm	Jacobs University Bremen, Germany
Herbert Woratschek	University of Bayreuth, Germany

Content

Special Sessions were organized by the following colleagues:

- **Bioinformatics and Biostatistics**: organized by *Dominik Heider* (University of Marburg, Germany),
- **Complexity, Data Science and Statistics Through Visualization and Classification**: organized by *Carmela Iorio* and *Roberta Siciliano* (University of Naples Federico II, Italy),
- **Consumer Preferences and Marketing Analytics**: organized by *Reinhold Decker* (University of Bielefeld, Germany) and *Winfried Steiner* (TU Clausthal-Zellerfeld, Germany),
- **Data Analysis in Finance**: organized by *Krzysztof Jajuga* (Wroclaw University of Economics, Poland),
- **Data Analysis Models in Economics and Business**: organized by *Józef Pociecha* (Cracow University of Economics, Poland),
- **Interpretable Machine Learning**: organized by *Johannes Fürnkranz, Eneldo Loza Mencía* (both: TU Darmstadt, Germany), and *Ute Schmid* (University of Bamberg, Germany),
- **Statistical Learning**: organized by *Angela Montanari* (University of Bologna, Italy).

Many thanks for this wonderful engagement.

1 Plenary Talks

Big Data Science in Systems Medicine – A Disruptive View on Current Medicine

Jan Baumbach (TU München)

One major obstacle in current medicine and drug development is inherent in the way we define and approach diseases. We discuss the diagnostic and prognostic value of (multi-)omics panels. We have a closer look at breast cancer survival and treatment outcome, as case example, using gene expression panels – and we will discuss the current „best practice“ in the light of critical statistical considerations. In addition, we introduce computational approaches for network-based medicine. We discuss novel developments in graph-based machine learning using examples ranging from Huntington's disease mechanisms via Alzheimer's drug target discovery back to where we started, i.e. breast cancer treatment optimization – but now from a systems medicine point of view. We conclude that multiscale network medicine and modern artificial intelligence open new avenues to shape future medicine. We will also have a short glimpse on novel approaches for privacy-aware sensitive medical data sharing. We quickly introduce the concept of federated machine learning and blockchain-based consent management to build a Medical AI Store ensuring privacy by design and architecture.

Co-Clustering – A Survey on Models, Algorithms and Applications

Hans-Hermann Bock (RWTH Aachen University)

When a data set is represented by a two-dimensional two-mode matrix it can often be useful to cluster the rows and the columns of this matrix. Clustering can be performed either separately or simultaneously. This second approach is called two-way clustering, biclustering or co-clustering (e.g. Bock 2003, Govart, Nadif 2018, Schepers et al. 2017) and will be the subject of this paper. Major applications are known, e.g., from gene analysis, marketing, social sciences and text mining. There exists a large number of co-clustering methods, surveys are given, e.g., by Bock 2014, Govart, Nadif 2013, Pontes et al. 2015, Van Mechelen et al. 2004. While some methods are based on empirical argumentations only, others are derived from probabilistic models and/or proceed by optimizing a suitable bi-partitional clustering criterion. Different methods have been developed for data matrices with real-valued, binary or qualitative entries, eventually corresponding to different practical purposes. Methods are also distinguished by the type of domains that should be bi-partitioned: the two index sets of data vectors x_{ij}, the two coordinate spaces of two-dimensional data $\boldsymbol{x}_i = (x_{i1}, x_{i2})$, the row and column categories of a contingency table $\boldsymbol{N} = (n_{ij})$, etc. From a methodological point of view models can be classified into block models (yielding bi-partitions) or latent block models (using mixture models), combined with maximum likelihood approaches, eventually leading to generalized k-means or EM algorithms. Contingency tables are typically bi-clustered by maximizing information-type measures. The paper presents a structured survey on the most important co-clustering models, related numerical algorithms and various recent extensions, e.g., for multiway tables or functional data.

References

Bock, H.-H. (2003): Convexity-based clustering criteria: theory, algorithms, and applications in statistics. Statistical Methods & Applications 12, 293-317.

Bock, H.-H. (2014): Probabilistic two-way clustering approaches with emphasis on the maximum interaction criterion. Archives of Data Science 1 (1), 3-20.

Govaert, G., Nadif, M. (2013): Co-clustering: models, algorithms and applications. Wiley, 2013.

Govaert, G., Nadif, M. (2018): Mutual information, phi-squared and model-based co-clustering for contingency tables. Adv. in Data Analysis and Classification 12, 455-488.

Pontes, B., Giráldez, R., Aguilar-Ruiz, J.S. (2015): Biclustering on expression data: a review. Journal of Biomedical Informatics 57, 163-180.

Schepers, J., Bock, H.-H., Van Mechelen, I. (2017): Maximal interaction two-mode clustering. Journal of Classification 34 (1), 49-75.

Van Mechelen, I., Bock, H.-H., De Boeck. P. (2004): Two-mode clustering methods: a structured review. Statistical Methods in Medical Research 13, 363-394.

Analyzing Retail Market Basket Data by Unsupervised Machine Learning Methods

Harald Hruschka (University of Regensburg)

We compare the performance of several unsupervised machine learning methods, namely binary factor analysis, two topic models (latent Dirichlet allocation and the correlated topic model), the restricted Boltzmann machine and the deep belief net, on a retail market basket data set. We shortly present these methods and outline their estimation. Performance is measured by log likelihood values for a holdout data set. Binary factor analysis vastly outperforms topic models. Both the restricted Boltzmann machine and the deep belief net on the other hand attain a similar performance advantage over binary factor analysis. We also show how to interpret the relationships between the most important hidden variables and observed category purchases. To demonstrate managerial implications, we compute relative basket size increase due to promoting each category for the better performing models. Recommendations which product categories to promote based on the restricted Boltzmann machine and the deep belief net not only have lower uncertainty due to their better predictive performance, they are also more convincing than those derived by binary factor analysis, which leave out most categories with high purchase frequencies.

Relations Among K-means, Mixture of Distributions, and Fuzzy Clustering

Sadaaki Miyamoto (University of Tsukuba)

The method of K-means is best known among different clustering algorithms. Another class of algorithms using statistical models is EM algorithms, a typical of which is based on Gaussian Mixture Model (GMM). Yet another idea for clustering is to use fuzzy models. The best-known fuzzy method is fuzzy K-means (FKM), a fuzzy extension of the original K-means, which uses an alternate minimization of an objective function. Although it is obvious that fuzzy K-means are an extension of K-means, a relation between FKM and GMM is generally unknown, or they are considered to have no relations, since a statistical model and a fuzzy model are very different. In spite of this general understanding, we show that they have close methodological relations in this talk. The main point to be noted is that we have many generalizations of fuzzy K-means, among which some methods are equivalent or similar to statistical mixture models. Indeed, KL-information based clustering proposed by Ichihashi et al. (2000) is formally equivalent to the iterative solution of GMM using EM algorithm. Generalized method of fuzzy K-means includes two more variables for adjusting cluster sizes (or cluster volumes) and cluster covariances. It should be noted that they are not exactly those in statistical models with prior distributions and covariances in clusters but have similar functions. We hence overview the idea of these three classes of algorithms and show a number of methods of generalized fuzzy K-means which are equivalent or similar to statistical mixture of distributions. We moreover introduce what we call a classification function, whereby we uncover theoretical properties of clustering results not only in fuzzy methods but also those by statistical models. To conclude, what we emphasize in this talk is that the above three classes of methods are not independent but have close relations that should be noted for considering theoretical insights and also useful for further methodological development in data clustering.

Meta-Analysis and Symbolic Data Analysis

Masahiro Mizuta (Hokkaido University)

We point out an important relationship between meta-analysis and symbolic data analysis (SDA) and discuss methods to support meta-analysis with SDA. SDA was proposed by Professor Edwin Diday in 1988. In most of statistical analysis methods, the unit of analysis is an individual, but the unit of analysis in SDA is an object (or class, set, concept), which has a description, e.g. interval value, distributional value, multi values. Diday insisted that symbolic data is “any data taking care on the variation inside classes of standard observation.” A role of SDA is to analyze internal variations of the data or concepts. In the early stages of SDA’s development, we study interval valued data. Nowadays, there are so many types of symbolic data, including modal interval data, distributional data, multi-valued data. SDA is used in many fields.

Meta-Analysis, proposed in 1976 is a method to derive results with several studies and is widely used in many fields, including social science, medical science, drug approval. In medical research, meta-analysis has been placed at the top of the evidence pyramid (Oxford Center for Evidence-Based Medicine's levels of Evidence and Recommendation). There are two models used in meta-analysis, the fixed effect model and the random effects model. Heterogeneity in meta-analysis refers to the variation in study outcomes between studies.

Meta-analysis and SDA have been developed independently, but there is something in common between them. In a meta-analysis, the unit of analysis is an individual study rather than an individual study participant. This framework is almost the same as that of SDA. Meta-analysis is in the top of the evidence pyramid. However, even if formally conducting analytic meta-analysis of multiple studies, it is difficult to find a hidden structure. SDA is a powerful tool for exploratory meta-analysis. There are many tools in SDA community, and most of them are useful for exploratory meta-analysis; symbolic clustering is an effective tool for subgroup analysis in meta-analysis for example.

Archetypes, Prototypes and Other Types: The Current State and Prospects for Future Research

Francesco Palumbo (University of Naples Federico II)

Statistical and machine learning can significantly speed up human knowledge development, helping to determine the basic categories in a relatively short amount of time. Exploratory data analysis (EDA) can be considered the forefather of statistical learning; it relies on the mind's ability to learn from data and, in particular, it aims to summarize data-sets through a limited number of interpretable latent features or clusters offering cognitive geometric models to define categorizations. It can also be understood as the implementation of the human cognitive process extended to huge amounts of data: Big Data. But EDA alone cannot answer to the questions: "How many, and what are the categories to retain?" and „What are the observations that can represent a category better than others in human cognitive processes? ". The concept of categorization implies data summarization in a limited number of well-separated groups that must be maximally and internally homogeneous at the same time. This contribute aims to present an overview of the most recent literature on the archetypal analysis (AA) and its related methods as a statistical categorization approach. At the same time some most recent approaches in prototypes identification when dealing with large and huge data-sets are presented. In combination with consistent clustering approaches, AA helps to identify those observed or unobserved prototypes that satisfy Rosch's definition. Those small number of groups that are maximally homogeneous within the units belonging to the same group and maximally heterogeneous among groups, and which allow the development of the human knowledge by the relationships between prototypes and a new unknown object.

CB-SEM or PLS-SEM? Five Perspectives and Five Recommendations

Marko Sarstedt (Otto-von-Guericke University Magdeburg)

To estimate structural equation models, researchers can draw on two main approaches: Covariance-based structural equation modeling (CB-SEM) and partial least squares structural equation modeling (PLS-SEM). Concerns about the limitations of the different approaches might lead researchers to seek reassurance by comparing results across approaches. But should researchers expect the results from CB-SEM and PLS-SEM to agree, if the structure of the two models is otherwise the same? Differences in philosophy of science and different expectations about the research situation underlie five different perspectives on this question. We argue that the comparison of results from CB-SEM and PLS-SEM is misleading and misguided, capable of generating both false confidence and false concern. Instead of seeking confidence in the comparison of results across methods, which differ in their specific requirements, computational procedures, and imposed constraints on the model, researchers should focus on more fundamental aspects of research design. Based on our discussion, we derive recommendations for applied research using SEM.

Discrete Time-to-Event Analysis – Methods and Recent Developments

Matthias Schmid (University of Bonn)

Discrete time-to-event analysis comprises a set of statistical methods to analyze the duration time until an event of interest occurs. In contrast to classical methods for survival analysis, which typically assume the duration time to be continuous, discrete-time methods apply to situations where time is measured (or recorded) on a discrete time scale t = 1, 2, …. These situations are likely to occur in longitudinal studies with fixed follow-up times (e.g., epidemiological studies or panel studies) in which it is only known that events have happened (or not) between pairs of consecutive points in time. Unlike the classical approaches, discrete-time methods are able to handle potentially large numbers of tied duration times. Furthermore, estimation is greatly facilitated by the fact that the log-likelihood of a discrete time-to-event model is closely linked to the log-likelihood of a model with binary outcome. The talk will provide an overview of the most popular methods for discrete time-to-event modeling. In addition, it will cover recent developments on model validation, tree-based approaches, and methods for discrete-time competing risks analysis.

Composite Indicators and Rankings

Andrzej Sokolowski (Cracow University of Economics)

Rankings are popular both in scientific research and in everyday life. Many institutions rank different objects from the best to the worst. So, we have ranking of countries, universities, cities, hospitals, jobs, provinces, books, songs, actors, footballers etc. If such a ranking is based on just one variable measured in strong scale, then the task is trivial. The only thing we should decide is the direction – the bigger the better or the smaller the better. The problem is more complicated with multidimensional case, even with just two diagnostic variables. The aim of the paper is to present and discuss issues connected with the consecutive stages of the process of constructing composite indicators. They are as follows:

- Definition of the general criterion and possible sub-criteria
- Choosing diagnostic variables – initial and final list. Some unjustified statistical procedures
- Setting relations between general criterion and diagnostic variables (stimulants, destimulants, nominants). Merit and automatic identification methods
- Transformation of variables – normalization, standardization, other transformations
- Weighting systems
- Aggregation formulas
- Robustness and sensitivity

Linear ordering methods are sometimes divided into those with benchmark and without. It can be argued that every method has some benchmark – assumed or calculated from the data.

Short review of the literature will be given, and some example provided and critically discussed – such as Human Development Index or 200 Best Jobs in USA. Finally, new propositions will be presented such as – iterative procedures, method for choosing the best composite indicator for w given set of variables, Multidimensional Scaling approach, and non-linear ordering.

Zero-Sum Regression

Rainer Spang (University of Regensburg)

The performance of Machine Learning algorithms can critically depend on the preprocessing of the input data. This is the case for many types of molecular data used in biomedicine including transcriptomics, proteomics and metabolomics profiles. Biases from experimentation and measurement can strongly affect performance, even after state-of-the-art data normalization protocols where applied. Some Machine Learning algorithms propagate these biases more than others. For supervised learning problems we propose “zero-sum“ regression as a tool designed to be fairly robust against frequent biases. This is achieved by using generalized linear models combined with the zero-sum constraint, which causes the resulting models to be scale invariant.

Using Microblogging to Predict IPO Success by Means of Apache Spark Big Data Analytics

Dirk Van den Poel (Ghent University)

This research investigates the influence of Twitter social media messages on the success/failure of Initial Public Offerings (IPOs). We analyze the (potential) impact of the number of tweets, their sentiment, and many other features on (1) the difference between the IPO price and the closing price of the stock at their first day of trading, and (2) the difference between the closing price on the first day of trading and the closing price after three months of trading.

Tutorials

Partial Least Squares Structural Equation Modeling (PLS-SEM)

Marko Sarstedt (Otto-von-Guericke University Magdeburg)

Partial least squares structural equation modeling (PLS-SEM) has recently received considerable attention in a variety of disciplines, including marketing, strategic management, management information systems, and many more.PLS is a composite-based approach to SEM, which aims at maximizing the explained variance of dependent constructs in the path model. Compared to other SEM techniques, PLS allows researchers to estimate very complex models with many constructs and indicator variables.Furthermore, PLS-SEM allows to estimate reflective and formative constructs and generally offers much flexibility in terms of data requirements.This full-day workshop introduces participants to the state-of-the-art of PLS-SEM using the SmartPLS 3 software. After a brief introduction to the basic principles of structural equation modeling, participants will learn the foundations of PLS-SEM and how to apply the method by means of the SmartPLS software. The workshop will cover various aspects related to the evaluation of measurement and structural model results. For this purpose, the instructor will make use of several examples and exercises.

Sponsors of ECDA2019

Springer-Verlag GmbH, Heidelberg, Germany

Springer is a leading global scientific, technical and medical portfolio, providing researchers in academia, scientific institutions and corporate R&D departments with quality content through innovative information, products and services. Address: Tiergartenstrasse 17, 69121 Heidelberg, Germany.

Betriebswirtschaftliches Forschungszentrum für Fragen der mittelständischen Wirtschaft e. V. (BF/M), Bayreuth, Germany

BF/M is a non-profit association founded in 1979 with the goal of interlinking science and industry by transferring business research results into companies and conducting empirical research. The main focus is clearly on small and medium-sized enterprises. Address: Mainstraße 5, 95444 Bayreuth.

Fraunhofer-Institut für Angewandte Informationstechnik FIT Projektgruppe Wirtschaftsinformatik, Schloss Augustin, Augsburg and Bayreuth, Germany

The Research Center Finance & Information Management (FIM) and the Project Group Business & Information Systems Engineering of the Fraunhofer Institute for Applied Information Technology (FIT) have established themselves at the Universities of Augsburg and Bayreuth as renowned research centers at the interface of financial and information management. In innovative projects, they have been supporting numerous companies from financial service and IT sectors as well as industrial companies for years. The success is based on the close cooperation with eight renowned chairs in the fields of banking/finance, financial mathematics and business informatics as well as the synergetic interaction of research, teaching and practical projects. Address: 53754 Sankt Augustin, Germany.

Wissenschaftscampus E-Commerce, Burgkunstadt, Germany

The science campus E-Commerce is a joint project of the Upper Franconian economy and science and is located in the premises of a former shoe factory: 7,000 square meters of usable space, which has been renovated and made usable by “Baur-Versand” for several million Euros for this purpose. Experts from trade and academics from the participating universities and the Fraunhofer Institute for Applied Information Technology (FIT) will jointly tackle problems and develop solutions. The aim is to remain internationally competitive by means of innovative ideas and mutual exchange with the mail order business, which is traditionally strong in Germany and especially Bavaria. Address: Bahnhofstraße 10, 96222 Burgkunstadt, Germany.

BAUR-Group, Burgkunstadt, Germany

The BAUR-Group focuses on online business and trade-related services. Core business of the BAUR Group is BAUR with its online shop www.baur.de. The business area “Online Trade” is completed by “I’m walking”, an online shop for shoes and clothes. Additionally, the Austrian subsidiary “UNITO” is responsible for brands like “Universal”, “Otto Österreich”, “Quelle” (Germany, Austria and Switzerland) as well as “Ackermann” and “Alpenwelt”. The business area “Services” contains “BFS”, specialized in the online and mail order value chain, logistics provider “2. HTS”, the online marketing specialist “octobo” and the e-commerce service provider “empiriecom”. BAUR is an important member of the Otto Group. Address: Bahnhofstraße 10, 96222 Burgkunstadt, Germany.

CELONIS, Munich, Germany

Celonis was founded in Munich in 2011 and is one of the world’s leading providers of Process Mining Software, as well as one of the fastest growing technology ventures in Germany. Celonis received a 50-million-dollar investment in their Series B. The well-known US investors “Accel” und “83North” value the start-up from Munich at one billion dollars. This puts Celonis into the club of so-called “Unicorns”, companies with an estimated value of at least 1 billion dollars. Address: Theresienstr. 6, 80333 München, Germany.

2 Special Sessions and Regular Sessions

Biostatistics and Bioinformatics 1

An Effects Size Measure Based on Probabilities and Covering Data Sets where Cohen's D is not Applicable

Alfred Ultsch[1], Jörn Lötsch[2] ([1]: Philipps University of Marburg; [2]: Goethe-University)

Calculation of the extent of group differences or of effect sizes is common throughout quantitative science (Cohen, 1988). Many different measures of effect sizes have been proposed and their use in biomedical research continues to be an active research topic until now (Monsarrat and Vergnes, 2017). One of the effect size measures most frequently used in biomedical research is Cohen's d, which calculates the effect size as the difference in group means divided by the joined standard deviation (Cohen, 1960). This raises several limitations (Grice and Barrett, 2014), for example making it obviously unsuitable when the variance approaches zero such in cases where the measurement happens to obtain only the same single value in all cases.

Data can usually be represented as probabilities, such as the expression of genes in biological samples where, for the comparison of different diseases affecting many genes, Cohen's d leads to implausible results. As an alternative measure of effect sizes, "Separation" (S), is proposed. Separation is obtained as the difference of the means of two data subsets (delta) with the probability that the two subsets have no common elements as a weighting factor.

For parametric distributions such as the normal distribution, this effect size can be determined analytically. For nonparametric distributions, the intersection of the probability densities of the two groups can be easily obtained iteratively. Separation possesses a finite range spanning the interval between 0 and delta. Separation is largest (S=delta), if the two subsets have no elements in common. Separation approaches zero if either there is no difference in means or the two subsets share a large portion of common values.

As Separation provides plausible values where Cohen's d approaches infinity or is not defined, it is particularly suitable as an effect size measure for automated analyses of large and high-dimensional data sets (Ultsch and Lötsch, 2017). Using artificial data and bioinformatics experiments, it will be presented how the effect sizes quantified using Separation compare with other effect size measures, with a focus of a direct comparison with Cohen's d. In addition, real biomedical data will be shown where Separation outperforms Cohen's d in providing biologically plausible estimates of effects.

References

Cohen, J. (1960): A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20.

Cohen, J. (1988): Statistical Power Analysis for the Behavioral Sciences (New York: Routledge).

Grice, J.W., Barrett, P.T. (2014): A Note on Cohen's Overlapping Proportions of Normal Distributions. Psychol Rep 115, 741-747.

Monsarrat, P., Vergnes, J.-N. (2017): The Intriguing Evolution of Effect Sizes in Biomedical Research over Time: Smaller but more Often Statistically Significant. GigaScience 7, 1-10.

Ultsch, A., Lötsch, J. (2017): Machine-learned Cluster Identification in High-dimensional Data. J Biomed Inform 66, 95-104.

Incorporating Foreign Classes in Feature Selection Processes

Robin Szekely[1], Ludwig Lausser[1], Hans A. Kestler[1] ([1]: Ulm University)

Datasets generated by molecular high-throughput experiments are mainly characterized by their extremely high dimensionalities and their relatively small sample sizes ($n \gg m$). While feature profiles typically comprise several thousand molecular markers, sample sets rarely contain more than a few dozens of samples, due to ethical and economic reasons. Often multi-centric studies are required to achieve a reasonable number of samples. General strategies to cope with this setting can be the use of sparse classification models or the acquisition of additional data resources.

In this work, we investigate the possibility of supplementing samples of an original binary classification task by samples of foreign but related classes (Lausser et al. 2018a, 2018b). We assume such extra classes to occur in multi-class datasets that were collected for a common research question. More specifically, we utilize the foreign samples for feature selection. We characterize the quality of individual features in indirect experiments between original and foreign classes. We show that certain constellations directly imply a high-quality measure for the original classification task. We evaluate this indirect strategy empirically in cross-validation experiments on multi-class microarray and RNA-Sequencing datasets. In our evaluations this strategy outperformed the direct (original) feature selection in up to 67.78% of all experiments.

References

Lausser, L., Szekely, R., Kessler, V., Schwenker, F., Kestler, H. (2018a): Selecting Features from Foreign Classes. In: Pancioni, L., Schwenker, F., Trentin, E. (eds.) Artificial Neural Networks in Pattern Recognition. pp. 66–77. Springer International Publishing, Cham.

Lausser, L., Szekely, R., Schirra, L.R., Kestler, H. (2018b): The influence of multi-class feature selection on the prediction of diagnostic phenotypes. Neural Processing Letters 48(2), 863–880.

A Comparative Study on Peptide Encodings for Biomedical Classification

Sebastian Spänig[1], Dominik Heider[1] ([1]: University of Marburg)

The key mechanism of many diseases can be encountered on protein level, including cancer, HIV, and neurodegenerative diseases. Moreover, cases involving multi-resistant pathogens have increased to a threatening level. Thus, in the last decades computational biologists developed encodings and machine learning models for automated and accurate classification of peptides and proteins. By engaging research with this respect, even the pharmaceutical industry acknowledges the potential yield from gained insights to medical treatments (Mahlapuu et al., 2016). Since these models require a fixed-length, numerical input, an essential part of the classification pipeline involves the engineering of representative encodings of the peptide sequence. To this end, several sequence-based encodings have been introduced so far as part of antimicrobial peptides classification studies, including sparse, compositional, and physicochemical encodings (Veltri et al., 2017). In contrast, others explored informative descriptors of the secondary structure, mainly to tackle structure related issues, e.g., the prediction of HIV-1 co-receptor tropism (Löchel et al., 2018).However, an appropriate literature search results in a variety of promising encodings, leading to the question, which encoding is suitable for a particular classification task and which of the encodings performs better on a particular data set, respectively. Consequently, we compared the performance of state-of-the-art amino acid encodings on independent, biomedical data sets. Due to their high impact, the data encompasses, e.g., antimicrobial peptides as well as HIV-1 co-receptor tropism, but also biological data, such as protein-protein interaction and cell-penetrating peptides, to examine the generalization capabilities of the proposed encodings. By collecting and testing available encodings, we incur the time-consuming literature search and, in particular, the feature selection. This will aid computational biologist to focus on the actual results instead of occupying themself with choosing an appropriate encoding beforehand. Finally, the overall machine learning part is simplified. In conclusion, our findings will significantly promote classification accuracy specifically for active peptides as well as proteins in general.

References

Löchel, H. F., Riemenschneider, M., Frishman, D., Heider, D. (2018). SCOTCH: Subtype A Coreceptor Tropism Classification in HIV-1. Bioinformatics 34, 2575–2580.

Mahlapuu, M., Håkansson, J., Ringstad, L., Björn, C. (2016). Antimicrobial Peptides: An Emerging Category of Therapeutic Agents. Front. Cell. Infect. Microbiol. 6, 194.

Veltri, D., Kamath, U., and Shehu, A. (2017). Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming. IEEE/ACM Trans. Comput. Biol. Bioinform. 14, 300–313.

Biostatistics

Semantic Multi-Class Classifier Systems in Precision Medicine

Lea Siegle[1], Ludwig Lausser[1], Hans A. Kestler[1] ([1]: Ulm University)

Molecular high-throughput technologies have made the measuring of vast sets of markers possible. The downside: this high-dimensionality makes it nearly impossible for the datasets to be interpreted by human experts. Machine learning techniques like classification are required for the evaluation of these high-dimensional datasets, be it in diagnostics (for personalised medicine) or in sciences (to create new hypotheses). However, the patterns used to classify samples are often as uninterpretable as the data itself. We employ a new technique named ”Semantic multi-class classifier system” (SMCCS) to train biologically meaningful classifiers and to increase model interpretability. SMCCS combine (non-data-driven) semantic feature selection (SFS) on the basis of established vocabularies (e.g. GO and KEGG) with a multi-class classification (MCC) method (Lausser et al. 2016). We tested two types of SFS on the well-known one-against-one (OaO) and one-against-all (OaA) MCCs (Lausser at al. 2018: In the first strategy, a common set of features selected and provided to all base classifiers. In the second one, features are selected individually for each base classifier. The four combinations are evaluated on a 4-classed breast cancer dataset and characterised by their accuracy and feature selection stability. We additionally provide evidence on the selected semantic terms. Interestingly, support can even be found for terms selected for individual two-class comparisons.

References

Lausser, L., Schmid, F., Platzer, M., Sillanpää, M.J., Kestler, H.A. (2016): Semantic Multi-classifier Systems for the Analysis of Gene Expression Profiles. Arch. Data Sci. Ser. A (Online First) 1(1).

Lausser, L., Szekely, R., Schirra, L.R., Kestler, H. (2018): The Influence of Multi-class Feature Selection on the Prediction of Diagnostic Phenotypes. Neural Processing Letters 48(2), 863–880.

Attributable Fraction in Multifactorial Situations – Concepts and Properties of Commonly used Methods

Carolin Malsch (University of Würzburg)

Attributable fractions (AF) are used to estimate the amount of an outcome associated with exposition to a risk factor. Levin’s formula for AF, first introduced in 1953, is a function of the risk ratio and the risk factor prevalence. The formula allows consideration of only one risk factor, but adjusted risk ratios are often entered to obtain adjusted AF estimates. However, this strategy is criticized due to non-additivity of AF when applied to multiple risk factors successively, resulting in problems of interpretability. Various approaches are available to estimate the AF considering multiple expositions simultaneously, including sequential, average sequential and proportional AF. These approaches ensure additivity, but magnitudes of AF can differ noticeably, which again leads to uncertainty regarding interpretation.

A possible strategy to define and calculate AF in case of multiple risk factors consists in the disjoint decomposition of the population with regard to these factors, i.e. the resulting strata are associated with the appearance of exactly one subset of all risk factors, respectively. The partial AF are calculated for each stratum individually. AF for each risk factor is subsequently calculated from strata where the risk factor occurs alone or together with other risk factors. In strata with multiple risk factors, the partial AF are systematically divided into fragments and are subsequently assigned to the risk factors of interest. The sequential, average sequential and proportional approaches can be described in the context of partial AF and come along with individual partitioning and assignment strategies. This causes differences in the magnitudes of the AF.

This study illustrates the different assignment strategies that come along with the sequential, average sequential and proportional approaches. A matrix-based presentation of the partitioning and assignment scheme is developed, and the bias from adopting an assignment strategy that does not match a risk factors’ role in the disease mechanism is quantified using computer simulation. The properties of the different assignment strategies are investigated regarding symmetry, marginal rationality, internal marginal rationality and additivity and their interpretation in epidemiologic research is addressed. This study aims to facilitate the understanding of different calculation strategies in order to

foster a reasonable and prudent interpretation of study results to improve communication on AF in future research.

Improving Classification of Single-Cell Data Using Consensus Clustering Based on Different Laplacian Eigenmaps

Cornelia Fütterer[1], Thomas Augustin[1], Christiane Fuchs[2] ([1]: LMU München; [2]: Helmholtz Zentrum München and Universität Bielefeld)

The analysis of single-cell genomic data is of high importance in cancer research. It challenges data scientists especially in the context of personalized medicine. The single-cell RNA sequencing (scRNA-seq) technique, introduced by Tang et al. (2009, Nature Methods), enables access to gene expression of single cells, which allows classification of single cells based on highly resolved genomic profiles. While sequencing technologies advance, more and more clustering approaches are developed to be applied directly to single cells as main entity. In their comparison paper on classifying single-cell data, Kiselev et al. (2017, Nature Methods) conclude that their single-cell consensus clustering approach performs best in many settings. On publicly available data sets of the Sanger Institute, SC3 turned out to be more robust and accurate than the methods t-SNE + k-means, pcaReduce, SNN-Cliq, SINCERA and SEURAT.The focus of the SC3 method lies on aggregating clustering results based on different dimension reduction techniques such as the principal component analysis (PCA) or Laplacian eigenmaps. We challenge the PCA-based classification as part of the SC3 method with regard to the underlying structure of single-cell data and show that SC3 can be improved by applying the consensus clustering based on Laplacian eigenmaps.

In simulation studies, we observed that the performance of SC3 highly depends on the underlying (usually unknown) cluster sizes. The performance of SC3 is good for equally sized partitions of single cells in each group whereas it declines rapidly in case of more and more unequal partitions. Targeting the conflicting results of individual clustering, we propose an evaluation of the obtained consensus score. This allows to perform a final clustering on single-cells which promises to be trustworthy. Thus, especially single cells showing a high accordance of the individual clustering results are selected for the final clustering. All in all, our adapted consensus clustering approach is comparable or even substantially better than SC3 with regard to accuracy.

Biostatistics and Bioinformatics 2

A Novel Privacy-preserving Software Framework for Federated Machine Learning in Clinical Settings

Marta Lemanczyk[1], Dominik Heider[1] ([1]: Philipps-Universität Marburg)

Sharing of patient data between cooperation partners is an important component of biomedical research as prediction models can be improved by using a larger amount of data. Due to privacy regulations, such as the General Data Protection Regulation (GDPR) by the EU (2018), sharing of sensitive biomedical data of patients is not possible without permissions. Thus, keeping data centralized and making it only available for cooperation partners is not an option since it would increase the risk of cyber-attacks. Furthermore, anonymization of individual data is in most cases not sufficient. It has been shown that it is possible to trace back the identity of one patient by reproducing it from data (Sweeney 2002). However, collaborations between institutions require a secure solution for sharing information gained from patient data. Federated Machine Learning (FML) enables information exchange between users by sharing parameters and meta-information of machine learning models, without violating the privacy requirements. Each user stores their own data locally and has no access to other user's data. Without sharing any sensitive information, a user can transfer parameters and meta-information of a locally trained model and can combine it with the other models towards a more generalized model. In this study, we focused on the development of a secure privacy-preserving software framework, which can be used for developing federated machine learning models. Moreover, we analyzed multi-party model combinations. Depending on data size and class balance, different

combinations of models were evaluated, with a particular focus on logistic regression and random forests (Breiman 2001), due to the fact that these models are well-accepted in the biomedical community. For training and evaluation of the models, two real-world data sets were used, namely diabetes prediction (Kälsch et al. 2015) and prediction of antimicrobial efficacy (Pirtskhalava et al. 2015). Additionally, the concept of differential privacy (Pathak et al. 2010) is applied on the models to prevent reproduction of patient identities.

References

Sweeney L. (2002): k-anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570.

Breiman L. (2001): Random Forests. Machine Learning, 45(1): 5-32.

Kälsch J., et al. (2015): Normal Liver Enzymes are Correlated with Severity of Metabolic Syndrome in a Large Population Based Cohort. Scientific reports, 5: 13058.

Pirtskhalava M. et al. (2015): DBAASP v.2: An Enhanced Database of Structure and Antimicrobial/Cytotoxic Activity of Natural and Synthetic Peptides. Nucleic acids research 44(D1): D1104-D1112.

Pathak M., Shantanu R., and Bhiksha R. (2010): Multiparty Differential Privacy via Aggregation of Locally Trained Classifiers. Advances in Neural Information Processing Systems 23:1876-1884.

Inverse Document Frequency as the Distance for Clustering Gene Sets

Michael Christoph Thrun[1], Catharina Lippmann[2], Alfred Ultsch[1] ([1]: University of Marburg; [2]: Fraunhofer Institute of Molecular Biology and Applied Ecology)

The analysis of gene expression data plays a key role in disease diagnosis. However, such analysis is rather complex due to the high-dimensional gene space (Acharya et al. 2017). Semantically related genes can be grouped using biological knowledge contained in the Gene Ontology (GO) (Ashburner et al. 2000). This work proposes the usage of the inverse document frequency(idf) (Sparck 1972) as a distance measure between two genes. For each gene, the idf logarithmically counts the number of GO terms in which any gene of the set is annotated, divided by be the number of GO terms to which the specific gene is annotated.

The inverse document frequency applies concepts of Information Retrieval (IR) to gene similarity based on the GO. IR distances have been developed to find and compare written documents, e.g. papers or books, which are described by a large vector of descriptors (terms). GO terms are assumed to represent documents in the IR sense and the description terms are the genes of the set.

The absolute difference between two counts is the distance measure between two genes. Contrary to the PAM clustering (Acharya et al. 2017), the Databionic Swarm (DBS) is applied (Thrun 2018). DBS is able to find cluster structures of any shape instead of being restricted to spherical cluster structures. Cluster analysis is performed on gene sets causally associated with pain and the chronification of pain (Ultsch et al. 2016), hearing loss (Gene Testing Registry 2018), cancer (Futreal et al. 2018) and drug addiction (Li et al. 2008) showing a distinctive cluster structure. Using overrepresentation analysis (Lippmann et al. 2018), the knowledge regarding pain processes and the chronification of pain shown in (Lötsch et al. 2013) was reproduced. The cluster structures found in the gene sets of the other three diseases are under investigation. This work indicates the successful application of Information Retrieval distance measures for clustering gene sets based on knowledge bases.

References

Acharya, S., S. Saha, and N. Nikhil (2017): Unsupervised Gene Selection Using Biological Knowledge: Application in Sample Clustering. BMC Bioinformatics 18: p. 513.

Ashburner, M., et al. (2000): Gene Ontology: Tool for the Unification of Biology. Gene Ontology Consortium. Nature genetics 25(1): p. 25-29.

Sparck Jones, K. (1072): A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of documentation 28(1): p. 11-21.

Thrun, M.C. (2018): Projection Based Clustering Through Self-Organization and Swarm Intelligence. 2018, Heidelberg: Springer.

Ultsch, A., et al. (2016): A Data Science Approach to Candidate Gene Selection of Pain Regarded as a Process of Learning and Neural Plasticity. Pain.

Gene Testing Registry (2018). OtoGenome Test for Hearing Loss 2018, Retrieved from: https://www.ncbi.nlm.nih.gov/gtr/tests/509148/.

Futreal, P.A., et al. (2018): A Census of Human Cancer Genes. Nature reviews cancer 4(3): p. 177.

Li, C.-Y., X. Mao, and L. Wei (2008): Genes and (Common) Pathways Underlying Drug Addiction. PLoS computational biology 4(1).
Lippmann, C., et al. (2018): Computational Functional Genomics-based Approaches in Analgesic Drug Discovery and Repurposing. Pharmacogenomics 19(9): p. 783-797.
Lötsch, J., et al. (2013): Functional Genomics of Pain in Analgesic Drug Development and Therapy. Pharmacology & Therapeutics 139(1): p. 60-70.

Extracting Ordinal Substructures from Multi-Categorial Datasets

Lisa M. Schäfer[1], *Ludwig Lausser*[1], *Hans A. Kestler*[1] ([1]: Ulm University)

Various datasets comprise multiple classes of the same domain. These multi-categorial datasets can be analysed on different levels. Hereby, not only the characteristics of the categories themselves but also the inter-class relations might be of interest. Often relations are described on a semantic level, but it is unknown whether the assumption holds for the analysed feature representation. Eliciting sequences of these relations from data can lead to the confirmation or rejection of existing order hypothesis but also the findings of new data inherent processes that were not observable before. As multi-class classifier systems characterize concepts by discriminative patterns, they allow hypotheses on the class inherent properties and their mutual dependencies. In this work, we provide an explorative data analysis procedure for detection of ordinal class structures in multi-categorial data collections. Our approach is based on classification experiments with ordinal classifier cascades that are applied exhaustively to all subsets of classes (Lattke et al. 2015). These cascades are constrained by a predefined class order. If this order is not reflected in the feature representation, the corresponding candidate cascade shows a diminished generalization performance. As the performance of a candidate cascade depends on the performance of its subcascades, we focus on the detection of the longest cascades that pass a predefined quality threshold on the minimal class-wise sensitivity. In our experiments, we show that long ordinal subcascades can be detected in data collections of different research fields. We apply our exhaustive screening approach to datasets comprising up to 26 classes, ranging between 16 and 54000 features, indicating that the method is applicable to diverse data collections. Most of the final candidate cascades provide a sequence of ordinal decision regions that not only allow an accurate classification of the cascade classes but also show a non-random assignment of the remaining classes. Data inherent information revealed by our method allows for data interpretation on a class neighbouring level and might generate new hypotheses about class underlying dependencies.

References

Lattke, R., Lausser, L., Müssel, C., Kestler, H.A. (2015): Detecting Ordinal Class Structures. In: Schwenker, F., Roli, F., Kittler, J. (eds.) Multiple Classifier Systems (MCS). LNCS, vol. 9132, pp. 100–111. Springer (2015)

Clustering 1

A Multivariate Characterisation of Clustering Quality and Validity

Christian Hennig (University of Bologna)

I have argued (Hennig 2015) that there are various different aims of cluster analysis, for which different clusterings may be optimal even on the same dataset. I present a collection of indexes that measure different aspects of interest in clustering (such as within-cluster homogeneity, between-cluster separation, representation of the underlying distance structure by the clustering, correspondence to high density regions, good representation of clusters by centroids etc.). There are a number of cluster validity indexes proposed in the literature (Valkidi et al. 2015). Most if not all of them attempt to give a one-dimensional assessment of the overall quality of a clustering, which does not provide insight into how the trade-off between the specific characteristics that could be potentially desirable plays out. The proposed collection of indexes can be used for various aims such as comparing different clusterings on a dataset including estimating the number of clusters or giving an empirical multivariate

characterisation of the behaviour of popular clustering methods in order to assist users to choose an appropriate method in a given application. I may also touch on some theoretical issues.

References

Hennig, C. (2015): What are the true Clusters? Pattern Recognition Letters 64, 53-62.

Hennig, C. (2017): Cluster Validation by Measurement of Clustering Characteristics Relevant to the User. arXiv:1703.09282.

Valkidi, M., Vazirgiannis, M. and Hennig, C. (2015): Method-Independent Indices for Cluster Validation and Estimating the Number of Clusters. In: Hennig C., Meila M., Murtagh F. and Rocci R. (eds.) Handbook of Cluster Analysis. Chapman and Hall/CRC, USA.

On a Class of Minimum Distance Aggregation Clustering Algorithms Based on Bi-Partial Objective Function

Jan W. Owsiński (Polish Academy of Sciences)

The paper presents a class of hierarchical merger algorithms, based on merging of the closest clusters, which, while being analogous to the known class, based on the Lance-Williams formula and its extensions, offers a natural stopping criterion, indicating the sub-optimal solution to the clustering problem.

As it is well known, the classical aggregation algorithms work for a set ***X*** of *n* objects or observations, characterised by vectors $x_i = (x_{i1}, \dots, x_{ik}, \dots, x_{im})$ of values of m variables, in such a way that starting from the partition P_1 into n clusters, equivalent to individual objects, at each step in the matrix of distances d_{ij} between clusters the smallest element is found. The corresponding two clusters are merged, and the distance matrix is updated for the newly formed cluster. The known algorithms differ exactly in the latter point, i.e. the way inter-cluster distances are updated for the newly formed cluster. Otherwise, the mergers proceed down to the partition P_n, composed of just one cluster. The proper solutions to the clustering problem are then based on some “external” criteria, applied to the results obtained in the course of the procedure (values of minimum distance, characteristics of clusters, of the entire partition, etc.).

Although there are some works towards the association of certain objective functions with some of the particular progressive merger algorithms, there is no comprehensive methodology that would allow for (1) formulation of such an objective function; (2) derivation of the stopping criterion therefrom. The paper shows, how, on the basis of the bi-partial precepts, i.e. the objective function, which expresses both intra-cluster cohesion and inter-cluster dissimilarity, it is possible to construct a general scheme of an aggregation algorithm, which includes the stopping rule. Several examples of the algorithms are shown, related to different formulations of the general bi-partial objective function.

Cluster Validation for Mixed-Type Data

Rabea Aschenbruck[1], Gero Szepannek[1] ([1]: Hochschule Stralsund)

For cluster analysis based on mixed-type data (i.e. data consisting of numerical and categorical variables), only a few cluster methods are available, including the extension of the k-Means algorithm (Huang, 1997). This so-called k-prototype algorithm is implemented in the R-package "clustMixType" (Szepannek, 2018). It is known that the selection of a suitable number of clusters k is particularly crucial in partitioning cluster procedures. Many implementations of cluster validation indices in R are not suitable for mixed-type data. This presentation examines the transferability of validation indices, such as e.g. Gamma index, average silhouette width or Dunn index, to mixed-type data. The selection of the indices considered is mainly based on the paper by Milligan and Cooper (1985). Furthermore, the package "clustMixType" is improved by these indices and their application is demonstrated. Finally, the behaviour of the adapted indices is tested by a short simulation study while using different data situations.

References

Huang, Z. (1997), „Clustering Large Data Sets with Mixed Numeric and Categorical Values “, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore: World Scientific, pp. 21–34.

Milligan, G. W., Cooper, M. C. (1985). „An Examination of Procedures for Determining the Number of Clusters in a Data Set “, Psychometrika, 50(2), pp. 159-179.
Szepannek, G. (2018). clustMixType: k-Prototypes Clustering for Mixed Variable-Type Data. R package version 0.1-36, https://CRAN.R-project.org/package=clustMixType.

Clustering 2

Bounds for the Number of Partitions of a Graph

Fabian Ball[1], Andreas Geyer-Schulz[1] ([1]: Karlsruhe Institute of Technology)

We know from combinatorics that the number of partitions of a set of n elements is given by the Bell number $B(n)$. This number increases rapidly, even for smaller n. The Bell number can be written as the sum of the Stirling numbers of the second kind, $S(n,k)$, over all $k = 1$ to $n - 1$. $S(n,k)$ determines the number of partitions of an n −element set into k non-empty parts. However, in graph clustering we are interested in a 'good' partition, which, independent of the used algorithm, normally depends on the adjacency of the graph. This means in particular that each cluster of a graph partition contains only nodes which are connected. The constraint highly restricts the number of possible partitions.

Our contribution provides examples of the possible partitions of very small graphs $(n \leq 13\, nodes)$. Furthermore, we propose an estimate upper bound of the actual number of partitions, which is based on the number of edges, and present the exact number of possible partitions of trees.

Trees are the sparsest possible connected graphs and, therefore, the number of tree partitions is a lower bound.

This lower bound is 2^m and thus still exponential.

A Bipartite-Graph Based Approach for Split-Merge Evolutionary Clustering

Veselka Boeva[1], Milena Angelova[2], Elena Tsiporkova[3] ([1]: Blekinge Institute of Technology; [2]: Technical University of Sofia; [3]: The Collective Center for the Belgian technological industry)

In many practical applications such as healthcare decision support systems the information available in the system database is periodically updated by gathering new data. The available data elements, e.g. patient profiles, are usually partitioned into groups of individuals with similar clinical conditions in order to facilitate the diagnosis and initial patient treatment. It is becoming impractical to re-cluster this large volume of available information when a new portion of data arrives. Profiling of users with wearable applications with the purpose to provide personalized recommendations is another example. As more users get involved one needs to re-cluster the initial clusters and also assign new incoming users to the existing clusters. In the context of profiling of machines (industrial assets) for the purpose of condition monitoring the existing original clusters can become outdated caused by aging of the machines and degradation of performance due to influence of changing external factors. This outdating of models is in fact a concept drift and requires that the clustering techniques, used for deriving the original machine profiles, can deal with such a concept drift and enable reliable and scalable model update.

We propose a split-merge evolutionary clustering approach that is suited for applications affected by concept drift. It is a bipartite correlation clustering technique that can be used to adapt the existing clustering solution to clustering of newly arrived data elements. The proposed technique is supposed to provide the flexibility to compute clusters on a new portion of data collected over a defined time period and to update the existing clustering solution by the computed new one. Such an updating clustering should better reflect the current characteristics of the data by being able to examine clusters occurring in the considered time period and eventually capture interesting trends in the area. For example, some clusters will be updated by merging with ones from newly constructed clustering while others will be transformed by splitting their elements among several new clusters. The proposed clustering algorithm is evaluated and compared to another bipartite correlation clustering technique (PivotBiCluster) on two different case studies: expertise retrieval and patient profiling in healthcare.

The split-merge evolutionary clustering algorithm has shown better performance than the PivotBi-Cluster in most of the studied experimental scenarios. For future work, we plan to pursue further evaluation and comparison of our technique with other clustering approaches suited for concept drift scenarios in different application domains.

A Study on the Improvement of the Overlapping Cluster Analysis

Satoru Yokoyama (Aoyama Gakuin University)

Several overlapping cluster analysis or soft clustering models have been suggested, and various kinds of data were analysed by these models.

ADCLUS model suggested by Shepard and Arabie (1979) and related models are one of the overlapping cluster analyses, and these models can be adapted to proximity data.

ADCLUS is analyzed for one-mode two-way data, INDCLUS (Carroll and Arabie, 1983) is for two-mode three-way data, and GENNCLUS (DeSarbo, 1982) is the generalized model for two-way data. Moreover, the author and co-researchers suggested a one-mode three-way model (Yokoyama et al., 2009).

These models are based on MAPCLUS (Arabie and Carroll, 1980) which is the most famous and general algorithm for overlapping cluster analysis. This algorithm consists of an alternating least squares approach and a combinatorial optimization. Therefore, it takes a lot of time to obtain the result and it is likely to obtain local results when the large number of objects such as marketing data is analyzed.

In the present study, the author attempts to improve this problem.

Clustering Ranking Data via Copulas

Marta Nai Ruscone (LIUC Università Cattaneo)

Clustering of ranking data aims at the identification of groups of subjects with a homogenous, common, preference behavior. Ranking data occurs when a number of subjects are asked to rank a list of objects according to their personal preference order. The input in cluster analysis is a distance matrix, whose elements measure the distances between rankings of two subjects. The choice of the distance dramatically affects the final result and therefore the computation of an appropriate distance matrix is an issue. Several distance measures have been proposed for ranking data (Alvo & Yu, 2014). The most important are the Kendall's t, Spearman's r and Cayley distances (Critchlow et al., 1991; Mallows, 1957; Spearman, 1904). When the aim is to emphasize top ranks, weighted distances for ranking data should be used (Tarsitano, 2005). We propose a generalization of this kind of distances using copulas. Those generalizations provide a more flexible instrument to model different types of data dependence structures and consider different situations in the classification process. Simulated and real data are used to illustrate the pertinence and the importance of our proposal.

Complexity, Data Science and Statistics through Visualization and Classification 1

Enhancement on Probabilistic Boosted-Oriented Clustering

Roberta Siciliano[1], Giuseppe Pandolfo[1], Antonio D'Ambrosio[1] ([1]: University of Naples Federico II)

The framework of this paper is probabilistic boosted-oriented clustering suitable for time series (Iorio et al., 2015). The novelty behind this methodology is to adopt a boosting prospective for unsupervised learning problems. The boosting approach proposed by Freund and Schapire (1997) is based on the idea that a supervised learning algorithm (weak learner) improves its performance by learning from its errors. From this approach we took the idea to weight each instance according to some measure

of badness of fit in order to define a fuzzy clustering process based on a weighted re-sampling procedure. In performing the fuzzy clustering, we adopt the Probabilistic Distance clustering proposed by Ben-Israel and Iyigun (2008). The proposed methodology allows us to overcome the well-known issues of the fuzzifier parameter choice. We do not have degrees of freedom to determine the membership matrix, since the probability of each instances to belong to any cluster is related to the distance of each instance to cluster center. To assign each instance to a cluster, we assume the representative instance of a given cluster as a target instance, a loss function as a synthetic index of the global performance and the probability of each instance to belong to a given cluster as the individual contribution of a given instance to the overall solution. The larger is the probability of a given instances to be member of a given cluster, the larger is the weight of that instances in the resampling process. In this paper, we extend the above approach to the analysis of standard multivariate data and to the setting of spherical data. This latter type of data is of particular interest since occurs in many areas, e.g. Earth Sciences, Meteorology, Biology and Medicine. Such data are characterized by some peculiar features (such as the lack of a natural ordering and reference direction) that make usual statistics not appropriate. For example, the standard mean provides misleading results in such setting. Suppose we have angular measurements 1°, 0°, 359°, then the arithmetic mean of these three numbers is 120°, but it is clear that a more sensible result is 0°. As a result, we define a more general probabilistic boosted oriented clustering method suitable for standard multivariate data as well as spherical data. The global performance of the proposed method is investigated by means of experiments.

References

Ben-Israel, A., Iyigun, C. (2008): Probabilistic D-Clustering. Journal of classification, 25(1); 5-26. Freund, Y., Schapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences , 55(1), 119-139.

D'Ambrosio, A., Frasso, G., Iorio, C., Siciliano, R. (2015): Probabilistic Boosted-oriented Clustering of Time Series. In Mola, F. and Conversano, C.(Eds.), CLADAG 2015 10th scientific meeting of the Classification and Data Analysis Group of the Italian Statistic Society, Book of Abstracts,(Flamingo Resort, S. Margherita di Pula, CAGLIARI, 8-10 Ottobre 2015), ISBN: 9788884677499 pp. 61-64.

Iorio, C., Frasso, G., D'Ambrosio, A., Siciliano, R. (2015): Boosted-Oriented Probabilistic Smoothing-Spline Clustering of Series. arXiv preprint arXiv:1507.04905.

Interactive Visualization of Story-boards: A Case Study of Long Time Series About Agriculture in Scotland

Michele Staiano[1], Giuseppe Pandolfo[1], Richard Aspinall[2] ([1]: University of Naples Federico II; [2]: Independent, c/o James Hutton Institute)

To characterize and understand the dynamics of change in provisioning services from agriculture in Scotland over the period 1940 to 2016, a wealth of datasets describing agricultural land use, production, and financial and energy inputs and outputs are analyzed against drivers of change in land use. By adopting an accounting framework that links funds of natural, human, physical and financial capital, with flows of goods and services it is possible to identify ways in which funds of capitals and flows of inputs and outputs of goods are linked to land management practices and policies at a national scale. The simultaneous analysis of the long time series helps us to capture true change points, and also changes between quinquennia (to accommodate lags in the land system) and are highlighted by means of multiple metrics Sankey diagrams.

Since the stories that such a composite set of data and relations embed can be captured only by an approach rooted in complexity, showing them by simple visualization requires a strategy for reduction that is tailored specifically to the relations in the stories, and by avoiding any level of compression unsuited to highlighting the relevant relations. Along with the composition of multiple graphs into a dashboard and the integration of some degrees for interaction with the single graphs, we introduce active links among pieces of evidence and propose the use of a guided sequence of clues, in the form of an interactive story-board, to sustain a process of quantitative story-telling about Scotland agriculture evolution over the last 76 years.

References

Aminikhanghahi, S., Cook, D. J. (2016): A Survey of Methods for Time Series Change Point Detection. Knowledge and Information Systems, 51(2), 339-367.

Aspinall, R., Staiano, M. (2018): Ecosystem Services as the Products of Land System Dynamics: Lessons from a Longitudinal Study of Coupled Human–environment Systems. Landscape Ecology, https://doi.org/10.1007/s10980-018-0752-7.

Heer, J., Shneiderman, B. (2012): Interactive Dynamics for Visual Analysis, Communications of the ACM, 55(4), 45-54.

Heer J., van Ham F., Carpendale S., Weaver C., Isenberg P. (2008): Creation and Collaboration: Engaging New Audiences for Information Visualization. In: Kerren A., Stasko J.T., Fekete JD., North C. (eds) Information Visualization. Lecture Notes in Computer Science, vol 4950. Springer, Berlin, Heidelberg.

Wilhelm, A., Kestler, H. A. (eds.) (2016). Analysis of Large and Complex Data. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg.

Nonparametric Clustering of Spherical Data

Giuseppe Pandolfo[1], Antonio D'Ambrosio[1] ([1]: University of Naples Federico II)

The analysis of spherical data is gaining great interest in last decades. Such data arise in many scientific fields such as bioinformatics, astronomy, environmetrics and text mining to cite just a few. Standard statistical methods may provide misleading results when used in this setting because of the peculiar features of such data. Hence, specific techniques are required and this holds true also for clustering. We propose a new nonparametric procedure based on suitable statistical data depth functions to cluster spherical objects. Such clustering method has an great potential to get insights from existing data which lie on the surface of the (hyper)sphere. Several clustering algorithms based on data depths already exist for standard Euclidean data. However, to the best of authors' knowledge, no algorithm of such type was proposed for spherical data. In the proposed algorithm the medoids of the clusters are obtained by using the concept of data depth. The performance of the algorithm called Data Depth Based Medoids Clustering Algorithm (DBMCA) is evaluated first by means of simulated datasets and then by a real dataset in text clustering. The comparison with the leading state-of-the-art alternatives demonstrates that the proposed algorithm yields good results and is robust to outliers. In addition, it is rotational invariant.

Complexity, Data Science and Statistics through Visualization and Classification 2

Spatially Weighted Exploratory Regression Tree: A Quantitative Story-telling Proposal

Carmela Iorio[1], Giuseppe Pandolfo[1], Roberta Siciliano[1] ([1]: University of Naples Federico II)

This paper was designed to deal with a real problem of statistical analysis in Geographic Information System (GIS) field. The setting up was the analysis of the Almeria GIS data matrix consisting of 376 instances, designated as all the irrigation communities in Almeria (Spain) that are grouped geographically by 18 water management areas and technically by 38 distinct water sources patterns. The data, collected by Violeta Cabello, belonging to a set of pilot case studies approached during the first year of the H2020 MAGIC project (G.A. n. 6896669). The real GIS dataset was exploited as a test for the fruitful interaction of domain experts and statisticians. The main purpose is to find a partition of irrigation communities such to predict the water consumption for each area. Within the framework of recursive partitioning algorithms by tree-based methods proposed by Breiman at al. (1984), we propose to build an explorative Regression Tree Geographically Weighted. Classification And Regression Trees (CART) can be considered as the ancestors of supervised statistical learning paradigm introduced by Vapnik (2013). Tree-based methods have been proposed for both prediction and explora-

tory purposes. Since the entire population is surveyed, we explore the irrigation communities of Almeria based on water consumption per hectare. Exploratory trees belong to data mining methods where an important role is the visualization of the results because of it helps the analyst to better understand the phenomena under study (Fayyad et al., 2002). The aim is classifying the specific consumption of water (per hectare) of the farming communities based on either water management areas and different mix of sources for irrigation water (surface, groundwater, waste water, desalination). Even if the regression tree is a device simple to be presented and understood, main issue is the tailoring of the standard approach (being the observations weighted geographically). The results show that the water consumption per hectare in some management areas is quite more spread than in some others. Moreover, the role of profiles of water sources varies in different sets of water management areas, highlighting as the two predictors interact. Finally, this proposal reducing the gap between theory and practice (domain experts) points out the usefulness of our contribution in the knowledge discovery process from databases and collaborative quantitative story-telling for MAGIC project addresses for informing nexus security.

References

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J. (1984): Classification and Regression trees. CRC press, Boca Raton.

Fayyad, U.M., Wierse, A., Grinstein, G.G. (2002): Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann in Data Management Systems.

Siciliano, R., Mola, F. (2000): Multivariate Data Analysis and Modeling Through Classification and Regression Trees. Computational Statistics & Data Analysis 32 (3), 285{301).

Vapnik, V. (2013): The Nature of Statistical Learning Theory. Springer science & business media, New York.

Median Constrained Bucket Order: A Way to Think About Tied Rankings

Antonio D'Ambrosio (University of Naples Federico II)

The rank aggregation problem can be summarized as the problem of aggregating individual preferences expressed by a set of judges to obtain a ranking that represents the best synthesis of their choices. Suppose to have a set of n items to be ranked by m judges. When a judge gives a complete and strict precedence ranking of the items, producing in fact a permutation of the first n integers, the resulting ranking is defined complete (or full). Sometimes some individuals assign the same integer to two or more items, producing a tied ranking. Sometimes tied rankings are called bucket orders, in the sense of a set of items ranked in a tie at a given location. Depending on the reference framework, the problem of aggregating individual preferences is known as social choice problem, rank aggregation problem, median ranking, central ranking, Kemeny problem. This (NP-hard) problem has gained increasing importance over the years both as a main research topic and as an essential starting point for other types of analysis. It has been tackled with several different approaches. Most of the time the detection the consensus ranking is based on the minimization of a distance measure suitably defined for preference rankings. We are adopting Kemeny's axiomatic framework. We assume that the consensus ranking is the median ranking defined as that/those ranking/s that minimize/s the sum of the Kemeny distances between itself, and the rankings expressed by a set of m judges.

This presentation concerns the median constrained bucket order, namely a specific solution of the rank aggregation problem in which the median ranking is forced to have a pre-specified number of buckets. Our proposal is motivated by the attempt to find a solution for some real data problems, one of them concerning the evaluation of a set of nurses within the so-called triage prioritization.

Clustering of All that is Exceptional and Anomalous, Counterposed to Commonality, in Big Data Analytics

Fionn Murtagh (University of Huddersfield)

From our previous work (Murtagh 2017), an important analytical issue is the resolution scale of the data and what can be the ethical issues relating to all data that is aggregated. Data aggregation may be at issue when visualization of data, in particular Big Data, is at issue. It can be very relevant for

the analysis being both effective in the outcome, and efficient computationally to have the data visualized. Often used is this expression for the analysis: visualization and verbalization of data. Here, another important theme in analysis, perhaps oriented towards clustering, is that exceptional data characteristics, semantically exceptional, in the geometric understanding of the data, that can perhaps also be characterized as anomalous data characteristics, can well be counterposed to the predominant commonality and typicality in the data that is being analyzed. A very important role in the analytics here is to have the data re-encoded, such as using p-adic data encoding, rather than real-valued data encoding. In the case study example here from text mining, the analysis is focused on, or oriented towards, a divisive, ternary (i.e. p-adic where p = 3) hierarchical clustering from factor space mapping. Hence the topology is related to the geometry of the data, and we start with the important role of the Baire ultrametric in so many application domains. A conclusion of this presentation will be the differentiation in Big Data analytics of what is both exceptional and quite unique relative to what is both common and shared, and predominant. Therefore, the analytics seeks both the typical and standard data characteristics, as well as orienting the analysis towards the exceptional and atypical data characteristics. The geometry of the factor space can define the former here, and the latter here will be quite likely to proceed towards the topology relating to, or defined by, hierarchical clustering.

References

Murtagh, F. (2017): Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics, Chapman and Hall, CRC Press.

Detecting Unobserved Heterogeneity in Latent Growth Curve Models

Laura Trinchera[1], Katerina M. Marcoulides[2] ([1]: NEOMA Business School; [2]: University of Florida)

Latent growth curve modeling is frequently used in social and behavioral science research to analyze complex developmental patterns of change over time. Although it is commonly assumed that individuals in an examined sample will exhibit similar growth trajectory patterns, there can be situations where typological differences in development and change are present. In such instances, it is important to treat the sample as stemming from unobserved heterogeneous populations. Unobserved heterogeneity is commonly analyzed using growth mixture models or group-based trajectory models. These methods are designed to identify clusters of individuals that follow a similar developmental trajectory on an outcome of interest. The methods utilize a combination of a latent growth curve model and a finite mixture model by assuming that the underlying population consists of a fixed but unknown number of groups or classes, each with distinct growth trajectories. Because group membership is not known and no observed variable is available to identify homogenous groups, group membership must in some manner be inferred from the data.

We propose a new approach to growth mixture modeling where the number of growth trajectories is determined directly from the data by algorithmically grouping or clustering individuals who follow the same estimated growth trajectory based on an evaluation of individual case residuals.

The identified groups are assumed to represent latent longitudinal segments or strata in which variability is characterized by differences across individuals in the level (intercept) and shape (slope) of their trajectories and their corresponding individual case residuals. The illustrated approach algorithmically enables the data to determine both the number of groups and corresponding trajectories. The approach is illustrated using both empirical longitudinal and simulated data.

Consumer Preferences and Marketing Analytics 1

Ranking, Rating, and Pairwise Comparisons: On Transformations Between Representations of Preference Structures

Andreas Geyer-Schulz (Karlsruhe Institute of Technology)

In this contribution we consider formal transformations between ranking, rating and pairwise comparison data. We discuss the role of the rationality of a decision-maker as captured by the von Neumann-

Morgenstern utility axioms for the existence of such transformations with minimal information loss. Potential applications of such transformations are the combination and integration of ranking, rating and pairwise comparison data from conjoint experiments or product configurators.

Multimodal Preference Heterogeneity in Choice-Based Conjoint Analysis: A Simulation Study

Nils Goeken[1], Peter Kurz[2], Winfried Steiner[1] ([1]: Clausthal University of Technology; [2]: bms marketing research & strategy)

The most commonly used variant of conjoint analysis is choice-based conjoint (CBC). Here, preferences for attributes and attribute levels are derived through choice decisions, rather than by ranking or rating tasks. This closely resembles actual market behavior and the imitation of real shopping behavior increases the external validity. In marketing research, the multinomial logit (MNL) model is widely used for preference estimation. One limitation of the basic MNL model is its inability of addressing market heterogeneity. Methods that recover market heterogeneity are of particular importance in marketing.

One method to capture heterogeneity is a choice model that partitions the market into a small number of homogeneous preference segments or latent classes. The recent development of Hierarchical Bayesian (HB) estimation methods for MNL models enables the estimation of part-worth utilities at the individual level. Very common is the MNL model with a normal distribution to describe the variation in part worth heterogeneity. Although the normal distribution is often used to model consumer heterogeneity in marketing research, this distribution is unable to represent multimodal distributions. A mixture of multivariate normal distributions as a first-stage prior allows the estimation of multimodal and skewed preference heterogeneity. Similar to the Latent Class approach consumers are assigned to one segment, which, however, allows for within-segment heterogeneity. A new and more flexible approach to modal multimodal preference heterogeneity in the context of CBC is a MNL model with a Dirichlet Process Prior (DPP) for the distribution of heterogeneity. With additional layers of flexibility part worths are drawn from a distribution with an unknown form. Therefore, this function follows a Dirichlet Process. The number and masses of segments are influenced by the model and the prior settings and do not have to be predisposed like in the Latent Class model and in the Mixture of Normals model, respectively. On the basis of eleven datasets Voleti et al. (2017) show that the MNL model with a DPP works best in terms of prediction accuracy.

The aim of this research is to compare the following choice models in a detailed simulation study (N=3120): the Bayesian MNL model (with and without heterogeneity), the Bayesian Mixture of Normals model, a Bayesian approximation of the Latent Class model and the Bayesian DPP model. To the best of our knowledge, no simulation study has yet systematically explored the performance of these Bayesian choice models. Using statistical criteria for parameter recovery, goodness-of-fit and predictive accuracy we evaluate the performance of these models under varying the number of segments, different segmentation structures (symmetric vs. asymmetric), different separations, different levels of within-segment heterogeneity, different numbers of parameters at the individual level and varying number of choice sets. For statistical inference we use analysis of variance.

References

Voleti, S., Srinivasan, V., Ghosh, P. (2017): An approach to improve the predictive power of choice-based conjoint analysis. International Journal of Research in Marketing 34, 2, 325–335.

Optimizing Tickets for Sport Event Spectators

Mario Kaiser[1], Herbert Woratschek[1] ([1]: University of Bayreuth)

Protests of sport fans against ticket prices with slogans like “£nough is £nough” or “Football without fans is nothing” show the ongoing disconnect between fans and sport organizations. In order to avoid fan protests, sport managers need to know how much potential spectators are willing to pay for a sport event ticket. However, in most sport organizations ticket prices are intuitively fixed and rarely based on estimations of spectators’ willingness-to-pay (WTP).

Existing research in sport management applies average ticket prices and homogeneous spectators' WTP (Rascher et al., 2007), although various scholars in sport management literature have addressed the heterogeneity of sport fans. Research showed heterogeneity in socio-demographic and psychographic variables (Hunt, Bristol, & Bashaw, 1999 Wann & Branscombe, 1990). To the best of our knowledge, there is a lack of research focusing on benefit segmentation of sport fans. The purpose of this research is to examine sport event spectators' preferences for tickets and their WTP with an additional focus on innovative ticket features.

Therefore, this research contributes to open research gaps in spectator segmentation and ticket pricing and addresses the following research questions: Which spectator segments can be identified regarding ticket preferences and WTP? How do ticket preferences and WTP differ between the segments?

We applied choice-based conjoint analysis with Latent Class Analysis within the case of a German basketball club. Within our research design, we used the attributes "price" and "seat category" and "opposing teams" with five attribute levels each. Other variables are standardized. We conducted an online survey. The online link was sent to fans of the investigated club via social media and relevant online forums. In total, a convenience sample of 370 completed surveys is used for further analyses.

Results lead to four different spectator segments, described by the relative importance of the attributes. The biggest segment of "Price-Sensitive Spectators" has a size of 31.6 %, whereas "Seat Quality-Oriented Spectators" are the smallest segment with a size of 10.3 %. It is noteworthy that "price" is the most important attribute (89 %) for price-sensitive spectators and seat category (56%) for seat quality-oriented spectators.

The results show that spectator preferences are heterogeneous. Therefore, club managers need to know the preferences of their spectators in order to adjust their ticket offerings better. Therefore, we calculate the maximum WTP for each segment, in each seating category, and against any opposing team, including innovative ticket features like new opponents in a league. As a consequence, we can simultaneously optimize seat categories, ticket prices, and top game surcharges. Also, we develop a ticket pricing tool for sport managers as a practical contribution. Based on the results of the latent class analysis, this ticket pricing tool enables managers to identify optimal ticket price-performance ratios.

Consumer Preferences and Marketing Analytics 2

Adaptive CBC: Are the Benefits Justifying its Additional Efforts Compared to Traditional CBC?

Benedikt Martin Brand[1], Daniel Baier[1] ([1]: University of Bayreuth)

Currently, there is a big discussion ongoing among practitioners and scientists whether the benefits of the Adaptive Choice-Based Conjoint (ACBC) analysis in comparison to the traditional Choice-Based Conjoint (CBC) analysis are justifying the additional costs and efforts of ACBC. To answer this question, recent studies in literature are reviewed and a conducted ACBC (n=205) about e-commerce in an international context is analyzed with regards to several aspects, e.g. excluded attribute levels and stimuli used for the Choice Tasks section. The results indicate that CBC is generally able to provide the main information about the most preferred attribute levels with less effort compared to ACBC. However, ACBC is very suitable for more complex products or services and for gaining deeper insights, such as information about the second-best options or completely unacceptable features. Furthermore, CBC requires a bigger sample size and is often less precise. Still, the related context will always remain the main factor for or against the usage of one or the other method.

Versioning, Price Fairness and Purchase Decision – An Empirical Investigation

Bastian Werner[1], Ines Brusch[1], Larissa Bürks-Arndt[1] ([1]: Brandenburg University of Technology Cottbus-Senftenberg)

Prior research suggests that companies often offer different variants of the initial product to fit as many heterogenic consumer needs and wants as possible to capture the differential willingness to pay of consumers. The production method at which existing parts of a product are improved, reduced, corrected or degraded is called Versioning (De Sordi et al. 2016).

Pursuant to rational choice theory consumers weigh benefits relative to their costs in evaluating a product and generate the purchase decision. Consequently, the production method should be irrelevant. The empirical evidence of this study contradicts this thought. Based on Equity-Theory (Adams 1963; Nguyen et al. 2014), Dual-Entitlement-Theory (Kahneman et al. 1986; Chen et al. 2018) and Value Function (Kahneman/Tversky 2013) a quantitative survey of 211 subjects is carried out.

In this context, the four versioning methods were examined to determine whether they appear fair to consumers and how/if they influence their purchasing decisions. In addition, consumer expectations regarding price-quality, price-life span, price-costs, and price-production steps were reviewed. The empirical evaluation is carried out using various multivariate analysis methods (e.g. cluster analysis). The tested products (tights, printers and washing machines) have different life expectancies, degrees of involvement, and qualities. A total of three product types with three different qualities each and four product variants each were tested. Then the perceived price fairness of products with different life expectancies and production processes as well as their effects on the purchase decision are analyzed.

The results provide new insights for researchers from a theoretical and practical point of view, e.g. price fairness and ethical convictions have enormous effects on purchasing decisions. Finally, the paper gives some general implications and recommendations for future research.

References

Adams, Stacy J. (1963): Towards an Understanding of Inequity, in: The Journal of Abnormal and Social Psychology, 67(5), 422.

Chen, Haipeng A., Bolton, Lisa E., Ng, Sharon, Lee, Dongwon, Wang, Dian (2018): Culture, Relationship Norms, and Dual Entitlement, in: Journal of Consumer Research, 45 (1), 1-20.

De Sordi, José O., Reed, Elliot N., Meireles, Manuel, da Silveira, Marco A. (2016): Development of Digital Products and Services: Proposal of a Framework to Analyse Versioning Actions, in: European Management Journal, 34(5), 564-578.

Kahneman, Daniel, Knetsch, Jack L., Thaler, Richard H. (1986): Fairness and the Assumptions of Economics, in: Journal of Business, 285-300.

Recommender Systems for Personalized Advertising: Success Factors for Designing Product Recommendations from a Customers' Perspective

Timo Schreiner[1], Alexandra Rese[1], Daniel Baier[1] ([1]: University of Bayreuth)

Nowadays, recommender systems are widely used in various contexts and across different areas of ecommerce providing benefits for both, firms and customers by increasing product sales and allowing for cross- and upselling as well as supporting and facilitating users' decision-making.

We present an overview of popular and commonly used recommender algorithms as well as current developments in research and state-of-the-art applications in practice. In addition, success factors of recommender systems beyond the algorithm are discussed based on a literature review serving as a starting point for our empirical research.

In a study within the apparel industry, we examine consumers' preferences regarding product recommendations in advertisements across different media channels applying choice-based conjoint analysis where all stimuli are presented visually via specifically designed advertisements. The findings of two distinct experiments for young male (n=170) and female (n=162) consumers show that the recommender algorithm is not necessarily of upmost importance. In contrast, the advertising channel is of highest relevance with banner advertising being the least preferred channel.

Differences between male and female respondents are also clearly outlined: While males prefer a rather small set of product recommendations (four at a time) and favour the presentation of product recommendations in package inserts females rather wish for larger recommendation sets (twelve at a time) presented within email messages.

Finally, we present and discuss suggestions for designing product recommendations and provide an outlook for future research possibilities.

Keywords: Recommender Systems, Personalization, Multichannel, Electronic Commerce, Conjoint Analysis, User-centric Evaluation.

Measuring Preferences for Complex Products – Self-explicated Methods vs. Pairwise Comparison-based Preference Measurement (PCPM)

Jörgen Eimecke[1], Madeline Hecht[2], Julia Clara Buchold[2], Daniel Baier[2] ([1]: BF/M-Bayreuth; [2]: University of Bayreuth)

Preference measurement is a widely used method to collect the importance of product variants that customers ascribe. Furthermore, as products getting more and more complex the number of relevant attributes and attribute levels is increasing. For this reason, there is a need for methods which on the one hand are able to process a great deal of information and on the other hand do not overburden the subjects with the number and complexity of the tasks to be evaluated (e.g., Park, Ding and Rao 2008, 562). Conjoint Analysis – as a representative of decompositional methods – was the most famous tool for preference measurement for years but it does not seem to be suitable for products with more than six attributes because of (e.g.) the use of heuristics by the subjects (Green and Srinivasan 1990, 8). Self-explicated approaches – as representatives for compositional methods – could be used for preference measurement of complex products but direct measurement of preferences has been criticized as being imprecise. Facing these problems Scholz, Meißner and Decker (2010) presented a new compositional approach based on paired comparisons (PCPM). In this and later publications the PCPM was compared to mainly new adaptive decompositional approaches.

The present study compares the PCPM with two approaches of two-staged Self-explicated methods (rank order; constant-sum method and rank order; rating scale with one anchor). The object of the investigation are student accommodations. To determine the more suitable method for measuring preferences of complex products direct and indirect validity criteria are analysed. The results are compared with previous studies to give recommendations which method should be used for complex products. Moreover, indirect validity criteria show partially superior outcomes of SEM, which reaffirms its applicability in this context.

References

Green, Paul E. und V. Srinivasan (1990): Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice, Journal of Marketing, 54 (4), 3–19.

Park, Young-Hoon, Min Ding und Vithala R. Rao (2008): Eliciting Preference for Complex Products: A Web-Based Upgrading Method, Journal of Marketing Research, 45 (5), 562–574.

Scholz, Sören W., Martin Meissner und Reinhold Decker (2010): Measuring Consumer Preferences for Complex Products: A Compositional Approach Based on Paired Comparisons, Journal of Marketing Research, 47 (4), 685–698.

Consumer Preferences and Marketing Analytics 3

Investigating Machine Learning Techniques for Solving NP-hard Product-line Optimization Problems

Sascha Voekler[1], Daniel Baier[2] ([1]: Brandenburg University of Technology Cottbus-Senftenberg; [2]: University of Bayreuth)

Product-line optimization using consumers' preferences measured by conjoint analysis is an important issue to market researchers. Since it is a combinatorial NP-hard optimization problem, several meta-heuristics have been proposed to ensure at least near-optimal solutions. This work presents already used meta-heuristics in the context of product-line optimization like genetic algorithms, simulated annealing, particle-swarm optimization, and ant-colony optimization. Furthermore, other promising approaches like tabu search, harmony search, multiverse optimizer and memetic algorithms are introduced to the topic. All of these algorithms are applied to three different objective functions for optimizing buyers' welfare, market share and firms' profit. The performance of the meta-heuristics is measured in terms of best and average solution quality, variance, needed CPU time, and robustness testing for part-worths under uncertainty. To determine the most suitable meta-heuristics for the underlying objective functions, a Monte Carlo simulation for hundreds of problem instances with simulated data is performed. Simulation results suggest the use of genetic algorithms, simulated annealing and memetic algorithms for product-line optimization.

References

Albritton, M. D., McMullen, P. R. (2007): Optimal Product Design Using a Colony of Virtual Ants, in European Journal of Operational Research, 176 (1), 498–520.

Baier, D., Gaul, W. (1999): Optimal Product Positioning Based on Paired Comparison Data, in: Journal of Econometrics, 89 (89), 365–392.

Balakrishnan, P. S., Jacob, V. S. (1996): Genetic Algorithms for Product Design, in: Management Science, 42 (8), 1105–1117.

Belloni, A., Freund, R., Selove, M., Simester, D. (2008): Optimizing Product Line Designs: Efficient Methods and Comparisons, in: Management Science, 54 (9), 1544–1552.

Kohli, R., Sukumar, R. (1990): Heuristics for Product-line Design Using Conjoint Analysis, in: Management Science, 36, 1464–1478.

Tsafarakis, S., Marinakis, Y., Matsatsinis, N. (2011): Particle Swarm Optimization for Optimal Product Line Design, in: International Journal of Research in Marketing, 28 (1), 13–22.

New Notion of Constructing Brand Switching Matrix with Diagonal Elements

Akinori Okada[1], Hiroyuki Tsurumi[2] ([1]: Rikkyo University; [2]: Yokohama National University)

A brand switching matrix represents the frequency of changes from one brand to others in two consecutive purchase occasions among a set of brands in a group of consumers. When we derive a brand switching matrix of durable consumer goods such as automobiles or household electric appliances, it is not difficult to define the brand switching simply by comparing two brands purchased at the first and the second occasions respectively. When we derive a brand switching matrix of non-durable consumer goods such as soft drinks or snacks, it is desirable to define the brand switching differently, because the non-durable consumer goods are frequently purchased and sometimes more than one brand are purchased simultaneously in one occasion. One possible way of dealing with this is to define the brand switching based on the most purchased brand (amount of money or quantity) in a certain period. The (j,k) element of the brand switching matrix based on the most purchased brand in a period represents the number of consumers whose most purchased brand in the first period is brand j and the most purchased brand in the second period is brand $k\|2$.

The brand switching matrix based on the most purchased brand has a drawback; it depends only on one brand in each of the first and the second periods. The other brands which were second most purchased, the third most purchased, …, are ignored in deriving a brand switching matrix. The procedure of deriving a brand switching matrix based on all brands was introduced earlier (Okada and Tsurumi, 2018). The procedure has another drawback; the diagonal element is not defined. The diagonal element of a brand switching matrix shows how loyal consumers are to the brand. In the

present study, a procedure of defining diagonal elements of a brand switching matrix, which is compatible with the definition of off-diagonal elements, is introduced. The procedure is based on the comparison of the amount of purchase of a brand in the first and second periods. The brand switching matrix on potato snacks derived by the present and the earlier procedures are compared by using asymmetric multidimensional scaling based on the singular value decomposition. The two-dimensional configuration was adopted as the solution for both procedures. The first dimension derived from the brand switching matrix formed by the present procedure represents similar relationships among brands which is represented by the two-dimensional configuration derived from the brand switching matrix formed by the earlier procedure. The second dimension derived from the brand switching matrix formed by the present procedure represents relationships among brands which are not fully represented by the other procedure.

Keywords: Asymmetry, Brand Switching, Diagonal Element, Multidimensional Scaling, Non-durable Consumer Goods.

References

Okada, A., & Tsurumi, H. (2018): A New Notion of Constructing Brand Switching Matrix and its Application [summary]. Abstract of the 7th German-Japanese Symposium, TU Dortmund University.

Bored out or Stressed out – Respondents‘ Aggregation and Simplification Patterns in Preference Measurement

Alexander Sänn[1], Jörgen Eimecke[2] ([1]: University of Bayreuth; [2]: BF/M Bayreuth)

Today the use of surveys to determine customers’ preferences for marketing is challenged by the digital revolution that may have an impact on the quality of survey results. Applying common methodologies of preference measurement like conjoint analysis in its methodological variety and others with durations up to 15 minutes and with more than 10 choice tasks requires mental effort and time from the respondent. Both requirements became rare in recent times of permanent information push and information overload. Today, not only the way how customers are using their communication devices is altered, but also perception of traditional surveys as a reliable source for business development is in doubt. Previous studies already showed that the validity of such surveys decreased in the last decade. Though, this longitudinal observation supports the discussion on reliability of pure survey results to create services and/or products, to invest substantial financial effort, and to bind further resources for development in practice.

The basic circumstances lead to thoughts on how results of online surveys are affected by “permanent” distractions within a constant competition for a respondent’s attention. Does the fundamental methodology cause bored out and/or stressed out respondents (or in other words: is the methodology the root cause to lose the competition for a respondent’s attention) and is it negative or positive to lose the competition? How does the respondent cope the situation and how can this be turned into a positive value to be challenged in future applications?

Therefore, in this contribution multiple observations and studies are discussed. First, it is analysed if respondents show basic signals for boredom or stress within a preference measurement and it is examined what kind of behavioural patterns are applied by the respondent. Based on standard strategies the respondent may either concentrate on a minimal set of individual attributes or falls into a “click-through” mode and use heuristics. It is further evaluated how the result of a preference measurement is altered in terms of validity by the applied coping strategy.

In this scenario we used choice based conjoint analysis (CBC) as the methodological base for preference measurement, since it is commonly applied in practice. We applied eye-tracking analysis to evaluate areas of interest to determine coping strategies, if boredom or stress was identified per respondent. The basic identification is done with EDA measurement. The study was applied within a laboratory setting.

Second, this contribution is focused on a possible solution tactic from a marketing perspective to gain back attention from respondents by using medial stimuli as incentives. This is set up against results with video stimuli.

The general thought on a stressed or bored state to unveil respondents' true preferences and implications for market research to respect the coping mechanism in methodological parts for future preference measurement are discussed to finalize this contribution.

Data Analysis in Finance 1

Minimum MSE Hedging of Complex Financial Retail Structured Products in Discrete Time

Philip Rosenthal[1], *Rainer Baule*[1] ([1]: University of Hagen)

Bonus certificates are amongst the most important structured financial retail products sold by banks to private investors. Banks need to hedge their positions against market risk, especially price movements of the underlying asset, as they make their profits by incorporating margins in their prices and not by trading and speculating against their own customers. Delta hedging is a standard method for hedging vanilla call options. However, a straightforward implementation may lead to undesirable hedging errors when exotic path-dependent barrier options are embedded, such as down-and-out puts in the case of bonus certificates. Further complications arise when hedging is only possible in discrete time, which is the case in any practical implementation. This is especially true when considering overnight risk.

We propose a method that takes into consideration both possible price jumps as well as the nonlinear and discontinuous nature of the certificate's payoff function. In contrast to classical delta hedging, our approach not only considers infinitesimally small price movements of the underlying, but the whole range of possible price movements during a short-term hedging time interval, for example a trading day or the overnight period when exchanges are closed. The goal is to minimize the mean squared error (MSE) of the hedging portfolio, consisting of the certificate (or, equivalently, a down-and-out put option) and the hedging instrument, in discrete time.

One major benefit of our proposed method is that it allows for different hedging instruments to be used. We therefore not only use the underlying itself but also look at vanilla call options with strike equal to the down-and-out put's barrier as suitable candidates. We show that deltas obtained by our method are lower than standard Black-Scholes deltas and reduce the MSE by a factor 2–3, depending on the actual setup. The usage of call options instead of the underlying further reduces the hedging error. However, efficiency depends greatly on the option's time to maturity. Short expiry times yield the best results. Nevertheless, even if the call option's time to expiry is greater than the down-and-out-put's one, an improvement in hedging performance is still possible. Hedging is most challenging when the intrinsic value of the down-and-out put is high near the barrier and when its remaining time to maturity is short. Interestingly, higher volatility leads to smaller hedging errors. The level of the risk-free rate and its effect on hedging performance is negligible.

The Volatility of Hourly Electricity Contracts on the Continuous Intraday Market

Michael Naumann (University of Hagen)

We analyze the volatility of electricity contracts on the continuous intraday market of the power exchange "EPEX SPOT". "EPEX SPOT" is the most important European power exchange for the short-term trading of electricity. This power exchange offers several products, in particular the day-ahead auction and the intraday continuous trading with different periods of the electricity contracts (15, 30, and 60 minutes). In contrast to the majority of the existing literature, which focuses on the day-ahead auction, we analyze the volatility of the hourly electricity contracts of the intraday continuous market. This market segment has experienced a growing importance in the recent past, with respect to both absolute and relative trading volume. The study is based on an empirical data set from October 2015 to September 2018. For various reasons, the classical measure of volatility as a standard deviation of log returns known from financial time series cannot be directly applied to the electricity market.

Therefore, we propose five different risk measures: the volume-weighted standard deviation, the outlier-adjusted volume-weighted standard deviation, the absolute distances, the interquartile range and the range between the 90 % and 10 % quantile. These measures turn out to be equally well suited to describe the risk on the intraday continuous market, since they are highly correlated. We furthermore analyze the influence of renewable energies (solar and wind energy) on the risk measures, finding evidence that in particular the generated wind energy (besides the trading volume) significantly increases the volatility measures. Furthermore, there are significant influences of different seasons (spring, summer, autumn and winter) and days (workday and holiday) on electricity price volatility.

The Impact of Financial Speculation on Commodity Prices: A Meta-Granger Analysis

Jerome Geyer-Klingeberg[1], Marie Hütter[1], Andreas Rathgeber[1], Florian Schmid[1], Thomas Wimmer[1]
([1]: University of Augsburg)

The market environment of commodities trading has undergone substantial changes over the last decades. Often termed as “financialization of commodity markets”, commodities have become an increasingly attractive asset class for investors. In this context, also financial speculation, amplified by the emerging popularity of index related financial products, has dramatically increased the trading activity in commodity futures markets. The impact of financial speculation on commodity markets attracts enormous attention in both public media and academic research. Despite the rapid expansion of research publications, empirical evidence on the impact of speculation is still contradictory and without clear bottom line.

In this study, we apply meta-regression analysis on a sample of 70 empirical studies reporting 3,892 estimates from Granger causality tests to disentangle which study characteristics determine the wide heterogeneity of previous findings. This approach extends previous reviews in the field (Boyd et al. 2018, Grosche 2012, Haase et al. 2016, Shutes and Meijerink 2012, Will et al., 2016) and contributes to the literature in several ways: (i) We provide the first statistical integration of the accumulated research and show that the literature as a whole does not detect speculation effects. In contrast, the analysis of certain subgroups, e.g. whether authors are affiliated with an inter- or non-governmental organizations or studies focusing on agriculture, provides evidence that speculation drives commodity prices under certain conditions. (ii) We apply the meta-regression model for Granger causality testing by Bruns and Stern (2018) to assert the lack of an overall publication selection bias. In addition, the model shows that primary studies only rarely suffer from overfitting via lag selection. (iii) Within the meta-regression model, we explicitly test the joint impact of various aspects of study design, such as the analyzed commodity type, measurement differences, and methodological characteristics of the primary study models. We conclude that heterogeneous findings of the previous literature are largely attributable to those characteristics, especially to the commodity type as well as the use of report-based data. In contrast to soft commodities like corn and wheat, metal markets are largely unaffected by speculation impacts. Furthermore, studies using publicly available, report-based data to measure speculation find significantly less commodity price distortions than non-report based analyzes. Future studies might be evaluated against the benchmark given by our meta-findings. Moreover, insights from the meta-granger analysis support policy makers and related organizations, as well as participants in commodity markets at figuring out under which conditions speculation genuinely impacts commodity prices.

Data Analysis in Finance 2

Systemic Risk and Measures of Connectedness in Financial Systems for Chosen Countries

Katarzyna Kuziak[1], Krzysztof Piontek[1] ([1]: Wrocław University of Economics)

The recent global financial crisis has emphasized the importance of connectedness as a key dimension of systemic risk. Systemic risk involves the financial system, a collection of interconnected institutions that have mutually relationships through which losses can quickly propagate during periods of financial distress. In this paper, we use two econometric methods to capture this connectedness – principal components analysis and Granger-causality tests – and apply them to evaluate systemic risk in financial system. We will use principal components analysis to estimate the number and importance of common factors driving the returns of chosen financial institutions, and we will use pairwise Granger-causality tests to identify the relations among these institutions. Empirical study will be conducted for financial institutions in Poland and Germany.

Keywords: Systemic Risk, Financial Institutions, Financial Crises.

References

Acharya, V., Pedersen, L. H., Philippon, T., Richardson, M. (2011): Measuring Systemic Risk. Unpublished working paper. New York University.

Billioa, M., Getmansky, M., Lo, A.W., Pelizzona, L. (2012): Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors, Journal of Financial Economics, Volume 104, Issue 3, pp. 535-559.

Diebold, F. X., Yılmaz, K. (2015): Financial and Macroeconomic Connectedness: A Network Approach to Measurement and Monitoring.

Glasserman P., Young H. (2014): How Likely is Contagion in Financial Networks? Journal Bank. Financ., Volume 50, pp. 383-399.

Li, F., Perez-Saizb, H. (2018): Measuring Systemic Risk Across Financial Market Infrastructures, Journal of Financial Stability, Volume 34, pp. 1-11.

Silva, W., Kimura, H., Sobreiro, V.A. (2017): An Analysis of the Literature on Systemic Financial Risk: A Survey, Journal of Financial Stability, Volume 28, pp. 91-114.

Low Volatility Anomaly and Value at Risk: Risk-adjusted Returns of Stock Portfolios Based on VaR Predictions

Bernhard Lange[1], Matthias Gehrke[1] ([1]: FOM University of Applied Sciences)

The low volatility anomaly can be observed in many developed and emerging markets around the globe since about 1970. It states that an investment strategy which optimises asset allocation towards stocks with low volatility of daily returns leads to higher risk-adjusted returns than a market or equal weight portfolio. Substituting volatility with Value at Risk (VaR) as a measure of risk avoids unrealistic assumptions about investor behaviour and symmetric return distributions. This paper contributes to the discussion and extends the research on the continued existence of the low volatility anomaly. Furthermore, we investigate the impact of this strategy during different market phases and the question whether the VaR forecast accuracy influences the risk-adjusted return positively. Therefore, we forecast the one-day-ahead VaR of stocks based on three quantitative models (Historical Simulation, HAR-QREG, GARCH) and construct portfolios comprised of stocks with the lowest projected VaR. We test the hypothesis of higher risk-adjusted returns of a low-VaR portfolio vis-à-vis an equal weight portfolio with stocks from three large cap indices (Euro Stoxx 50, S&P 100, Nikkei 225) covering the years 2004 to 2016. The risk-adjusted return is illustrated by the Sharpe Ratio.

The results show that the assumptions of the low volatility anomaly are only valid when considering tranquil market phases, i.e. phases of low stock volatility, in all tested stock markets. However, the results indicate no relationship between the accuracy of forecast models and the respective risk-adjusted return. We show that low-VaR portfolios lack the ability to follow steep bullish markets compared to equal weight portfolios, which is the reason for their lower Sharpe Ratios in market phases of high stock volatility. Moreover, we find no evidence for a relation between forecast accuracy and

risk-adjusted return. The missing link can be explained by the - almost arbitrarily distributed - absolute returns of low-VaR portfolios with respect to their forecast accuracy. However, all tested models enable a significant reduction of portfolio volatility. While the results are in line with comparable studies regarding market phases suffering from low volatility, the failing in times of high volatility is not widely documented.

What Makes the Crowd Wise? An Empirical Analysis of Fundamental Information, Investor Sentiment, and Stock Market Performance

Brigitte Eierle[1], Sebastian Klamer[1], Matthias Muck[1] ([1]: University of Bamberg)

Prior research has found that sentiment derived from online discussions predicts stock market movements and attributes this observation to the concept of the “Wisdom of Crowds”, where the aggregate opinion of a crowd could be more precise than the opinion of experts. We analyse empirically whether traditional corporate disclosure about fundamentals drives the daily sentiment of users on Twitter and whether the drivers of sentiment matter for the ability to predict next day stock returns. For the purpose of this analysis, tweets about individual stocks included in the S&P 500 Index are collected for a period of 9 months and the aggregate sentiment regarding individual stocks is determined on a daily level using a word counting approach. Earnings announcements, dividend announcements, and 8-K filings are identified as corporate disclosure events, which are potential drivers of individual investors’ sentiment on online discussions. The results show that daily Twitter sentiment is partly driven by these events at the day of the announcement and several days afterwards. However, results also show that traditional corporate disclosure explains only a minor portion of the dispersion in daily sentiment on Twitter. Hence, we conclude that sentiment is mainly driven by other non-fundamental information. In a subsequent analysis, we examine whether the ability to predict next day stock returns differs between normal sentiment justified by existing fundamental information and abnormal sentiment not explained by existing fundamental information. To measure normal and abnormal sentiment, fitted values and residual values are estimated with a linear model, which controls for fundamental performance and the normal activity of users on Twitter. The results of this analysis show that only abnormal sentiment has predictive power. Hence, the ability of sentiment to predict next day stock returns is mainly driven by sentiment not justified by existing fundamental information. Our findings provide new insights into the drivers of sentiment from online discussions on Twitter and highlight that especially sentiment not justified by a firms’ fundamental performance provides value relevant information.

The Degree of Corporation Diversification: Insights from Ownership Concentration and Ownership Identity

Joachim Rojahn[1], Florian Zechser[1] ([1]: FOM Hochschule Essen)

Numerous studies reveal valuation discounts for diversified companies. In this analysis we take a step back and ask about the determinants of corporate diversification. Specifically, we extend previous research by examining the impact of owner concentration and identity (e.g. strategic investor, financial investor, foreign vs. domestic ownership, etc.) on the degree of corporate diversification that is measured by various proxies.

We do so by analyzing a sample of firms listed on the German Prime Standard segment, as the ownership structure of German issuers is quite heterogenous. Covering the years 2006 to 2016, the unbalanced panel is comprised of about 2,800 yearly firm observations. While controlling for a wide range of additional explanatory variables, we apply a nonlinear Tobit procedure adopted for panel data as well as gradient boosting techniques.

Data Analysis in Finance 3

Negations and Newspaper Sentiment – Improving and Testing the BPW-Dictionary

Matthias Pöferlein (University of Bamberg)

The sentiment-analysis of textual documents is a growing field in financial research, where one of the most common and widely used approaches is the bag-of-words-technique. It compares the words used in a textual document to word lists (dictionaries) containing words that are regarded to have a certain sentiment, e.g. positive or negative. This classification of words used and their fraction of text can be used afterwards for further analyses.

General dictionaries, like the Harvard Dictionary or the Linguistic Inquiry and Word Count (LIWC) lead to a misclassification of common words in financial related texts. This is attributable to an aim at a different type of sentiment. Loughran McDonald (2011) analyzed that almost three-fourth of the words identified as negative by the Harvard Dictionary are words typically not considered negative in a financial context (e.g. the word “liability” or “tax”). To analyze such texts, they created and tested an own domain-dependent dictionary (LMD-Dictionary) that has been widely used on different texts, like financial disclosures, press releases or newspaper articles.

Due to the existence of the LMD-Dictionary and the lack of an equivalent German dictionary, most research focuses on English texts. The findings of other studies reinforced the need for a German domain-dependent dictionary like the LMD-Dictionary. Until 2017 only general dictionaries like the German adaption of the LIWC and SentiWS (based on the Harvard Dictionary) were available in the German language. In 2017/2018 Bannier, Pauls and Walter introduced the German BPW-Dictionary based on the LMD-Dictionary (Bannier et al. 2018). This important contribution makes it possible to answer a wide range of scientific questions regarding the analysis of financial texts in the German language.

In this paper the BPW-Dictionary is complemented with a German adaption of the negations introduced by Loughran and McDonald (2011), based on an analysis of annual reports of German DAX and MDAX companies. This modified-BPW-Dictionary is tested against its initial version and the general dictionaries LIWC and SentiWS. The tests will be applied on a corpus of CEO-Speeches (analog to Bannier et al. 2017) and newspaper articles of leading financial newspapers. This contribution will show if the German adaption of the LMD-Dictionary introduced by Bannier et al. (2018) is superior in classifying other corpuses than CEO-Speeches and if an allowance for negations is able to improve the results.

References

Bannier, Christina E., Pauls, Thomas, Walter, Andreas (2017): CEO-Speeches and Stock Returns. Working Paper.

Bannier, Christina E., Pauls, Thomas, Walter, Andreas (2018): Content Analysis of Business Specific Text Documents. Introducing a German Dictionary. Working Paper.

Loughran, Tim, McDonald, Bill (2011): When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. In: The Journal of Finance 66 (1), S. 35–65.

Bubble Detection in Error Correction Models

Leopold Sögner[1], Martin Wagner[2] ([1]: Institute for Advanced Studies; [2]: TU Dortmund)

We develop consistent monitoring procedures, with the goal to detect bubbles in a Johansen-type error correction model. In particular, we consider breaks where the cointegration rank remains constant as well as breaks changing the cointegration rank. We develop Lagrange multiplier tests allowing to monitor these kinds of breaks. The monitoring procedure is used to detect possible bubbles in the triangular arbitrage parity.

Data Analysis in Finance 4

Extended Construction of the Credibility Premium

Anne Sumpf (Technische Universität Dresden)

This article generalizes the credibility framework to define the p-credibility premium by adding higher exponents of the past observations in the structure of the premium. We present a system of equations from which the p-credibility premium can be calculated and show that the premium error is decreasing in p and converges for every $p \text{ in } \mathrm{N}$ in the number of observations. For the Bühlmann-Straub model, we present formulas for the 2-credibility premium and 3-credibility premium. Estimators for the structure parameters of the Bühlmann-Straub model are also shown. Finally, we illustrate the behaviors of p-credibility premium for varying p with simulated data.

Pricing options via Fourier Transform - Analysis of Computational Speed and Accuracy

Arkadiusz Orzechowski (Warsaw School of Economics)

Assumptions concerning the Brownian motion and the normal distribution play a significant role in the financial literature on pricing options. Due to empirical investigations two arguments pertaining valuation of this type of contracts have been raised, i.e. (1) return distributions of stocks do not always follow normal distributions and (2) there is a volatility smile or smirk in option pricing. In consequence, alternative models of pricing options appeared (to the F. Black and M. Scholes approach), e.g. jump - diffusion models, pure jump models and stochastic volatility models (with or without jumps). All these models can be solved via Fourier transform. The aim of the article is to propose two alternative methods of determining Fourier transforms, which can be applied to the valuation of European options. As a part of conducted research computational speed and accuracy of each of the approaches are analyzed. For this purpose, the usefulness of various numerical schemes of calculating inverse Fourier transform is investigated. The final conclusion is that there is still room for improvement of the valuation of European options both in terms of computational speed and accuracy, especially in the stochastic volatility models.

Algorithmic Trading Using Long Short-Term Memory Network and Portfolio Optimization

Riccardo Lucato[1], Edgar Jimenez[1], Eduardo Salvador Rocha[1], Yang Qi[1], Marina Gavrilyuk[1], Rafael Rêgo Drumond[1], Lars Schmidt-Thieme[1] ([1]: University of Hildesheim)

Investors typically rely on a mix of experience, intuition, knowledge of economic fundamentals and real-time information to make informed choices and try to get as high a rate of return as possible. Their decisions are typically more instinct-driven than methodical. Propelled by the need for numerically inspired judgments, ever stronger within the financial community, in recent years the usage of computational and mathematical tools has been taking root. In this paper, we use a Long Short-Term Memory (LSTM) Network trained on historical prices to predict future daily close prices of several stocks listed on the New York Stock Exchange (NYSE). We compare the predictions of our LSTM network with those produced by another state-of-the-art approach, the Hidden Markov Model (HMM), in order to validate our findings. Subsequently we feed our forecasts into a Markowitz Portfolio Optimization (MPT) procedure to identify the best trading strategy. The purpose of MPT, which allows for simultaneous and optimal trading of multiple stocks, is to compute a set of daily weights representing the portion of initial capital to be invested in each company. Our empirical results highlight two facts: firstly, our LSTM model achieves higher accuracy than the standard HMM approach. Secondly, by trading various stocks at the same time we can obtain a higher rate of return than is possible by using the single stock strategy, while also greatly enhancing the real-world applicability of our model.

Data Analysis in Industrial Automation 1

Data Analytics for Process Improvement in Industrial Settings

Stefan Thalmann[1], Matej Vukovic[2], Jürgen Mangler[3], Christian Huemer[3], Gerti Kappl[3], Stefanie Lindstaedt[4] ([1]: University of Graz; [2]: Pro2Future GmbH; [3]: ACDP GmbH; [4]: Graz University of Technology)

Digitization in industry is an answer to shorter product life cycles, increased customer demand for individualized products and to globalization of supply chains. However, this demands more flexible and interconnected supply chains and especially an intensive data exchange between all involved supply chain partners. Also due to cheap and advanced sensors on industrial equipment it is easier to collect data in industrial settings. However, it is still a challenge to use analytics beyond tracking historical performance of single machines. However, just knowing what happened for a single machine is not enough to address the demands induced by digitisation. Especially industrial processes across organizations require a more holistic and connected perspective on industrial data analytics.

The introduction of process management technology to orchestrate business processes from ERPs all the way down to machine control seems an appropriate answer and produces much more interconnected data. Insights from data analytics can be used to improve business process at design time but also to improve the execution of business processes during run time.

In the scope of this paper we present a design study on how advanced data analytics can be used for process improvement in industrial settings. The design study is conducted in a high precision metal manufacturing production line. In the center of our technical solution is a process engine performing the machine control on the one hand and collecting and integrating the machine data on the other hand. Based on the data collected we developed a data analytics component in Python to gain insights from the collected data.

Based on our analysis we present results for three types of changes: (1) Insights of the data analytics component lead to changes in the process structure. This means we gain knowledge about how the process can be improved and this results in a remodeling of the process during design time. (2) Insights of the data analytics component lead to an adaptation of parameter settings. This means that insights about optimal machine parameters lead to a new configuration of parameter settings at design time. (3) Data analytics becomes part of the process logic during run time. This means that process steps that make use of data analytics are part of the process design. Thereby, advanced data analytics models are used at runtime to make decisions in the process logic, e.g. forecast of the quality of a product to judge on further process steps. Based on our design study we demonstrate how advanced data analytics can be used for process improvement of industrial processes during design and run time. Specifically, we demonstrate how data analytics techniques including machine learning and predictive algorithms can be used to gain useful insights about processes from industrial sensor data. We show for our case that this approach has the potential to improve industrial processes by reducing costs for quality control and maintenance activities, reducing the scrap rate, avoiding equipment failures and by improving product quality.

Applying Two Theorems of Machine Learning to the Forecasting of Biweekly Arrivals at a Call Center

Michael Christoph Thrun[1,2], Julian Märte[1], Peter Böhme[2], Tino Gehlert[2] ([1]: University of Marburg; [2]: Viessmann Werke GmbH & Co. KG)

The forecasting of arriving calls in call centers plays a crucial role in determining appropriate staffing levels and scheduling plans (Ibrahim et al. 2016). Usually, the number of calls is forecasted in a period and the best forecasting method is chosen by MAPE or MAE (Taylor 2008) which has two problems. First, the incoming calls are dependent on the background of the call center and customer leading to the comparison of several types of problems which makes any specific choice of one forecasting algorithm impracticable (Wolpert 1996). Thus, we propose to change the data representation of the problem from arriving calls to issues because a low service level can result in multiple calls regarding the same topic. Second, evaluation of forecasting results is always biased if a quality measure (QM) is seen as a similarity measure between the forecast curve and the test set curve of data because

any two different curves share the same number of properties (c.f. Watanabe 1969). Thus, the QM should be chosen accordingly to the goal, namely, capacity planning with the key performance indicator defined as the service level (Aksin et al. 2007). A forecast smaller than the real value leads to undesired understaffing. Therefore, a forecast should be more similar, if it lies above the real value. Special events are usually known prior by the call center manager. Hence less weight should be put on outliers. Capacity planning of the call center using five years of daily historical data and weather data was performed by an ensemble of an additive decomposition model (Geurts et al. 2006) combined with random forest regression (Geurts et al. 2006). For a forecast horizon of 14 days over a year of test data, the average forecasting quality is 91.3% outperforming (Taylor 2008, Ibrahim and L'Ecuyer 2013). For the decomposition model (Geurts et al. 2006), all parameters were optimized w.r.t. MRE and bias (Kourentzes et al. 2014). This Pareto optimization problem was resolved using a radial basis function surrogate that is successively improving around areas of importance using a recursive inversion formula.

References

Aksin, Z., Armony, M., Mehrotra, V. (2007): The Modern Call Center: A Multi-disciplinary Perspective on Operations Management Research. Production and operations management 16(6), p. 665-688.

Geurts, P., Ernst, D., Wehenkel, L. (2006): Extremely Randomized Trees. Machine Learning 63(1): p. 3-42.

Ibrahim, R., et al (2016): Modeling and Forecasting Call Center Arrivals: A Literature Survey and a Case Study. International Journal of Forecasting 32(3): p. 865-874.

Ibrahim, R., L'Ecuyer, P. (2013): Forecasting Call Center Arrivals: Fixed-effects, Mixed-effects, and Bivariate Models. Manufacturing & Service Operations Management 15(1): p. 72-85.

Kourentzes, N., Trapero, J.R., Svetunkov, I. (2014): Measuring the Behaviour of Experts on Demand Forecasting: a Complex Task, Department of Management Science: Technical report published in the Lancaster University Management School. p. 1-23.

Taylor, J.W. (2008): A Comparison of Univariate Time Series Methods for Forecasting Intraday Arrivals at a Call Center. Management Science 54(2): p. 253-265.

Watanabe, S. (1969): Knowing and Guessing: A Quantitative Study of Inference and Information. New York, USA: John Wiley & Sons Inc.

Wolpert, D.H. (1996): The Lack of a Priori Distinctions Between Learning Algorithms. Neural computation 8(7): p. 1341-1390.

Deep Neural Networks for Data-driven Modelling of a Tube Mill Using Automatic Feature Extraction

Christian Thiel[1], Carolin Steidl[2], Sarah Johannesmann[2], Bernd Henning[2] ([1]: BENTELER Steel/Tube GmbH; [2]: Universität Paderborn)

Modelling industrial processes requires very specific domain knowledge, is time consuming and often not feasible for complex applications. Meanwhile measuring and storing product and process data is of increasing importance for many companies in the manufacturing industry. To capture every detail in the production process, sensors provide locally and temporally resolved measurements for each produced item, thus producing vast amounts of time series like data. This research aims to use the rich data measured in modern industrial environments to build a data-based model of highly complex processes. This calls for automatic feature extraction methods capable of handling varying length time series for dimensionality reduction as well as a capable model to capture complex relations.

In this work we firstly extend feature extraction techniques based on convolutional autoencoders by adding masking. This enables handling of varying length input features, originating for example from tubes of different lengths or varying process speeds. Secondly an end-to-end trainable network architecture for the task of modelling the resulting wall thickness of a tube after hot reshaping is introduced. To further improve the model performance and reduce overfitting, we use a three-step training approach. Finally, we validate our model using real world data.

Data Analysis in Industrial Automation 2

Predictive Quality Control in Industrial Production Processes

Adalbert Wilhelm (Jacobs University Bremen)

This presentation summarizes our experiences in developing predictive models for quality improvement in industrial production processes. Covering all stages from data preparation over data analysis to communicating the results we will focus on evaluative comparisons of different prediction techniques. We will mainly look at ensemble methods, support vector machines, neural networks and other machine learning tools. Besides presenting various ideas of comparing global measures, like accuracy and the area under the ROC curve, we also illustrate the use of the LIME – local interpretable model-agnostic explanation – approach. We will specifically focus on two aspects: how to best assess the specific requirements of highly imbalanced data and how to intertwine the knowledge and expectations of domain experts with the statistical characteristics of the classification approaches. Participants will get deeper insights into the functioning of machine learning, into approaches of evaluative comparisons between machine learning results, and the alignments of domain expert knowledge with data-generated information.

Using Errors-in-Variables Regression to Model and Analyze Sequential Process Chains

Oliver Meyer[1], Claus Weihs[1] ([1]: TU Dortmund University)

A process chain comprises a series of sequential (production) process steps, mostly in the area of manufacturing engineering. It describes a consecutive sequence of activities, which together form one single system. Within this system the sub-processes are presumed to influence each other by transferring characteristics. The single process steps of such a system can easily be simulated using regression (or other statistical learning) methods. The main challenge in simulating entire process chains, however, is the handling of prediction uncertainty in the transferred characteristics. As we have shown earlier (see Meyer & Weihs 2016), this can be studied by using Errors-in-Variables models instead of ordinary regression models. Furthermore, this approach can be used to identify and understand process-product-interactions within a process chain since prediction uncertainty, at least in part, represents real fluctuation in the quality of the intermediate and final products. Understanding these kinds of interactions is an important topic especially for new technologies which are subject to complex process production chains like high energy lithium-ion battery cells (see Westermeier et al. 2013).

In this paper, we want to discuss how to use higher dimension Errors-in-Variables Regression models to accurately simulate process chains. We will especially focus on how uncertainty (measured by variance) develops through the process chain and how it influences the results along the process chain. Based on this we will present a method to adjust the Errors-in-Variables Regression Models, if necessary, to ensure unbiased estimations along the process chain and discuss how the presented methods can be applied to other statistical learning techniques.

Keywords: Process Chains, Errors-in-Variables Regression, Uncertainty Development, Process-Product-Interactions.

References

Meyer, O., Weihs, C. (2016): Statistical Analysis of Sequential Process Chains based on Errors-in-Variables Models, In: Proceedings of the IEEE Symposium Series on Computational Intelligence, ISBN: 978-1-5090-4240-1.

Westermeier, M., Reinhart, G., Zeilinger, T. (2013): Method for Quality Parameter Identification and Classification in Battery Cell Production Quality Planning of Complex Production Chains for Battery Cells, 2013, In: 3rd International Electric Drives Production Conference, EDPC-Proceedings, 1-10. 10.1109/EDPC.2013.6689742.

Kriging and FANOVA Graphs for Computer Experiments with Both Continuous and Categorical Inputs

Dominik Kirchhoff (Dortmund University of Applied Sciences and Arts)

This talk deals with several approaches to incorporate categorical input variables into Kriging models, which can be used for sensitivity analyses – e.g. with functional analysis of variance (FANOVA) graphs – or meta-model-based optimization, especially in the framework of the so-called Efficient Global Optimization (EGO) algorithm.

The original Kriging model can only cope with purely continuous input variables. Since many applications include also categorical variables, we consider some extensions that are able to deal with this case.

First, a brief overview of different existing and new approaches is given.

We then show how FANOVA graphs can be combined with meta-models. These graphs visualize the interactions of input variables and can for example be used to split and parallelize an optimization problem.

Another application can be to explore the effects and interactions of different levels of the categorical variables.

Data Analysis in Medicine and Health Care and Ecology

Analyses of Serum Fatty Acids and Vitamin D with Dimension Reduction Methods

Yifan Chen[1], Rojeet Shrestha[1], Zhen Chen[1], Hitoshi Chiba[2], Shu-Ping Hui[1], Emiko Okada[3], Shigekazu Ukawa[3], Takafumi Nakagawa[4], Koshi Nakamura[1], Akiko Tamakoshi[1], Yuriko Komiya[1], Hiroyuki Minami[1], Masahiro Mizuta[1] ([1]: Hokkaido University; [2]: Sapporo University of Health Sciences; [3]: National Institute of Health and Nutrition; [4]: The Hokkaido Centre for Family Medicine)

The study of fatty acids (FA) is important for nutrition and disease prevention. Because of the diversity and complexity of FA in structures, metabolism and food resources, it is difficult that we interpret FA in epidemiological studies. Here, we applied dimension reduction methods, e.g. principal component analysis (PCA), factor analysis, independent component analysis (ICA) to a dataset of FA and vitamin D obtained from a general population in Suttsu town of Hokkaido, Japan.

A total of 545 participants aged 35 to 79 years, 300 women and 245 men, provided their basic information, including date of participation, gender and age, and serum samples. The FA data consisted of two major types of FA: free FA containing 12 FA subtypes and total FA containing 16 FA subtypes. Dimension reduction methods were employed as general preparation for further statistical analysis of FAs.

We would like to introduce the result of the PCA. The result was generally divided into two parts: outcome of free FA dataset and outcome of total FA dataset. We found that the values of total FA 14:0, 16:0, 18:0, and 18:1 were high in size factor. As well, the values of free FA 14:0 to free FA 18:3 were high in size factor. We also found that the second principal component of total FA dataset could be explained as axis related to long-chain saturated FAs, for its loading values of FAs 20:0, 22:0, 24:0 and 26:0 were very high. Since these FAs are metabolically stable, the second axis might be a stability factor. Furthermore, FAs 20:5, 22:6 and Vitamin D seemed to have a special relationship in both data sets. Some of the results of the factor analysis and the ICA are almost same as that of the PCA.

In conclusion, dimension reduction methods are useful for analysis of FA data in epidemiological studies

Graph Theoretical Network Analysis for Diagnoses Based on Comorbidities

Reinhard Schuster[1], Timo Emcke[2] ([1]: Medizinischer Dienst der Krankenversicherung Nord; [2]: Kassenärztlichen Vereinigung Schleswig-Holstein)

Multimorbidity is a growing challenge for medical treatment against the background of demographic changes. Comorbidities have an impact on therapeutic decisions, while guidelines focus on single diseases. We analyze, which other diseases have the highest increase in probability if a certain disease occurs. Diseases are being set in relation to age and sex structures. As a first step we consider patients with an age between 65 and 69 years only.

We utilize the International Statistical Classification of Diseases and Related Health Problems [ICD-10-GM] on the three-character level. The analyzed datasets contain a pseudonymized patient-ID with age and sex information which have a quarterly resolution differentiating between acute and permanent diagnoses as well as levels of confidence. We analyze diagnostic data of all patients of the statutory health insurance of the most Northern Federal State of Germany (Schleswig-Holstein) from quarter 4/2017. There are 126,848 patients in the considered age category from overall 2.045 million patients treated in that quarter. On average the patients who are 65-69 have 10.17 diseases at the considered ICD level. We will analyze only those diagnoses with appear at least 1,000 times and combinations of diagnoses which occur at least 50 times in order to avoid random effects.

We consider the diagnoses as nodes of a graph and the connections of two diagnoses as graph edges represent the increased probability ordered by the amount of increase.

If we use the top n diagnoses that way, we get an ordered top n graph and a related unordered graph, which we will consider hereinafter. The top 1 graph has 12 components forming a tree structure. The largest one has 60 nodes and 60 edges. The determination of graph community structures using the modularity method of Mathematica generates 19 components.

The top 3 graph is connected with 239 nods, 639 edges and diameter of 7. It has 10 graph community structures with 61,61,35,20,19,14,10,8,6,5 nodes which form a diagnose structure having regard to comorbidities. The two largest community structures have 163 and 158 edges, and both have a graph diameter of 4. The largest cluster has four diagnoses as graph center: “Dorsalgia” (M54), “Injury of unspecified body region” (T14), “Somatoform disorders” (F45) and “Other functional intestinal disorders” (K59), the graph periphery consists of the diagnoses “Symptoms and signs involving emotional state” (R45), “Sleep disorders” (G47) and “Other rheumatoid arthritis” (M06). The second largest graph community structure has only one diagnosis as a graph center (“Essential (primary) hypertension”, I10), but 19 diagnoses as periphery. The smallest structure contains of five diagnoses from the Z-chapter of ICD “Persons encountering health services for examination and investigation” (Z00, Z23, Z25, Z26, Z27).

Graph methods lead to new combinations of diagnoses which are potentially of high interest with respect to treatment interactions. One can analyze those effects with age group and sex stratifications or adjustments. The procedure can also be applied to diagnoses completely classified in order to identify the specific potential of treatment interactions.

Comparison of Methods for Predicting High-cost Patients Captured Within the Oncology Care Model (OCM): A Simulation Study

Madhu Mazumdar[1], Jung-Yi Lin[1], Kavita Dharamrajan[1], Wei Zhang[1], Mark Liu[1], Mark Sanderson[1], Luis Isola[1], Liangyuan Hu[1] ([1:] Icahn School of Medicine at Mount Sinai)

The Centers for Medicare & Medicaid Services (CMS) developed the Oncology Care Model (OCM) as an episode-based payment model to encourage participating practitioners to provide better care at a lower cost for Medicare beneficiaries with cancer. CMS defined 6-month episodes for OCM that begin with the receipt of outpatient non-topical chemotherapy for cancer. The CMS risk-adjustment model includes 16 variables including age, gender, disease, and treatment.

Usually a log-link Gamma generalized linear model (GGLM) is used for estimating concurrent cost. However, Advanced Supervised Machine Learning Method (ASMLM) and Partially Linear Additive Quartile Regression (PLAQR) have recently emerged as modeling approaches with higher flexibility.

We conducted extensive simulations to examine the performance of these methods under various realistic scenarios. The methods are compared using the root mean square error (RMSE), classification accuracy (CA), and patient cost accuracy (PCA) as the performance metrics. The methods were also applied to the data on 4205 OCM episodes collected in 2012-2015 from the Mount Sinai Health System. We provide guidance on use of each method under specific setting.

Illuminating Interactions in Microbial Lake Ecology Using General Learning Chain Ensembles

Theodor Sperlea[1], Daniela Beißer[2], Jens Boenigk[2], Dominik Heider[1] ([1]: Philipps-Universität Marburg; [2]: University of Duisburg-Essen)

In the recent years, legislation such as the EU's Water Framework Directive is putting a focus on the issue of anthropogenic stressors heavily affecting aquatic biomes (Hering et al., 2010). Impacting both biotic as well as abiotic factors of ecosystems, these anthropogenic stressors lead to a global decline of ecosystem quality and increase the probability of extinctions. In order to slow down, halt, or even revert this process, it is necessary to develop methods to assess the status of an ecosystem as well as to identify the requirements that different organisms pose on their surroundings.

In the context of microbial lake ecology, it is straightforward to log and count the microorganisms in a water sample based on their genomic sequences (Tan et al., 2015; Grossmann et al., 2016). However, the connections between the number of these microorganisms and the status of their surroundings remain mostly unknown.

In the current study, we introduce General Learning Chain Ensembles (GLCEs), that are able to efficiently model multi-target datasets containing inhomogeneous data types. We further show that GLCEs are able to model the complex interactions between biotic and abiotic factors in microbial ecology. From these models, we were able to gather information on which parameters are important for the well-being of certain species as well as on microscopic bioindicators for water quality.

References

Grossmann, L., Jensen, M., Heider, D., Jost, S., Glücksman, E., Hartikainen, H., Mahamdallie, S. S., Gardner, M., Hoffmann, D., Bass, D., Boenigk, J. (2016): Protistan Community Analysis: Key findings of a Large-scale Molecular Sampling. The ISME Journal , 10(9), 2269–2279.

Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C. K., Heiskanen, A.-S., Johnson, R. K., Moe, J., Pont, D. (2010): The European Water Frameworkdirective at the Age of 10: A Critical Review of the Achievements with Recommendations for the Future. Science of The Total Environment, 408(19), 4007–4019.

Tan, B., Ng, C., Nshimyimana, J. P., Loh, L. L., Gin, K. Y.-H., Thompson, J. R. (2015): Next-Generation Sequencing (NGS) for Assessment of Microbial Water Quality: Current Progress, Challenges and Future Opportunities. Frontiers in Microbiology, 6.

Data Analysis in Psychology

Representing and Analyzing the Tripartite Emotion Expression and Perception Model with Conditional Linear Gaussian Bayesian Networks

Eva Endres[1], Thomas Augustin[1], Klaus Scherer[2] ([1]: LMU München; [2]: University of Geneva)

The tripartite emotion expression and perception (TEEP) model is an extension of Brunswik's lens model, which represents the vocal communication of emotion. On the one hand, it consists of four components displaying the expressed emotion, distal cues (acoustic measures), proximal percepts (subjective voice quality ratings) and the perceived emotions. On the other hand, it describes three phases between these components, illustrating the encoding, the transmission and the decoding processes of the corresponding emotion.

The architecture of this theoretical model gives reason to use graphical models for statistical analyses. First attempts were made by Bänziger, Hosoya and Scherer (2015, PLoS ONE) who apply path models for the empirical analysis of the emotion families anger, fear, happiness, and sadness. They investigate two data files originating from Munich and Geneva where German- and French-speaking actors enacted all the listed emotions in a controlled experiment.

Within this contribution, we embed the TEEP model into the framework of probabilistic graphical models. This kind of statistical models represents random variables (the components of the TEEP model) by nodes and dependencies among them by edges in a graph (phases of the TEEP model). The graphical representation yields a factorization of the joint distribution of all involved variables and forms the basis for all subsequent analyses, including for instance the computation of the marginal or posterior distributions (belief updating). Using conditional linear Gaussian Bayesian networks, we are able to visualize the process of vocal emotion communication in our statistical model and refine the theoretic knowledge about this process. For this purpose, we combine expert knowledge about the relations of the components of the TEEP model and data-driven learning algorithms to find more information about direct dependencies among these components within the different phases. Thereby, we also identify which objectively measured distal cues provide information about the subjectively measured proximal percepts. Furthermore, we achieve conditional linear Gaussian Bayesian networks for all emotion families, helping to develop a deeper understanding of the vocal emotion communication in future research projects. All our analyses are conducted on the data from Munich and Geneva and compared to the previous results by Bänziger, Hosoya and Scherer (2015, PLoS ONE).

What Drives AfD Election Success?

Sebastian Sauer[1], *Oliver Gansser*[1] ([1]: FOM Hochschule)

The party “Alternative für Deutschland” (AfD) is a right-wing to far-right political party in Germany. Founded in 2013 it has managed to gain seats in all states parliaments as well as in the federal parliament within a short period of time. The AfD became the third-largest party in Germany after the 2017 federal elections. Such striking success is quite unique in German post-war politics. Consequently, AfD’s upsurge has sparked an intensive debate as to the why´s and how’s of this success. Some explanations of AfD’s electoral success have been brought forward by scholars but also some “folk theories” circulate. In this talk, we test some folk theories highlighting potential causes of AfD’s electoral success such as unemployment, migration rate, age, and east/west cultural differences. Our data are based on the German federal election results (2017), alongside with structural data on each German electoral district (n=299). Our analysis is novel insofar as a more rigorous Bayesian multilevel modeling is applied. In addition, we include large-scale human behavioral data (n = 22,000) which was not available in previous studies. In sum, our result provides little evidence for the validity of typical folk theories. We interpret the results that such folk theories should be abandoned. Moreover, value driven human behavior trait did not explain much election behavior either. It appears that voting success, particularly with regard to the AfD, is still far from being understood properly. Explanations for this finding are discussed alongside with recommendations for future research.

A Single Peaked Multivariate Logistic Distance Model

Mark de Rooij (Leiden University)

We propose a multivariate logistic distance model where the probability of answering “yes” to an item depends on the Euclidean distance between a person position and an item position in a low dimensional space. Such a model gives rise to single peaked response functions. We contrast the model to the model of Worku and De Rooij (2018, Journal of Classification) where the probability is modelled in terms of the Euclidean distance of a person towards the categories of the response item. The latter gives rise to monotone increasing response function. For the new model, we show how to fit it to data using a quasi-likelihood function and present some ideas about model selection. Finally, an application will be presented based on the Dutch Parliamentary Election Study.

Data Analysis in Social Sciences 1

Decision Support System for Road Safety Improvement

Katharina Meißner (University of Hildesheim)

The Valletta Declaration on Road Safety 2017 by the EU transport ministers proclaims that the amount of people killed by road accidents shall be halved until 2020. To reach this goal, it is necessary to evaluate the circumstances of accidents and their change over time carefully. Only in this way, the police is enabled to decide upon targeted actions to improve road safety, e.g., new speed limit reductions, stop signs or investments in walking and cycling infrastructure. In order to assist the police, we propose a decision support system based on (i) frequent itemset mining, (ii) time series clustering, and (iii) change detection. Due to the layered structure of this system, we are able to optimize the single steps individually by considering the previously obtained results.

The statistical data used, depicting ten years of road accidents that entailed personal injuries, includes 65 attributes describing the accidents' external circumstances (e.g., time of day, weather, road condition). Firstly, we apply frequent itemset mining on these attributes to find frequent combinations and calculate the support for each itemset for each month. In this way, we generate thousands of possibly interesting attribute combinations. A combination of attributes (i.e., frequent itemset) is interesting or valuable, if its relative frequency varies from one month to the other, where a slow inclining or declining trend, an arbitrary change of direction, or any other kind of instability may occur. Since frequent itemset mining methods lead to numerous potentially interesting attribute combinations, an automated change mining approach is needed in order to find the most interesting ones. Therefore, a clustering of time series is performed in the second step in our system. Due to the amount of time series, we only cluster a suitable sample and classify the remaining time series according to the clusters found. In order to obtain the optimal parameters for clustering, we use multiple samples and consider several clustering methods, numbers of clusters, distance methods, and dimensionality reduction techniques. For every generated cluster of time series, we are then able to determine the optimal forecasting method and can therefore provide predictions on the future frequencies of a certain attribute combination. Based on this information, interesting attribute combinations can be presented on maps indicating hotspots concerning locations and traffic systems that help police analysts establishing further action plans to reduce road accidents in the future.

Clustering and Modeling Data. A Quantile Regression Approach

Cristina Davino[1], Domenico Vistocco[1] ([1]: University of Naples Federico II)

This paper exploits quantile regression to identify groups of units with a different dependence. The best model for each group will be estimated and inferential procedures will allow us to test if group structures are statistically different.

The conceptual premise of the work is represented by the consideration that the relationship between a response variable and a set of explanatory variables can be different if units belong to different groups. It is a matter of fact that if two units have similar features/behaviours, the dependence structure of a regression model is more alike.

The proposed approach is mainly based on the undoubted potentiality of QR to explore the entire conditional distribution of the response variable and it aims to discover groups in a dependence model and to identify the best model for each group. The approach can represent a valid tool to cluster units according to the dependence structure without a priori information but only using the observed similarities among them in terms of conditional quantile estimates.

The final results are easy to read and interpret as the coefficients associated to each group follow the same interpretation of any linear model; furthermore, the best quantile assigned to each group synthesizes the main location of the conditional distribution of the response variable where the group has an effect on. Classical inferential procedures can be used to compare the models because the group effects are identified using the whole sample.

Data Analysis in Social Sciences 2

Algorithmic Sources of Publication Bias in Political Science Research

Alrik Thiem[1], Lusine Mkrtchyan[1], David Sanchez[1] ([1]: University of Lucerne)

Despite recent attempts at improving research transparency, meta-analyses in political science continue to demonstrate the pervasiveness of publication bias. In this article, we reveal a hitherto undiscovered source of such bias. More specifically, we demonstrate why the uncritical import of optimization algorithms from electrical engineering into political science research employing the configurational method of Qualitative Comparative Analysis (QCA) has created publication bias on a large scale over the last quarter century.

Drawing on replication material for 145 peer-reviewed journal articles, we measure the extent this problem has assumed in empirical political science. We also present a solution that is guaranteed to eliminate this source of bias in future applications of QCA. More generally, our findings emphasize the importance of thoroughly evaluating the adequacy of foreign methods of data analysis before putting them to uses which they were not originally designed for.

The Involvement in Politics via the Social Media Channels: A Multivariable Analysis

Evangelia Nikolaou Markaki[1], Theodore Chadjipantelis[1] ([1]: Aristotle University of Thessaloniki)

With the use of conjoint analysis as well as of other multivariable analysis methods, we assess how the involvement in politics via the social media channels influences political engagement as well as the formation of political preferences. We used the scenario technique and we created an experiment so as to evaluate at the same time the influence of different factors on voting behavior and political engagement. Conjoint Analysis represents a hybrid type of technique to examine dependent relations and combines methods such as Regression or Anova permitting researchers to depict a person's preference about a concept, an idea or a product taking into account different characteristics or factors. The present technique analyses the components of the total preference where the researcher can estimate the relative importance for each characteristic or factor. These characteristics - factors are pre-defined by the researcher. The relative importance of each characteristic - factor shows its contribution to the "total preference" We also use a two-step procedure, computing firstly, via multivariate correspondence analysis, the principal axes and loadings and secondly, through cluster analysis, the attitudes that are grouped in clusters. Through this analysis, specific axes emerged, describing the data in less dimensions.

The Correspondence Analysis is a statistical method for the representation of rows and columns of a data table in a space of fewer dimensions than the original. Analyzing data in a space of fewer dimensions can reveal typological patterns of data and group the data into homogeneous clusters. This is a two-step process. The analysis is implemented through the use of two-way cross tabulation, contingency tables, and correspondence analysis by using the pioneer program "M.A.D." [Méthodes de l'Analyse des Données].

The involvement in politics is directly connected to the involvement in political competition. It refers to the way people choose to be in touch and involved in politics. Some of them choose an energetic action and dynamic support via physical presence or financial support. Some others prefer a passive observation of politics.

The present study examines also the use of new technologies in political behavioral analysis. Web 2.0 & 3.0 give people the opportunity to develop new ways of communication, interaction, diffusion of information, social connections, social involvement, and access to information and to entertainment. Many people use internet and the social media which during the last years were transformed to important factors of social activity.

The new media are the new context for communication and for social dialogue, where the information management and diffusion is open. The studies show that the influence exerted in social media is related to the political knowledge. The different types of social networking media (Facebook, Twitter,

YouTube, Myspace, Blogs) are involved in political marketing but each of them has its own penetration to people.

The use of these results as well as of the methodology can be widely used for strategic political marketing campaigns and help professionals to use the political agenda in an effective way.

Visualization of Latent Higher-order Interactions by Regions of Significance

Viktor Fredrich[1], Ricarda Bouncken[1] ([1]: University of Bayreuth)

Data visualization provides an easy way to communicate core research findings through effective graphics as it is easier for the brain to comprehend images versus words or numbers (Cleveland and McGill, 1984). A commonly accepted and well-established way to present results of moderator hypotheses includes simple slope analysis from multiple regression or structural equation modeling with latent measures. A series of freely available online tools (Dawson, 2014) allows the plotting of average main effects with one moderator (2-way), two moderators (3-way) or even non-linear moderators (e.g., quadratic 3-way). More recently, scientific research started to focus on regions of significance for marginal effects (i.e. confidence intervals for first derivatives of relevant parameters, Bauer and Curran, 2005). Particularly the increasing popularity of latent (i.e. imperfectly reliable) measures limits the application of traditional XY plots of means (–/+1 S.D.) due to their lack of absolute representation. The dominance of these traditional plots is persistent and under certain circumstances even misleading (e.g., out-of-sample extrapolation for highly skewed distributions; potentially non-significant aggregation of partially significant and insignificant parameters; assumption of constant standard errors for all values of moderators). We therefore suggest applying regions of significance to any higher-order latent or manifest interaction. We develop 2D-profiles of (linear or non-linear) moderators revealing regions of significance for a dependent variable (that could be a mean, an effect or any higher-order interaction itself) for all empirical combinations of two moderators.

This advanced visualization technique improves the precision of practical implications by revealing otherwise hidden information. Alternatively, we advocate testing the robustness of the identified regions of significance for parametrically nested models in order to avoid common statistical pitfalls (e.g., explicitly testing oftentimes unrealistic linearity assumptions of rather monotone relationships, Ganzach, 1998; avoiding statistical significance of parameters as artifacts of uncontrolled higher-order parameters, Cronbach, 1975; and thereby reducing typically unreported model specification bias, Kline et al., 2000).

References

Bauer, D.J., Curran, P.J. (2005): Probing Interactions in Fixed and Multilevel Regression: Inferential and Graphical Techniques. Multivariate Behavioral Research 40 (3), 373-400.

Cleveland, W.S., McGill, R. (1984): Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association 79 (387), 531-554.

Cronbach, L.J. (1975): Beyond the two Disciplines of Scientific Psychology. American Psychologist 30 (2), 116-127.

Dawson, J.F. (2014): Moderation in Management Research: What, Why, When, and How. Journal of Business and Psychology 29 (1), 1-19.

Ganzach, Y. (1998): Nonlinearity, Multicollinearity and the Probability of Type II Error in Detecting Interaction. Journal of Management 24 (5), 615–622.

Kline, T.J., Sulsky, L.M., Rever-Moriyama, S. (2000): Common Method Variance and Specification Errors: A Practical Approach to Detection. The Journal of Psychology 134 (4), 401-421.

Data Analysis Models in Economics and Business 1

Selecting the Optimal Multidimensional Scaling Procedure for Interval-valued Data with Symbolic-to-classic and Symbolic-to-symbolic Approaches

Marek Walesiak[1], Andrzej Dudek[1] ([1]: Wroclaw University of Economics)

In multidimensional scaling (MDS) carried out on the basis of interval-valued data table two approaches can be distinguished: symbolic-to-classic and symbolic-to-symbolic. The starting point is the data table in which each of the objects is described by m interval-valued variables. The article presents the mdsOpt package of the R program which helps to solve the problem of choosing the optimal multidimensional scaling procedure according to the normalization methods and iterative optimization algorithms in symbolic-to-symbolic approach and normalization methods, distance measures and MDS models in symbolic-to-classic approach. It uses two criteria for selecting the optimal multidimensional scaling procedure: Kruskal's Stress"-" 1 fit measure (I"-" Stress in case of symbolic-to-symbolic approach) and Hirschman-Herfindahl HHI index calculated based on Stress per point (Interval stress per box in case of symbolic-to-symbolic approach) values. The results are illustrated by an empirical example.

Keywords: Interval-valued Data, Multidimensional Scaling, Stress and I-Stress, Normalization of Variables, HHI Index.

References

Billard, L., Diday, E. (2006): Symbolic Data Analysis: Conceptual Statistics and Data Mining, John Wiley, Chichester. ISBN: 978-0-470-09016-9.

Borg, I., Groenen, P.J.F., Mair, P. (2018): Applied Multidimensional Scaling and Unfolding, Springer, Heidelberg, New York, Dordrecht, London. ISBN 978-3-319-73470-5. URL https://doi.org/10.1007/978-3-319-73471-2.

Groenen, P.J.F. Winsberg, S., Rodriguez, O., Diday, E. (2006): I-Scal: Multidimensional Scaling of Interval Dissimilarities, Computational Statistics & Data Analysis, 51(1), 360–378. URL http://dx.doi.org/10.1016/j.csda.2006.04.003.

Herfindahl, O.C. (1950): Concentration in the Steel Industry, Doctoral thesis, Columbia University.

Hirschman, A.O. (1964): The Paternity of an Index, The American Economic Review, Vol. 54, No. 5, 761-762. URL http://www.jstor.org/stable/1818582.

Mair, P., Borg, I., Rusch, T. (2016): Goodness-of-fit Assessment in Multidimensional Scaling and Unfolding, Multivariate Behavioral Research, Vol. 51, No. 6, pp. 772-789. URL http://dx.doi.org/10.1080/00273171.2016.1235966.

Mair, P., De Leeuw, J., Borg, I., Groenen, P. J. F. (2018): smacof: Multidimensional Scaling. R package ver. 1.10-8. URL https://CRAN.R-project.org/package=smacof.

R Core Team (2018): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Terada, Y., Groenen, P.J.F. (2015): smds: Symbolic Multidimensional Scaling. R package version 1.0. URL https://CRAN.R-project.org/package=smds.

Walesiak, M., Dudek, A. (2018a): clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package, version 0.47-3. URL https://CRAN.R-project.org/package=clusterSim.

Walesiak, M., Dudek, A. (2018b): mdsOpt: Searching for Optimal MDS Procedure for Metric and Interval-valued Symbolic Data. R package, version 0.3-3. URL https://CRAN.R-project.org/package=mdsOpt.

Risk Factor Analysis of Companies Listed on the Warsaw Stock Exchange

Lula Paweł[1], Paweł Cabała[1], Renata Oczkowska[1], Jakub Kanclerz[1] ([1]: Cracow University of Economics)

The analysis of essential risk factors associated with companies listed on the Warsaw Stock Exchange is the main goal of the presentation.

The research process involved two phases. During the first stage of the analysis main risk factors were identified. This problem was solved by using an ontology-based approach for analysis of company's prospectuses. The second stage of the research was focused on the analysis of risk factors

frequencies, their co-occurrences, and their relationships with economy sectors. The authors also performed cluster analysis of prospectus chapters containing risk factor description.

All algorithms used for risk factors analysis were implemented in R language.

Fuzzy Representation of Linguistic Measurement in Risk Perception, Preferences and Attitudes Classification

Jozef Dziechciarz[1], Marta Dziechciarz-Duda[1] ([1]: Wroclaw University of Economics)

The purpose of the presentation is to discuss the problem of ambiguity in the results of linguistic measurements and its fuzzy representation. Measurement of socio-economic phenomena, including risk perception, preferences and attitudes along with the perception of the material wealth of households may be conducted with the use of linguistic scale. Results of such measurement need to be transformed into quantitative representation.

Fuzzy numbers proved to be very helpful in representation and effective processing of imprecise (linguistic) information. For in-depth analysis, however, it is necessary to approximate given fuzzy measurement result using metric representation. For this purpose, a process of de fuzzing (sharpening) is usually used. For approximation, the triangular or trapezoid representation of fuzzy numbers may be used. Further variation may include the use of fuzzy symmetric and asymmetric representation, numbers of uneven length as well as unbalanced and overlapping shape. Discussion of various types of representation of fuzzy results is given. The analysis of the notions of the distance and orders between fuzzy numbers based on these representations is provided.

The problem of linear ordering and classification of phenomena measured with linguistic scale, with results converted into fuzzy representation is introduced and investigated. The comparison of results in decision making problems under fuzzy environment is discussed.

Keywords: Linguistic Scale, Fuzzy Representation, Attitudes, Risk Perception, Preferences, Classification.

Data Analysis Models in Economics and Business 2

Logit Leaf Model in Prediction of Corporate Bankruptcy

Barbara Pawełek[1], Józef Pociecha[1], Sebastian Grabarz[1] ([1]: Cracow University of Economics)

Data classification methods are frequently used for bankruptcy prediction. The logit leaf model is a new hybrid classification algorithm that enhances logistic correct regression and decision tree. The logit leaf model consists of two stages. In the first stage company sets are identified using the decision tree. In the second stage, the logit model is created for every leaf of this tree.

The aim of our study is to present the results of research on the usefulness of the logit leaf model for bankruptcy prediction. The added value of the work is the use of the logit leaf model in forecasting of corporate bankruptcy on the basis of the set of data cleared from outliers observed for firms continuing business.

The study uses the 64 financial indicators for industrial processing sector in Poland. The following methods were applied: logit leaf model, decision trees, logistic correct regression, random forests, logistic model trees. To measure the predictive accuracy of the models, the following measures have been used: overall accuracy, sensitivity, specificity and AUC. Calculations were performed in R.

References

De Caigny, A., Coussement, K., De Bock, K. W. (2018): A new Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees. European Journal of Operational Research, 269(2), 760–772, DOI: https://doi.org/10.1016/j.ejor.2018.02.009

De Caigny, A., Coussement, K., De Bock, K. W. (2018): LLM: Logit Leaf Model Classifier for Binary Classification. R package version 1.0.0., https://CRAN.R-project.org/package=LLM

Hornik, K., Buchta, C., Zeileis, A. (2009): Open-Source Machine Learning: R Meets Weka. Computational Statistics, 24(2), 225–232, DOI: https://doi: 10.1007/s00180-008-0119-7

Landwehr, N., Hall, M., Frank, E. (2005): Logistic Model Trees. Machine Learning, 59(1–2), 161–205, DOI: https://doi.org/10.1007/s10994-005-0466-3

Witten, I. H., Frank, E. (2005): Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco.

Empirical Bayes Method for the Estimation of Wages of Small Enterprises

Grażyna Dehnel[1], Łukasz Wawrowski[1] ([1]: The Poznań University Of Economics And Business)

The paper presents an empirical study, which was designed to test a non-classical small area estimation (SAE) method in assessing one of the parameters of entrepreneurship.

The aim of the study is to estimate average monthly wage in small enterprises at the low level of disaggregation using empirical Bayes method. The domain of interest is a unit resulting from joint cross-classification by district and economic activity category (NACE). The study involved enterprises employing from 10 to 49 employees and relied on data from a survey conducted by the Statistical Office in Poznan and administrative registers.

In Poland, sample surveys are considered the most important source of information about small enterprises. Only classical methods of estimation are used to produce monthly estimates of the parameters of entrepreneurship. Information, based on the received results, is published only at country and provinces level (including NACE sections division).

For lower level of disaggregation, direct estimates are unstable in the sense of having very large sampling errors for units with small sample size. However, an access to a variety of information about local economic condition is required for further development and support of entrepreneurship. Therefore, an attempt was made to estimate the monthly wages of employees for a more detailed domain of interest. Using Empirical Bayes method of indirect estimation, the study is expected to provide information about differences and characteristics of the distinguished areas.

The Smart Development of German Government Regions (Regierungsbezirke) 2005-2017

Andrzej Sokołowski[1], Małgorzata Markowska[2] ([1]: Cracow University of Economics; [2]: Wrocław University of Economics)

After some administrative changes, Germany is currently divided into 38 statistical regions of NUTS 2 level. The aim of this study is to measure the smart development level of these regions in 2008-2017. Smart development goals are defined for the whole European Union and also nation wise. There are four variables defined by Eurostat to measure the smart development level:

- Employment rate for age group 20-64 (as % of population aged 20-64):
  EU goal 75; German goal 77
- Gross domestic expenditure on R&D (as % of GDP):
  EU goal 3; German goal 3
- Early leavers from education and training (as % of population aged 18-24):
  EU goal 10; German goal 10
- Tertiary educational attainment (as % of population aged 30-34):
  EU goal 40; German goal 42

Special trimmed standardization procedure is used in the paper, to avoid the borrowing of effects between variables. The composite indicator measuring of the smart development is a newly proposed method which aggregates four different approaches, weighted and unweighted. Positions of German government regions in the set of 263 European NUTS 2 regions are shown. Time series of composite indicators are clustered in order to find typical paths toward smart development goals. Coherence is analysed through the trend of standard deviation. Finally, forecasts for achieving smart development goals in Germany are presented.

Data Science Education

Causal Modelling for Data Literacy - In Intro Stats?

Karsten Lübke[1], Matthias Gehrke[1], Jörg Horst[2], Sebastian Sauer[1] ([1]: FOM University of Applied Sciences; [2]: Bielefeld University of Applied Sciences)

Data literacy has been defined as "the ability to collect, manage, evaluate, and apply data, in a critical manner" (Ridsdale et al., 2015). But data literacy is nothing new to statistical education; as authors such as Gould (2017) put it: Data literacy "is statistical literacy".

Through the "Focus on conceptual understanding" (GAISE, 2016) by teaching techniques such as Simulation Based Inference (Bootstrapping, Permutation Test) and by de-emphasizing more traditional inference techniques, free space is added to the curriculum of introductory statistics and modelling can be used as a leitmotiv in teaching statistics (Stigler & Son, 2018).

However, the results of multivariate data modelling can be misleading through the presence of confounding variables such as in the well-known Simpson's or Berkson's Paradox.

The basic ideas of Causal Inference like, such as a) using Directed Acyclic Graphs, b) highlighting the difference between observing and manipulating data, and c) counterfactual evaluation, may foster a deeper understanding of what can and – maybe even more important – what cannot be deduced from the analysis of (observational) data. Moreover, knowledge of ideas of causal modelling may help to refrain from over-simplified conclusion based on "Big-Data" analysis.

The notion to integrate causal modelling in introductory statistics is supported by many, e.g. Ridgway, 2016, Angrist and Pischke, 2017, Kaplan, 2018, or the ASA "Causality in Statistics Education Award".

In this talk, we discuss how modelling and causal modelling may help students to think with and beyond data in introductory statistics courses. In addition, we will provide an instructional rationale and some preliminary evaluation data will be presented.

References

Angrist, J.D., Pischke, J.S. (2017): Undergraduate Econometrics Instruction: Through our Classes, darkly. Journal of Economic Perspectives 31(2), 125–144.

GAISE College Report ASA Revision Committee (2016): Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016.

Gould, R. (2017): Data Literacy is Statistical Literacy. Statistics Education Research Journal, 16(1), 2-25.

Kaplan, D. (2018): Teaching Stats for Data Science. The American Statistician 72(1), 89–96.

Ridgway, J. (2016): Implications of the Data Revolution for Statistics Education. International Statistical Review 84(3), 528–549.

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., Wuetherick, B. (2015): Strategies and Best Practices for Data Literacy Education: Knowledge synthesis report.

Stigler, J.W., Son, J.Y. (2018): Modeling First: A Modeling Approach to Teaching Introductory Statistics. In: M.A. Sorto, A. White, L. Guyot (eds.) Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics.

Industrial Data Science: Implementation of a Qualification Concept for Machine Learning in Industrial Production

Nadja Bauer[1], Daniel Horn[1], Malte Jastrow[1], Claus Weihs[1] ([1]: TU Dortmund University)

The advent of industry 4.0 and the availability of large data storage systems lead to an increasing demand for specially educated data-oriented professionals in industrial production.

The education of such specialists is supposed to combine elements from three fields of industrial engineering, data analysis and data administration. In order to extract knowledge from the stored data the proficient handling of data analysis tools - especially machine learning - is essential. Finally, industrial domain knowledge is important to identify possible applications for machine learning algorithms. However, a comprehensive education program incorporating elements of all three fields has not yet been established (in Germany).

In the context of the "Industrial Data Science" (InDaS) project we developed a qualification concept for machine learning in industrial production which we presented in detail on the last ECDA conference. Now we aim to provide our experience during implementation of the first part of the concept: an interdisciplinary lecture. Furthermore, we will introduce an elaborated plan for the second part: a practical seminar. This includes a set of real use cases provided through several industrial partners involved in the project. Lastly, we will discuss how the InDaS concept will be established in TU Dortmund University also after the end of the project funding phase.

'Optimal' Data Reduction in Non-Stationary Systems

Jakob Krause (MLU Halle-Wittenberg)

Starting from the premise that economic systems change fundamentally over time and that, consequently, data from the past is only partially representative for the current situation this paper aims to identify data sets yielding minimal bias estimators. I formalise and discuss the trade-off between data quality and data quantity in a system that gradually changes over time and target the question which level of understanding one can achieve in a changing environment. As an application an impact study of a paragraph in the banking regulation framework is performed. I show that this paragraph is incompatible with ergodicity, one of the fundamental assumptions of risk management and economics, and consequently highlight a severe cognitive dissonance that is in the currently active regulations.

Debate: Data Science - Occupation or Profession?

Ursula Maria Garczarek[1], Detlef Steuer[2] ([1]: Cytel Inc.; [2]: Helmut-Schmidt-Universität Hamburg)

In this session we want to have a debate with the audience on the question of the social responsibility of people engaged in “data science”. The ethical challenges arising from the (blind) use or trust in algorithms gets increasing attention in the public debate, so what do people from the existing professions engaged in data science – statistics, computer science, engineers, mathematicians, …- do think about how we want to address them?

To prepare a common ground, the session begins with a talk defining data science (Donoho, 2017) and an introduction to a system of six types of ethical issues around data science (CNIL, 2018), with examples. We will explain that to our perception, those that currently “do” data science are not prepared to take the responsibility to address the challenges, as they do not form a professional community.

We then want to start the debate with the statement: Data science is in the focal point of current societal development. Without becoming a profession with professional ethics, data science will fail in building trust in its interaction with and its much-needed contributions to society!

References

Donoho, David (2017): 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, doi: 10.1080/10618600.2017.1384734

Commission Nationale Informatique & Liberte, Algorithms and Artificial Intelligence (2018): CNIL’s Report on the Ethical Issues, May 25, 2018, https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues. Cited 2 Nov 2018.

Dimension Reduction 1

Redundancy Analysis for Sparse Component Loadings

Yuki Yamagishi[1], Hiroshi Yadohisa[1] ([1]: Doshisha University)

Redundancy analysis (RDA) is a useful method for analyzing the relationship between two sets of multivariate data by solving reduced rank parameters in multivariate regression. This is in contrast to canonical correlation analysis, which analyzes the relationship between two sets of variables but extracts components from each set that is maximally correlated. There are many fields where RDA is used, such as psychology, econometrics, and biology. Besides, RDA has been used in recent years to analyze functional MRI data as a special model of the constrained principal component analysis. This is because with RDA it is possible to analyze the functional networks involved in experimental tasks, by using the activation of voxels as a matrix of dependent variables and the experimental tasks (flagged by 0-1) as a matrix of independent variables. Here, the estimated low-rank parameter matrices are interpreted as principal component scores for each experimental condition and principal component loadings for each voxel.

It takes a lot of effort to interpret the principal component loadings for voxels consisting of 10,000-100,000 columns corresponding to the brain regions. Therefore, a threshold is set after the estimation, where some principal components loadings degenerate to zero if their absolute values are low. However, this procedure is potentially misleading in various respects.

This problem can be solved by introducing regularization penalties in reduced rank parameters. Nevertheless, existing methods are related to the selection of independent variables and methods that estimate the whole parameters as sparse, and it is not possible for the estimate of principal component loadings to be sparse.

Therefore, we propose column-wise sparse RDA (csRDA), which introduces regularization penalties to promote sparsity in principal component loadings to facilitate interpretation. Sparsity and orthogonality are incompatible constraints. Therefore, we impose both the constraints on different matrices for principal component loadings. The proposed method contains RDA and is considered as an efficient iterative algorithm for computation.

Hierarchical Disjoint Principal Component Analysis

Carlo Cavicchia[1], Maurizio Vichi[1], Giorgia Zaccaria[1] ([1]: University of Rome La Sapienza)

Dimensionality reduction has been often considered in the last years due to the use of Big Data. Frequently the process of reduction has a hierarchically nested form which can be represented with a graphical configuration of a tree. Leaves correspond to manifest indicators (MIs), i.e., the portfolio or scoreboard of observed variables, while internal nodes denote components (that is, linear or non-linear combinations of MIs) which synthesize common information in the data. The root of the tree is a general indicator.

The hierarchy is a property which can be attributed to a manifold of different phenomena, from most general to most specific, in which a more general level includes more specific concepts.

In this paper starting from the data matrix X of size (n × J), corresponding to n objects and J quantitative variables, we propose a statistical model for hierarchical parsimonious disjoint dimensionality reduction. This new methodology induces a hierarchical parsimonious system of indicators, each one represented by a component. The hierarchy is defined starting from the DPCA solution with a predefined number of latent variables. The components which take shape in the hierarchy could entail the identification of theoretical concepts, that represent the intermediate levels of the operationalisation phase for the construction of a composite indicator. The model is estimated by using the LS method, optimizing a constrained quadratic problem. Optimal properties, such as uniqueness and identifiability, are investigated. Albeit the HDPCA problem is NP-hard, a coordinate descent algorithm is proposed, and it turns out to be computationally efficient in real case studies. The goodness of fit of the hierarchical parsimonious trees can be computed to assess the quality of the hierarchical partitions

Zero-Inflated Negative Binomial Matrix Factorization

Hiroyasu Abe (Kyoto University)

Nonnegative matrix factorization (NMF) is a matrix decomposition technique that uses a data matrix consisting of nonnegative entries and enables us to understand the hidden structures in data matrices. NMF is often applied to multivariate count data, because the count data matches the nonnegativity constraint of NMF.

Poisson distribution is one of the most commonly used probability distributions for modeling count data and there is an NMF based on the Poisson, namely the Poisson matrix factorization (PMF). If a data matrix has many zero entries, i.e., zero-inflated case, PMF may provide poor approximation results, which do not appropriately reflect the structure in the data matrix. For such zero-inflated cases, the zero-inflated Poisson matrix factorization (ZIPMF) has been proposed in the literature.

ZIPMF is the modified NMF technique using the zero-inflated Poisson model (ZIP); the data is drawn from the degenerate distribution at zero or from the Poisson. However, both PMF and ZIPMF do not account for the over dispersion of count data.

Recently, an NMF based on the negative binomial distribution (NBMF) has been proposed by Gouvert et al. (2018). The negative binomial distribution is compatible with overdispersed count data, and it is often applied in linear regression analysis.

Based on the NBMF technique, we propose the new NMF, namely ZINBMF, which is based on the zero-inflated negative binomial distribution; the data is drawn from the degenerate distribution at zero or from the negative binomial.

Dimension Reduction 2

Dimension Reduction for Qualitative Ideas in Innovation Search

Diana Schif[1]*, Ulrich Theodor Schwarz*[2] ([1]: BMW AG; [2]: Chemnitz University of Technology)

Workshops or meetings for innovation search often result in collections of several hundred ideas. Working with a large number of ideas as rather abstract quantities is difficult. The objective of the method described in this paper is to easily reduce an unmanageable number of ideas to a smaller number without losing essential content. Reducing the quantity without loss of meaning is possible by combining ideas that deal with the same or similar issues. An upcoming problem during a trivial reducing process is to keep track of all ideas. A three-step method proposes a structure for the process. As a first step it is necessary to group the single ideas regarding similarity using an iterative comparison. The objective is to identify and label all groups. The difficulty by grouping the single ideas is that one idea may contain different aspects and can therefore be assigned to multiple groups. In the first step this multiple assignment will not be taken into account, because this will be handled in the next step. For the second step a table is created with the group titles in the horizontal index and the single ideas in the vertical. Now the evaluator has to identify the group/groups for each idea and mark the cell/cells in the table. Next, a filter will be applied for every combination of groups to gather ideas that handle with the same group titles. Last step is now to merge such same or similar filtered ideas to a new idea that integrates all aspects of the single ideas. The result is a three-step strategy that simplifies the reducing process of several ideas. The output of this strategy is a reduced list of merged ideas without loss of essential information. This strategy can reduce ideas to less than 20 percent of the original ones without losing singular ideas of the original set.

Uncovering the Structure of Rank Data by its Coherent Groups

Vartan Choulakian (Université de Moncton)

Let n respondents rank order d items. Our main task is to decompose the n respondents into a finite number of coherent groups, where each coherent group is composed of a finite number of coherent clusters. A coherent cluster is characterized by the fact that its members have the same first TCA

factor score, where TCA designates Taxicab Correspondence analysis, an L_1 variant of correspondence analysis. Furthermore, each coherent cluster is interpreted and visualised by its matrix of first-order marginals. We also quantify the coherency in a group or in a cluster by its crossing index, which measures the extent of crossing of scores of voters between two blocks seriation of the items where the Borda count statistic provides consensus ordering of the items on the first axis. Examples are provided.

Keywords: Borda Count, Coherent Group, Coherent Cluster, First-order Marginals, Mixed Group, Shuffling, Crossing Index, Taxicab Correspondence Analysis, Masking.

A Feature Selection Method Based on the Naive Hotelling T^2 Statistic

Mitsuru Tamatani (Doshisha University)

This talk is concerned with pattern recognition in a High Dimension Low Sample Size context. Tamatani and Naito (2018; Communications in Statistics - Theory and Methods) showed that asymptotic normality of the naive Hotelling T^2 statistic under a High Dimension Low Sample Size setting is developed using the central limit theorem of a martingale difference sequence. On the other hand, Fan and Fan (2008; The Annals of Statistics, 36: 2605–2637) and Tamatani, Koch and Naito (2012; Journal of Multivariate Analysis, 111: 350–367) proposed variable ranking and feature selection methods.

In general, it is well known that high dimensional data include significant noise components that increase the misclassification rate. To avoid this problem, several approaches to feature selection for discrimination have been proposed by Tibshirani et al. (2002; proceedings of the National Academy of Sciences, 99: 6567–6572), and others. In particular, Fan and Fan (2008) introduced criteria for noise reduction using the upper bound of the misclassification rate. Using an estimator of this upper bound, we have a practical objective function for feature selection, and this was in fact used by Fan and Fan (2008) and Tamatani, Koch and Naito (2012).

In this talk, we propose a new criterion for feature selection based on asymptotic normality of the naive Hotelling T^2 statistic under a High Dimension Low Sample Size. In particular, we propose a new criterion to determine the number of features used in practical analysis.

Image and Text Mining

Verifying User Preferences - A Content Based Analysis of Images in an Online Travel Community

Ines Brusch (Brandenburg University of Technology Cottbus-Senftenberg)

The number of images has been increasing for many years. Within three years, the number of images increased from 660 billion to 1200 billion in 2017 (Statista 2018). And also, the images that are posted online are getting more and more. It is not for nothing that social media platforms like Instagram are enjoying great popularity. Today, photography has become part of our lives, including sharing images on social networking sites, sharing images instantly with smartphones, using images in blogs and forums and much more.

Images are particularly widespread in tourism. Other cultures, places, people and of course yourself and special experiences are captured and taken home as a souvenir. So it is not surprising that online travel communities are also very popular. In research on online travel communities there are often questions about the reasons and motivations for the use (Bronner and de Hoog 2011), positive and negative effects of word-of-mouth and its influence on customer loyalty (Sanchez-Franco and Rondan-Cataluña 2010). There are also some studies on the importance of images in travel communities. For example, some authors show that images or social media content can be used as a cost-efficient alternative to surveys to draw conclusions about user preferences (Daniel and Baier 2013, Hausmann et al. 2018). Further studies also deal with the contents of the images to illustrate activities and experiences (Garrod 2009).

However, automatic recognition of the image content does not take place. The plagiaristic question that arises from this is: Are you saying what you are showing? The following paper aims to answer this question. In this respect, two important research questions arise. First, can pictures of tourists be automatically analyzed for scenes? Second: Are there typical holiday pictures that describe different types of tourists and thus make it possible to predict future holiday destinations?

To answer these questions, the paper first addresses the users of online travel communities. Afterwards the possibilities of artificial neurological networks are presented. In the following experiment different content-based image analysis methods are applied and their quality is compared. Finally, it is shown which images are typical for different types of holidays.

References

Bronner, F., de Hoog, R. (2011): Vacationers and eWOM: Who Posts, and Why, Where, and What? Journal of Travel Research, 50(1), 15–26.

Daniel, Ines, Baier, Daniel (2013): Lifestyle Segmentation Based on Contents of Preferred Images versus Ratings of Items. Studies in Classification, Data Analysis, and Knowledge Organization, (45), 439-448.

Garrod, B. (2009): Understanding the Relationship Between Tourism Destination Imagery and Tourist Photography. Journal of Travel Research, 47(3), 346-358.

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., Di Minin, E. (2018): Social Media Data can be used to Understand Tourists' Preferences for Nature-based Experiences in Protected Areas. Conservation Letters, 11(1).

Sanchez-Franco, M. J., Rondan-Cataluña, F. J. (2010): Virtual Travel Communities and Customer Loyalty: Customer Purchase Involvement and Web Site Design. Electronic Commerce Research and Applications, 9(2), 171-182.

Statista (2018): Smartphones Cause Photography Boom. URL: https://www.statista.com/chart/10913/number-of-photos-taken-worldwide (23/11/2018).

Beer Brand Image Classification Using Deep Learning

Atsuho Nakayama[1], Daniel Baier[2] ([1]: Tokyo Metropolitan University; [2]: University of Bayreuth)

We want to use consumers' uploading habits on internet for marketing purposes. The distribution of posts on the internet has been increasing. Recently uploading habits have become part of our lives such as sharing photos on social networking sites or instant sharing with smartphones. People use images and text on the internet to represent their activities, interests and opinions. Despite different demographics, posts of different users contain similar interests. We would like to use these free photos or uploading habits for marketing purposes.

We collect images of brands in some category and classify images using deep learning. Based on the results, we will clarify what kind of image consumers have for each brand. Image recognition technologies using artificial intelligence such as deep learning is rapidly evolving in recent years. These technologies have great influence on marketing operations. Artificial intelligence is essentially concerned with automating classic human-engineered intellectual tasks. The approach to achieving this goal is hard-coded, comprehensive collection of explicit rules for processing information. The recent artificial intelligence trend relies mainly on great successes of deep learning algorithms, which in turn are a class of machine learning algorithms. Deep learning are neural networks with many layers and feature selection automated. Deep convolutional neural networks have replaced decision trees as the best method especially with little pre-structural problems such as image analysis. Deep learning feeds in large number of data and answers and independently determine a set of rules from this information. This process is referred to as training. The data used for training has features that will be used to determine the output. All training data has a label that includes the desired output. Learning the rules is done using a loss function that indicates the degree of difference in the current output of the algorithm from the desired output. The parameters are chosen so that this loss function minimizes, i.e. the difference is as low as possible or below a selected threshold. The result is expected to correctly map new data that does not contain a label. The process of multi-level filtering of the data provides a very useful representation of the information.

We searched German beer brands' TV-spots that are uploaded on the official channel of the beer brand on YouTube. We collected TV-spot videos of 13 German beer brands. We used OpenCV to

handle videos in Python. OpenCV is the most popular library for image processing and video processing. We divided the TV-spot movies to the image by the frame. We loaded the movie, confirmed the existence for each frame, and saved them as an image into the specified directory. The images of 13 German beer brands are analyzed by a deep convolutional neural network model using Keras with Python. Images are divided into train, and test data set. 80% of all data was used as training data and 20% was used as test data. Based on the results of this analysis, we aim to clarify the relationship among images of brands and to reveal the image of consumers behind them.

The Effect of Preprocessing on Short Document Clustering

Cynthia Koopman[1], Adalbert Wilhelm[1] ([1]: Jacobs University Bremen)

Document clustering has gained popularity due to social media and its large volume. Natural Language Processing is able to extract information from unstructured data which can be powerful for businesses. Social media, customer reviews and even military messages are all very short and therefore harder to handle than longer texts. Cluster analysis is essential in gaining insight from these unlabeled texts. Pre-processing often removes words, which can become risky in short texts, where the main message is made of only a few words. The effect of pre-processing and feature extraction on these short documents is therefore analysed in this paper. Six different levels of text normalization are combined with four different feature extraction methods. These setting are all applied on K-means clustering and tested on three different datasets. Anticipated results cannot be concluded, however other findings are insightful in terms of the connection between text cleaning and feature extraction.

2016 US Presidential Election Sentiment Analysis

Bryant Hwang (Korea International School)

Despite high rates of past success, the majority of polling agencies were unable to predict the outcome of the 2016 United States presidential election between Hillary Clinton and Donald Trump. Further research has hypothesized that these agencies underestimated the amount of support for Trump and overestimated the representation of recent college graduates. Sentiment analysis of posts on popular social media sites presents promising possibilities for better gauging public opinion. In our research, we analyze positive and negative sentiment in Reddit posts from key dates during the 2016 election that mention either candidate. The results of our study match the results of a poll conducted by the New York Times, which showed that public opinion for Trump tended to stay consistent while Clintons positive sentiment decreased over time. Our study also shows that public opinion expressed on Reddit changed drastically in accordance to major events, like debates and email controversies. Alone, our work in sentiment analysis does not produce statistically significant enough results to predict the final outcome of an election but having the ability to gauge public opinion through social networks is a powerful way that polling agencies and researchers can strengthen their predictions in real time.

Innovation

Forecasting Sales of Durable Goods – Does Search Data Help?

Carsten D. Schultz (University of Hagen)

Search data comprises information of users' search queries entered in search engines during the search process. Search data has previously been used to forecast automobile sales (Fantazzini & Toktamysova 2015), cinema admissions (Hand & Judge 2012), economic indicators (Choi & Varian 2012; Vosen & Schmidt 2011), housing prices (Dietzel 2016; Oestmann & Bennöhr 2015), influenza epidemics (Ginsberg et al. 2009), tourism demand (Önder & Gunter 2016; Park et al. 2017), and unemployment rates (D'Amuri & Marucci 2009; Ettredge et al. 2005). The present study explores the use of such search data to forecast sales of durable goods. Specifically, this study draws on panel

data over a two-year time period for seven product groups: audio books, Blu-ray-players, calendars, child books, Hi-Fi-systems, smartphones, and televisions. These seven groups are well represented by a single search term. The corresponding search data is retrieved from Google Trends. Google Trends reports a weekly index of the search volume for the corresponding search terms.

Explorative findings indicate that search data provides some value to forecast product group sales of durable goods. First, linear regression between search data and durable goods sales is tested with time lag from zero to five weeks. Across the seven groups, different time lags are suited to predict weekly sales (audio books t = -5, Blu-ray-players t = -1, calendars t = -5, child books t = -1, Hi-Fi-systems t = 0, smartphones t = 0, and televisions t = -3). R^2 is .188 for audio books, .371 for Blu-ray-players, .809 calendars, .460 for child books, .479 for Hi-Fi-systems, .263 for smartphones, and .258 for televisions. The Durbin-Watson tests show signs of autoregressive residuals but the coefficients in the first order autoregressive regressions are not significant in most cases and do not provide a benchmark forecast. The average product group price is found to be not a good predictor in a linear regression for group sales.

A New Algorithm to Optimize the Development of Innovative Products and Services

Gloria Gheno[1], Massimo Garbuio[2] ([1]: Innovative data analysis; [2]: University of Sydney)

To remain competitive, it is essential to know how to interpret, anticipate and satisfy the demands of a competitive and global market, such as the current one. An efficient connection and collaboration between innovation and marketing can be obtained using the data available on goods already developed and placed on the market, improving the process of development of new products or services. Until now the main algorithms proposed in literature, using cluster techniques, grouped the customers on the basis of their requests or gathered the particular characteristics of the products on the basis of the customers. Only few pioneering works, using bicluster techniques so as to group simultaneously customers and characteristics of the products, have managed to bring together the requests on the basis of the target customers. Using these works we develop an algorithm which determines target customers with the characteristics necessary for them and with those negligible for them. The innovative proposal of our algorithm is the determination of superfluous characteristics, which effectively leads to saving time and money in the development of new products or services. To interpret the trend of the market and its evolution in time, we also present a modified version of the tricluster, an algorithm never applied in economics, which allows to monitor, in different moments, for each group of customers, the specific required characteristics and the useless ones. An integral approach between marketing and innovation, allowing to understand and anticipate market demands, creates decisive factors to remain competitive. The goodness of the algorithm is demonstrated applicating it to a real case and comparing it to methods already present in literature with statistical tests. Our method groups the target customers in a more profitable way with the characteristics, which they require, and it gets much more information than other methods.

Constructing an Ideal Workstream Collaboration Tool for Coworking Spaces Using Single-product Choice-based Conjoint Measurement

Cristopher Siegfried Kopplin[1], Daniel Baier[1] ([1]: University of Bayreuth)

Coworking spaces require and foster communication and collaboration among members, providers' staff, and between members and providers. A variety of tools seeks to fulfill the needs of both coworking space providers and members in this regard. These tools mainly emerge from the field of unified communications and amplify the field's scope and practice. Consequently, various terms and denotations have appeared that leave both researchers and practitioners disoriented. The first goal of the study at hand thus is to clarify the relevant software terms. Recent developments in collaboration tools have created a new market where both incumbents and new companies are present. Applications share a large corpus of features and functionality, impeding the composition of a unique selling proposition. Thus second, a single-product choice-based conjoint (CBC) approach among coworking

space members is used to develop a concept of an ideal collaboration tool. Single-product CBC allows for a choice between a stimulus and the none-option, simulating a binary buying decision for a concrete product at a time. Seven attributes with three levels each are employed to characterize the dominant features. As coworking spaces are a global phenomenon, surveys are undertaken in Germany and the US.

Interpretable Machine Learning 1

Interpretable Decision Sets - An Analysis

Jiri Filip[1], Tomas Kliegr[1] ([1]: University of Economics)

The objective of our work is the analysis of the IDS (Interpretable Decision Sets) algorithm (Lakkaraju et al, 2016). From pre-mined set of association rules, IDS selects a subset of rules so that an objective function reflecting classifier comprehensibility as well as accuracy is optimized. Due to the nature of the optimization task, as we are selecting a subset of the pre-mined rules, the objective function is submodular and non-monotone. Maximizing such function can be achieved with smooth local search (SLS), which provides ⅖ approximation guarantee.

In our work, we investigate whether IDS really is efficient and usable on a large-scale basis. Our experiments show that it might not be so. We evaluate IDS empirically, using our Python implementation on several datasets, and analytically, investigating its performance as relating to time complexity. We also analyze the multi objective function used by DS and each of its comprising objective, such as number of rules or accuracy of the rule list. We propose an alternative way to define the objective that might lead to the optimization problem being easier to solve. We also demonstrate why interpretability of certain criteria used in the original objective function might be biased and discuss if rules on the output of IDS are indeed interpretable.

Explaining Interpretable Models: A General Process Model for Explanation Generation in Inductive Logic Programming

Mark Gromowski[1], Michael Siebers[1], Ute Schmid[1] ([1]: University of Bamberg)

With the ongoing integration of Artificial Intelligence into everyday life, it is becoming increasingly important to make users understand what a system is doing and, in particular, why it is doing certain things. Thus, a high classification accuracy alone is not sufficient anymore in the area of Machine Learning. Consequently, there is a growing focus on explaining a system's classification choices. The usage of symbolic Machine Learning approaches, as for example Inductive Logic Programming (ILP), allows a system to generate classification rules that can be interpreted by human users familiar with the given approach – but interpretability is only the first step. In order to be understood by a broad variety of users independent of their AI background knowledge, these interpretable rules have to be further refined. For this purpose, we developed a general process model for ILP-based applications providing users with comprehensive explanations of classification choices made by the system. These explanations can take different forms and empower users to evaluate and correct the system's decisions. Applications implementing the model can either learn their classifier from a given set of prelabeled training examples or start with a handcrafted classifier only derived from background knowledge and well-considered assumptions. In both cases, the overall performance of the model is constantly improved through user feedback in the form of corrections, constraints and the integration of valuable domain knowledge – this way, the classifier within the model advances over time during repeated learning procedures. By implementing the model's structure, a constant mutual exchange of information between system and user is ensured instead of having a one-way information stream. The model is applicable to a broad variety of different scenarios, reaching from automated pain detection from a person‘s facial expression (example: PainFaceReader) to the analysis of a user‘s file system in order to recognize redundant files (example: Dare2Del). Explanations are generated in verbal form on the basis of the ILP Rules that are applied by the classifier. Additional explanations

are generated individually depending on the given application scenario and the type of input data and can take different forms – for instance, in the case of a classification task based on images or videos, an explanation may be given in the visual form of annotated images highlighting features that are relevant for the classification.

Methods for Understanding the Influence of Input Variables on the Decision of a Deep Artificial Neural Network

Torsten Dietl (Technische Universität Darmstadt)

Machine learning (ML) and artificial intelligence (AI) are topics with high industrial impact, due to the enormous success in applying deep artificial neural networks (DANNs) to various pattern recognition tasks.

The results of DANNs are so convincing that neural nets are already getting tested in heavily regulated fields like medicine or finance. However, these autonomous systems are deployed without evaluating the reasoning behind their decisions. Despite all the accomplishments DANNs achieved, their decisions are still mostly a black box. Thus, the aim of current research is to explain the influence of input variables on the decision of a DANN.

A promising approach is the "Linear Weighting Scheme for the Contribution of Input Variables (LICON)" by Kasneci and Gottron (2016). The LICON method is able to calculate reasonable input contributions for a given input sample. However, it has a locality issue, because the calculated contributions are only applicable to the input they were calculated on.

The approach adopted in the presented work tackles this locality issue by combining the LICON method with the "Global Sensitivity Analysis (GSA)" (Cortez and Embrechts, 2011). The GSA produces input contributions by calculating and aggregating contributions for sampled inputs.

Due to the sampling, it is possible to examine multiple values of an input. This means, the contribution calculated locally by LICON can be assessed in relation to contributions calculated for sampled input values.

The effectiveness of the proposed approach was assessed through a comparison study of the involved explanation methods (LICON, GSA and the new proposed approach). The three methods were used to generate input contributions for neural networks that were trained on an artificially generated dataset (laboratory conditions), the German credit dataset ("real-world" example) and the MNIST database of handwritten digits (easily interpretable benchmark).

Despite the computational complexity, which has to be dealt with in the future, it was shown that the proposed approach generates reasonable input contributions on the tested datasets. Furthermore, it produces additional information (mean and variance values) for contribution of the examined input sample, when compared to the LICON method.

References

Kasneci, G. and Gottron, T. (2016): ‘LICON: A Linear Weighting Scheme for the Contribution of Input Variables in Deep Artificial Neural Networks’. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM 2016 10. Indianapolis, Indiana, USA: ACM, pp. 45–54.

Cortez, P. and Embrechts, M. J. (2011): ‘Opening Black Box Data Mining Models Using Sensitivity Analysis’. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 341–348

Comparing Regularization Methods for Activation Maximization

Antonia Hain[1], Ulf Krumnack[1] ([1]: University of Osnabrück)

Activation maximization is a popular method when it comes to interpreting features learned by deep neural networks. By creating an artificial input stimulus that maximally activates a selected unit, direction or subspace in such a network, one can get an impression of what semantics are coded by this entity. However, in most cases simple gradient ascent leads to unsatisfactory results, suffering from noise and other artifacts. For this reason, in recent years a multitude of approaches has been

proposed to achieve more natural (i.e. interpretable) input stimuli. However, the utility of different techniques, as well as their potential for combination has not been thoroughly investigated yet.

In our work we compare different combinations of regularization techniques that have been proposed in literature for networks operating on visual inputs. We implement a tool that allows to directly contrast individual techniques and analyze the effect of these regularizers on the interpretability of the generated images as well as on the optimization process with respect to contrast, stability, and computation time. The direct comparison shows that regularization helps to accelerate the optimization process.

In a pilot study we investigated the interpretability of generated images to untrained subjects. We applied activation maximization to visualize output neurons of the well-known AlexNet. Subjects were prompted to provide their interpretation of the image contents and a score was computed to assess the overall recognizability of an image.

While unregularized activation maximization leads to hardly recognizable stimuli, we collected evidence that image interpretations are much closer to the target class after the application of different regularization techniques. Especially regularizers penalizing high-frequency structures are shown to be effective and increase interpretability significantly, and best results are achieved when multiple regularizers are combined. Furthermore, we show that meaningful features can be extracted from hidden layers with the use of different regularization techniques. The investigated method could be adapted to be used on other networks as well. We expect that results achieved with several regularization techniques will translate to the stimuli produced using activation maximization on these networks.

Interpretable Machine Learning 2

Comprehensibility in Symbolic Decision Systems

Florian Lerch[1], Alfred Ultsch[1] ([1]: Philipps University of Marburg)

In recent years there was a great increase in interest for more human understandable interpretations of machine learning models. Especially with LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al. 2016) introduced in 2016, defining a popular framework for introducing explanatory elements in a model agnostic way, the search for comprehensibility became a more important issue in many fields. Substantial advances in the comprehension of neural networks and other "black box" algorithms have been made in recent years (Zhang and Zhu 2018). However, there is only slow progress in traditionally human interpretable decision systems (Fürnkranz et al. 2018) with symbolic representation.

The terms "comprehensibility" and "interpretability" are prominently used in titles and abstracts of recent papers concerning symbolic decision systems, like decision trees, lists, rules etc. However, in these papers only measures for some kind of model size are presented. These are not necessarily more comprehensible (Stecher et al. 2016, Murphy and Pazzani 1994, Huysmans et al. 2011) and miss a lot of other important factors (Vellido et al. 2012, Lipton 2016).

The strong increase in the interest in human comprehension of classifier systems and even legal demand for human interpretation of results calls for a new benchmark for the comprehension of classifier systems (Lipton 2016).

An overview and critical review of the approaches to comprehensibility is presented, focusing on symbolic classification systems. The approaches are summarized into a catalog of criteria which reflect the qualities a comprehensible rule system should offer. Existing systems are discussed within this framework.

References

Ribeiro, M. T., Singh, S., Guestrin, C. (2016): 'Why Should I Trust You? Explaining the Predictions of Any Classifier', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Zhang, Q., Zhu, S.-C. (2018): 'Visual Interpretability for Deep Learning: A Survey', Frontiers of Information Technology & Electronic Engineering, 19(1), pp. 27–39.
Fürnkranz, J., Kliegr, T., Paulheim, H. (2018): 'On Cognitive Preferences and the Interpretability of Rule-based Models', arXiv preprint arXiv:1803.01316v2.
Stecher, J., Janssen, F., Fürnkranz, J. (2016): 'Shorter Rules Are Better, Aren't They?', Proc. 19th International Conference on Discovery Science, 9956, pp. 279–294.
Murphy, P. M., Pazzani, M. J. (1994): 'Exploring the Decision Forest: An Empirical Investigation of Occam's Razor in Decision Tree Induction', Journal of AI Research, 1, pp. 257–275.
Huysmans, J. et al. (2011): 'An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models', Decision Support Systems. Elsevier B.V., 51(1), pp. 141–154.
Vellido, A., Martin-Guerrero, J. D. Lisboa, P. J. G. (2012): 'Making Machine Learning Models Interpretable', ESANN 2012 Proc., pp. 163–172.
Lipton, Z. C. (2016): 'The Mythos of Model Interpretability', ICML Workshop on Human Interpretability in Machine Learning, pp. 96–100.

An Analysis of Design Choices in the LIME Framework

Amir Hossein Akhavan Rahnama[1], Henrik Boström[1] ([1]: KTH - Royal Institute of Technology)

Several different approaches to finding explanations of machine learning models have been proposed. Some of them try to explain the models from a global (holistic) perspective, whereas others aim for explaining individual predictions. In addition, some approaches rely on specific properties of the underlying (black-box) models, while others are model-agnostic.

The recently proposed LIME framework for explaining predictions has attracted a lot of interest in the research community. The framework, which is model-agnostic, first projects a prediction from the original feature space into a new space, called an interpretable representation, and then learns an interpretable model, such as a logistic regression model, in the local neighborhood of the prediction. The original paper introducing the LIME framework presented rather impressive results on high dimensional data, such as text and image data sets, using random forests and deep neural networks as the underlying models, and the framework has also been demonstrated to work well for standard tabular datasets.

In this study, we highlight some of the design choices that (implicitly) underlies the LIME framework and discuss and analyze some of their corresponding effects on the produced explanations. These choices include: i) using uniform sampling from the new (interpretable) feature space, rather than sampling from the underlying distribution, e.g., by selecting a subset of the training instances, ii) when constructing the interpretable model, the instances are weighted based on their distance to the predicted instance within the original feature space, rather than within the interpretable feature space, iii) the employed concept of locality assumes that a similarity kernel is provided, however with no explicit connection to the underlying black-box model, e.g., it may not effectively capture the decision boundary of the underlying model, iv) the framework involves searching for an explanation that minimizes a specific loss function, which balances interpretability (model complexity) and local faithfulness, however, without using any criterion for deciding whether an explanation is of sufficient quality or not, and v) the framework exploits only the predicted labels of the black-box model, and ignores any measures of (un)certainty that may be provided by the underlying models, such as class probabilities.

Some of the above design choices are analyzed in more detail and illustrated using results from multiple datasets that are similar to the ones in the original study. Finally, we will discuss possible future research directions.

Explaining Relational Concepts: When Visualization and Visual Interpretation of a Deep Neural Network's Decision are not Enough

Bettina Finzel[1], Johannes Rabold[1], Ute Schmid[1] ([1]: University of Bamberg)

One of the challenges in Deep Neural Network based image classification is the explanation of a neural network's decision. Recent privacy policies and the use of autonomously deciding systems raise the demand for comprehensive, transparent and trust-worthy approaches. While Deep Learning

produces good results in image classification, decisions are not inherently transparent and comprehensive to humans. Factors that influence the comprehensibility of classification results are hidden in complex network architectures, calculated weights and parameter settings. State-of-the-art techniques for Explainable AI (XAI), like Local Interpretable Model-agnostic Explanation (LIME) or Layer-wise Relevance Propagation (LRP) are promising approaches to shed light into the black box. They make use of visual highlighting of pixels and pixel regions to denote which have been considered as relevant to a black box classifier during learning. However, visual highlighting is limited to being interpreted by a human. That means, complex features represented in images, such as relationships between objects, remain unexplained. Hence the human needs knowledge about the domain of discourse in order to understand the underlying meaning of the visualization. We want to make the inside of a neural network not only accessible to experts in the domain but also to laymen. We present a companion system that makes use of Inductive Logic Programming to learn relations from visual features generated by different XAI techniques. This companion may serve as an Intelligent Tutor or Expert Decision Support system. We use data sets with relational concepts as training input to a deep neural network. Then we apply state-of-the-art XAI methods. With the help of Qualitative Spatial Reasoning we derive spatial relations between objects that have been detected to be relevant. We then apply Inductive Logic Programming in order to find logical rules that explain classification based on spatial relationships. In a later step we transform these logical expressions into verbal explanations. We belief that our approach reduces the semantic gap by explaining relations and thus complements visualization.

Interpretable Machine Learning 3

An Interactive Visual Tool to Enhance Understanding of Random Forest Predictions

Ram Bahadur Gurung[1], Tony Lindgren[1], Henrik Boström[2] ([1]: Stockholm University; [2]: KTH - Royal Institute of Technology)

Random forests have been shown to often deliver good predictive performance. However, the predictions are not easy to understand as they are the result of averaging individual predictions of a large number of diverse decision trees. In order to provide support for the understanding of the prediction for a specific test instance, an interactive visual tool has been developed where a user can manipulate selected features to evaluate “what-if” scenarios and see how the results would change on the fly. The tool provides aggregate information of the paths the instance follows in each decision tree, including the number of times each feature is a part of condition along one of the paths, the distribution of threshold values of the features in these conditions and at what depth in the trees the conditions were evaluated. Furthermore, the tool allows the user to visualize not only the rank of the features according to their importance, but also shows how different feature values would affect the importance, through density and contour plots. The tool also provides the user with a simple decision rule as an explanation for the prediction. The explanation is generated by first building a decision tree using the training instances that fall into same leafs as the test instance, together with the labels as predicted by the random forest and extracting a rule from the path the test instance follows in the generated decision tree. In addition, the tool provides support for automatically adjusting feature values of the test instance with minimal cost to change the prediction into a preferred class label. In order to evaluate the usability of the tool, a case study was undertaken at a large truck manufacturing company in Sweden, targeting the prediction of failure of truck components. A set of domain experts were invited to use the tool and provide feedback. Findings from this evaluation will be presented, and, in the light of these, requirements on tools to support the understanding and interaction with black-box predictive models will be discussed.

Representational Capabilities of Distilled Soft Decision Trees

Benedikt Wagner[1], Tarek R. Besold[2] ([1]: City, University of London; [2]: Telefonica Innovation Alpha)

In recent years, Machine Learning-based systems, including Neural Networks, are experiencing greater popularity. A weakness of these types of models, which rely on complex representations, is that they are considered black boxes with respect to explanatory power. It has been suggested that there are unavoidable trade-offs between accuracy and interpretability. In an attempt at overcoming these limitations, Frosst and Hinton (2017) introduced Soft Decision Trees that incorporate distilled knowledge from a Neural Network in order to illustrate how a decision was made whilst still achieving reasonable predictive performance. This method aims to represent the implicit information contained in a Neural Network inside a sparser model, allowing to extract comprehensible information regarding the decision. Soft Decision Trees enable hierarchically distributed explanations by using visual masks, which provide insight into discriminatory features that are dependent on the local context.

Our findings suggest that the accuracy-intelligibility trade-off is not solved by Soft Decision Trees, but that these methods can rather be considered as a means to work in between the extremes.

When examining different tree depths, we discovered gradually decreasing interpretability due to the excessive number of nodes on the one hand, and the difficulty to discriminate between them using weaker regularities in the classified image on the other. Weak regularities are difficult to understand and do not allow for alignment with prior expectations formed based on background knowledge or intuition. Smaller trees, however, provided interesting insight into the model, making it comprehensible relative to the task and its related data.

We emphasise new applications for Soft Decision Trees, point out limitations, and show how Soft Decision Trees generally align with logic-based methods such as TREPAN.

In summary, we demonstrate that the Soft Decision Tree method can be useful to validate the findings of a trained Machine Learning system. We show that Soft Decision Trees are better suited to give necessary insights into the decision-making in particular cases and can, for instance, contribute to preventing faulty behaviour of the system. Nevertheless, the dataset type and image attributes were found to be essential concerning prediction and interpretability aspects. As such, Soft Decision Trees do not generally solve the black box challenge but can be a useful tool for selected types of applications.

Instance-based Explanations: Motivation, Overview, and the Evidence Counterfactual Approach

Yanou Ramon[1], David Martens[1] ([1]: University of Antwerp)

Predictive modeling applications using high-dimensional, sparse data are ample, and range from document classification to mining behavioral data. Examples include predicting product interest based on online browsing data or Facebook likes, predicting spam emails, and detecting objectionable web content.

The high-dimensional and sparse nature of the data brings serious transparency issues for predictive models that are built on such data: even linear models require investigating thousands of coefficients, while non-linear models amplify the problem even further. However, explainability of the predictions made is crucial for trust, to accept predicted outcomes and to gain relevant insights.

In this paper, we provide an overview of existing explanation methods, with a focus on instance-based explanations, that are particularly suitable for this data type. We compare four techniques (EDC, SHAP, LIME, anchors) using both quantitative and qualitative experiments. The evaluation criteria that are considered follow a framework that takes into account who the explanations are for (manager, customer or data science team), and include algorithmic efficiency, explanation fidelity and subjective preference.

Oracle Coaching for Informative Decision Trees

Cecilia Sönströd[1], Ulf Johansson[2], Henrik Boström[3] ([1]: University of Borås; [2]: Jönköping University; [3]: KTH - Royal Institute of Technology)

Generally, the best predictive performance is obtained using techniques that produce opaque models, like support vector machines, neural networks or ensembles, making it impossible for a human to inspect the relationships found. Transparent models, where relationships between input and output are made explicit, on the other hand, enable insights into the underlying domain to be gained. A transparent and relatively small model, in some easily interpretable representation, allows analysis of the entire model, whereas a larger transparent model at least makes it possible to follow the reasons for each prediction. Most often, interpretability, in the form of using a transparent model, comes at the price of reduced predictive performance, compared to an opaque model. The situation where predictive performance is sacrificed in order to obtain an interpretable model is known as the accuracy vs. interpretability trade-off and has been investigated extensively in data mining research.

Oracle coaching is a method for improving predictive performance of transparent models in the very common situation where the instances to be predicted, i.e., the production data, are known and available at the time of model building. The overall purpose is to reduce the accuracy vs. interpretability trade-off by producing interpretable models optimized for the specific production set at hand. The approach employs a strong opaque model, called the oracle, to guide the generation of a transparent model, specialized on the production set. This is accomplished by using the oracle to predict labels for the production data, and then learning the transparent model on this data, possibly in conjunction with the original training set. In previous studies, oracle coaching has been shown to significantly improve predictive performance for transparent models, both for classification and regression. This increase in predictive performance, however, often came at the expense of larger models, since more training data usually results in larger trees. In this paper, using random forests as oracles, we address the question of whether decision trees built using oracle coaching retain their superior predictive performance when tree size is restricted and, more importantly, if oracle coaching inherently leads to better quality top splits. This is achieved by comparing trees of the same depth, which are obtained by successively pruning full trees, down to a specific depth.

The experiments, using a large number of benchmark classification data sets, show that the addition of oracle data improves the quality of the transparent models, in the sense that splits high up in the tree are more discriminatory. Consequently, oracle data can be used to significantly improve the predictive performance of decision trees, without increasing tree size. Furthermore, the study provides a detailed investigation into the interpretability vs. accuracy trade-off, by step-wise shrinking tree models to smallest size and studying the loss of predictive performance for each step. Overall, the results show that even heavily pruned tree models learned using oracle coaching retain relatively high accuracy, thus making them very attractive for inspection and analysis.

Machine Learning 1

Multivariate Extrapolation – A Tensor-based Approach

Josef Schosser[1], Angelika Schmid[1] ([1]: University of Passau)

Relational data is common in different scientific domains and real-world applications. Unlike traditional data collected from individual objects, relational data violates assumptions of independence or exchangeability. The modeling of these interactions is crucial for a proper understanding of the studied phenomena. As relational data is usually evolving over time, the problem of temporal link prediction has received particular attention in recent years. Temporal link prediction is characterized as the task of predicting future edges between nodes in a network. Many application problems can be described in terms of temporal link prediction.

Literature on the topic provides simple approaches that combine tensor decompositions and time series extrapolation. As generalizations of matrix decompositions, tensor decompositions are able to extract meaningful, latent structure in multiway data. The periodic patterns identified serve as input

for time series methods. In the paper at hand, we contribute to existing literature in the following way. First, we connect state-of-the-art tensor decompositions with a general class of state-space models underlying exponential smoothing. In doing so, we offer a useful framework to summarize existing literature and provide various extensions to it. Second, through several numerical experiments, we demonstrate the effectiveness of the proposed method. We synthesize data that exhibits different periodic patterns and show that our approach is able to reveal these patterns in time. Moreover, we use real-world data to demonstrate the superiority of our method over traditional extrapolation approaches. The latter cannot capture dynamic inter-relationships between variables of interest. In contrast, the proposed method identifies these associations and significantly improves forecast accuracy. Finally, we offer guidance on model selection and coding for practical applications.

Time Series Forecasting Using Variants of Recurrent Neural Networks

Chettan Kumar[1], Kurt Safak[1], Stefan Schönig[1], Stefan Jablonski[1] ([1]: University of Bayreuth)

Time series forecasting has a fundamental importance in business, finance and economics. There are several traditional models for time series forecasting such as Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), Exponential Smoothing (ES) and Autoregressive Integrated Moving Average (ARIMA).Deep learning algorithms such as Recurrent Neural Network (RNN), its variants algorithms and models like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) were developed for times series analysis and other complex applications of data analysis. The goal of our study is to evaluate and compare RNN models for time series forecasting according to models like ARIMA, ARMA, and ES. The expected output of this study is a comprehensive analysis of both forecast accuracy and run time efficiency used to optimize results in weather forecasting and many different areas that need predictions.

Machine Learning 2

Amharic Handwritten Character Recognition Using Convolutional Neural Network

Mesay Samuel Gondere[1], Lars Schmidt-Thieme[2], Abiot Sinamo Boltena[3], Hadi Samer Jomaa[2] ([1]: Arba Minch University; [2]: University of Hildesheim; [3]: Mekelle University)

Amharic is the official language of the Federal Democratic Republic of Ethiopia. There are lots of historic Amharic and Ethiopic handwritten documents addressing various relevant issues including governance, science, religious, social rules, cultures and art works which are very rich indigenous knowledge. The Amharic language has its own alphabet derived from Ge'ez which is currently the liturgical language in Ethiopia. Handwritten character recognition for non-Latin scripts like Amharic are not addressed especially using the advantages of the state-of-the-art techniques. This research work makes a first attempt to model Amharic handwritten character recognition using convolutional neural networks. The model was further enhanced using multi task learning from the relationships of the characters. Promising results are observed from the latter model which can further be applied to word prediction. The dataset was organized from collected sample handwritten documents and data augmentation was applied for machine learning.

A Research Survey on Applications of Recurrent Neural Networks and its Variants

Chettan Kumar[1], Muhammad Hassaan[1], Stefan Jablonski[1] ([1]: University of Bayreuth)

This survey paper elaborates the performance of Recurrent Neural Networks (RNN) and its variations. Especially, we examine Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). The objective of this study is to show how RNN, LSTM, and GRU perform for different applications. In an

initial phase, we examine the performance of RNN and its variants. Further, we illustrate how LSTM solves the RNN's gradient vanishing problem and how GRU trains faster with less parameters as compared to LSTM. It also becomes clear that these algorithms are suitable for any application that deals with the sequential or time series data. As application fields we introduce time series forecasting, autonomous driving, human activity recognition, and handwriting recognition.

Machine Learning 3

Towards Automated Machine Learning for Multi-Label Classification

Marcel Wever[1], Felix Mohr[1], Eyke Hüllermeier[1], Alexander Hetzer[1] ([1]: Paderborn University)

Due to the ever-increasing demand for machine learning applications and the simultaneous lack of experts to meet this demand, automated machine learning (AutoML) has established itself as an important research topic in machine learning and artificial intelligence. For standard (single-label) classification and regression tasks, a wide range of AutoML tools have already been proposed and proved beneficial compared to handcrafted solutions (Bergstra et al. 2014). However, there is only very few work on other machine learning tasks. In particular, there is hardly any work on AutoML tools for multi-label classification (MLC), a generalization of multi-class classification that allows an instance to belong to several classes simultaneously (Zhang and Zhou 2014). In this work, we discuss the challenges of developing AutoML approaches for the task of MLC, in which the number of classes is often very large. Moreover, we present a first approach based on ML-Plan, a state-of-the-art AutoML tool for standard classification (Mohr et al. 2018). In an experimental study, we compare our approach with the only existing AutoML approach to MLC, which is a grammar-based genetic algorithm (Guimarães Cardoso de Sá et al. 2018). As additional baselines, we consider a simple grid search and a random search technique. Compared to these baselines, our approach is highly competitive and yields superior results.

References

James Bergstra, Rémi Bardenet, Yoshua Bengio, Balázs Kégl: Algorithms for Hyper-Parameter Optimization. NIPS 2011: 2546-2554

Min-Ling Zhang, Zhi-Hua Zhou: A Review on Multi-Label Learning Algorithms. IEEE Trans. Knowl. Data Eng. 26(8): 1819-1837 (2014)

Felix Mohr, Marcel Wever, Eyke Hüllermeier: ML-Plan: Automated Machine Learning via Hierarchical Planning. Machine Learning 107(8-10): 1495-1515 (2018)

Alex Guimarães Cardoso de Sá, Alex Alves Freitas, Gisele L. Pappa: Automated Selection and Configuration of Multi-Label Classification Algorithms with Grammar-Based Genetic Programming. PPSN (2) 2018: 308-320

Algorithm Selection as Recommendation: From Collaborative Filtering to Dyad Ranking

Alexander Hetzer[1], Marcel Wever[1], Felix Mohr[1], Eyke Hüllermeier[1] ([1]: Paderborn University)

Problem classes such as integer optimization, SAT, or classification can be tackled by a large variety of algorithms, the performance of which may differ depending on the concrete problem instance at hand. Research on algorithm selection and configuration (ASC) seeks to support and partly automate the selection of an algorithm that is most suitable for a given problem instance, and to set the parameters of that algorithm in the most appropriate manner (Bischl et al. 2016, Hutter et al. 2014). Based on the idea of treating ASC as a recommendation problem, some approaches on the basis of techniques such as collaborative filtering have recently been proposed (Cunha et al. 2018, Misir and Sebag 2017, Sun-Hosoya et al. 2018, Yang et al. 2018): Instead of recommending products to users, algorithms are recommended for problem instances. Going beyond standard collaborative filtering, we propose to tackle ASC as a problem of dyad ranking, for which methods have recently been

developed in the field of preference learning (Schäfer and Hüllermeier 2018). This approach is motivated by at least two potential advantages. First, treating problem/algorithm pairs as dyads allows the learner to exploit properties (features) of both the problem instances and the candidate algorithms. Second, providing recommendations in the form of rankings of a set of candidate algorithms is presumably easier and arguably more adequate than evaluating each of them in terms of an absolute score. These advantages are substantiated by first experimental studies in the field of automated machine learning, i.e., the recommendation of machine learning algorithms for given data sets.

References

Bischl, B. et al. (2016): Aslib: A benchmark library for algorithm selection. Artificial Intelligence 237: 41-58.

Hutter, F. et al. (2014): AClib: A Benchmark Library for Algorithm Configuration. International Conference on Learning and Intelligent Optimization. Springer, Cham.

Cunha, T. et al. (2018): CF4CF-META: Hybrid Collaborative Filtering Algorithm Selection Framework. International Conference on Discovery Science. Springer, Cham.

Mısır, M., Sebag, M. (2017): Alors: An Algorithm Recommender System. Artificial Intelligence 244: 291-314.

Sun-Hosoya, L. et al. (2018): ActivMetaL: Algorithm Recommendation with Active Meta Learning. IAL 2018 workshop, ECML PKDD.

Yang, C. et al. (2018): Oboe: Collaborative Filtering for autoML Initialization. arXiv preprint arXiv:1808.03233.

Schäfer, D., Hüllermeier, E. (2018): Dyad Ranking Using Plackett–Luce Models Based on Joint Feature Representations. Machine Learning 107.5: 903-941.

Well-Calibrated and Specific Probability Estimation Trees

Ulf Johansson[1], Henrik Boström[2], Tuwe Löfström[1], Cecilia Sönströd[3] ([1]: Jönköping University; [2]: KTH - Royal Institute of Technology; [3]: University of Borås)

When predictive modeling is used for decision support or automated decision making, it is vital that the models are trustworthy. One fundamental property, often associated with trust, is interpretability, i.e., it must be possible to inspect the model and understand the logic behind individual predictions. The FAT/ML Principles for Accountable Algorithms and a Social Impact Statement for Algorithms, however, list responsibility, explainability, accuracy and auditability as the components of an accountable algorithm, thus making it clear that interpretability is just one of several properties of trustworthy algorithms. Under accuracy, one guiding question in the social impact statement is “How confident are the decision outputs by your algorithmic system?" Proposed solutions include performing sensitivity analysis and, most importantly, determining how to communicate the uncertainty for each decision.

So, to enable user-confidence in the predictions from a model, the model must not only be comprehensible and accurate, but also capable of reasoning about its own competence, or at the very least distinguish between predictions where it is certain and not. One obvious way of communicating this concept of algorithmic confidence is to supplement every prediction with some measurement of belief in that prediction. Currently, most classifiers are able to output not only the predicted class label, but also a probability distribution over the possible classes. If these probabilistic predictions are well-calibrated, i.e., the predicted class probabilities reflect the true, underlying probabilities, they are, of course, the best possible measure of confidence. If, on the other hand, the probabilistic predictions are not well-calibrated, the model actually becomes misleading. At the same time, probabilistic predictions must not only be well-calibrated, but also specific. The recently developed framework for prediction with confidence equips any machine learning technique with the ability to produce perfectly calibrated probabilistic predictions, using so-called Venn predictors.

In this paper, we suggest a novel, but very natural, Venn taxonomy, producing Venn predictors specifically tailored for decision trees. While all Venn predictors are theoretically guaranteed to be well-calibrated, a standard Venn predictor, utilizing a label-based taxonomy, is inherently not very specific, returning the same confidence for every prediction of a certain class. When using the suggested taxonomy, however, the result of the calibration is a fixed probability estimation tree, generated using the training and calibration sets, where each leaf contains a specific prediction, consisting of a label and a valid probability interval. Clearly, this is an interpretable and very informative model. In the

empirical investigation, the proposed Venn predictor was compared to Laplace estimates, Platt scaling, isotonic regression and a standard Venn predictor. The results show that the two Venn predictors and Platt scaling were empirically well-calibrated, while in particular the Laplace estimates, but also isotonic regression, were intrinsically optimistic, i.e., poorly calibrated. The estimates from the suggested Venn predictor significantly outperformed the standard Venn predictor with regard to log and Brier losses, while performing well against most other metrics used for the evaluation.

Machine Learning 4

Locally Learned SVMs - The Regionalization Approach

Florian Dumpert (University of Bayreuth)

Support Vector Machines (SVMs) play a successful role in supervised learning, i.e. in classification and regression, in many areas of science. Unfortunately, SVMs usually need a lot of capacity in terms of computational power which is why they sometimes need too much time to be a realistic alternative to other methods. During the last years, we worked on the idea of regionalization to overcome this problem. Good solutions from the computational point of view is one aspect. On the other hand, we proved statistical properties, e.g. robustness. This talk will summarize these previous works and their assumptions and offer some simulation results to visualize the power of the regionalization approach.

Hyperplane Folding - A Method for Handling Non-linear Relations in Support Vector Machines

Lars Lundberg[1], *Håkan Lennestad*[1], *Veselka Boeva*[1], *Eva Garcia-Martin*[1] ([1]: Blekinge Institute of Technology)

Support Vector Machines (SVMs) use linear hyperplanes to separate different classes. In the basic case, there are two classes and one hyperplane that divides the n-dimensional space into two parts. However, when there are non-linear relations between the classes and the features, a linear hyperplane is obviously not the best way to separate the two classes.

We present a new method called hyperplane folding. Based on the location of the support vectors, the n-dimensional data points are projected onto a 2-dimensional surface using n-2 rotations. Based on the location of the support vectors, the method then splits the dataset into two parts; a new SVM is created for each part. Each of these two SVMs has a separating hyperplane. We calculate the intersecting point and angle between these two hyperplanes and rotate some of the data points so that all data points can be separated by one of the two hyperplanes (the fold operation). We finally expand all points back to the original n dimensions. The method may be repeated many times, and in each iteration, the margin in the SVM is increased as long the margin is smaller than half of the shortest distance between any pair of points from the two classes. Hyperplane folding is particularly useful when there are non-linear relations. Experiments with 3-dimensional data with non-linear relations (Blood Pressure as a function of Age, Height and Weight) show that the margin does indeed increase and that the prediction accuracy improves. Our method can use any standard SVM implementation plus some basic manipulation of the data points, i.e., splitting and rotation.

Previous ways to handle non-linear relations in SVMs include piecewise linear SVMs with hinging hyperplanes and kernel tricks. Compared to hinging hyperplane approaches, our method is easier to implement since it is based on existing SVM implementations and simple operations, such as rotation and splitting. Moreover, we use the location of the support vectors to determine where to fold the dataset, which in many ways corresponds to where to put the hinges. Existing hinging hyperplane approaches do not use this kind of information. Compared to kernel tricks, our method provides better interpretability of the results, since we do not extend the number of dimensions or introduce new composite features. In fact, our method increases the interpretability since it enables us to detect non-linear relations (the non-linear relations typically occur in the dimensions where we do the folds). It is also possible to combine our method with kernel tricks. For future work, we plan to pursue further

evaluations and validations of the proposed hyperplane folding method on different and larger datasets.

Predicting Wine Quality Using Machine Learning Techniques

Stavros Dimitri Ioannidis[1], Dimitrios Avraam Ioannides[2] ([1]: University of Patras; [2]: University of Macedonia)

Wine companies are using product quality certification for the promotion of their products. This is a time taking and expensive process because it requires wine testing by human experts. Machine Learning Techniques (MLT) may reduce the time and the cost of the process. We apply methods as Neural Networks (NN), Random Forest (RF) and Support Vector Machine (SVM) using libraries from the open source statistic language R. The dependence of the quality of the wine is determined by its physicochemical characteristics (features). In our paper we study a big dataset of wines from a big company in North Greece, and we predict the values of the quality variable by the above mentioned MLT. Finally, we conclude that the prediction is improved if selected features are being selected.

Machine Learning 5

Convolutional Neural Networks in Economics: Opportunities and Challenges

Marvin Schweizer[1], Abdolreza Nazemi[1], Andreas Geyer-Schulz[1] ([1]: Karlsruhe Institute of Technology)

In recent years, the volume, variety and velocity of available economic data have increased considerably. This is to a major extent due to the rise of the internet which has made different types of data available that allow researchers to use and merge them for different purposes. The availability of these datasets creates a host of new research opportunities in economics. At the same time, scientists face theoretical and computational challenges to integrate and operationalize these new kinds of datasets, such as image data, in their research studies.

A class of machine learning techniques called deep learning promises a solution for this computational challenge. More specifically, convolutional neural networks (CNNs), a specialized type of deep learning architecture tailored to the use with image data, have demonstrated superior performance in image and video classification, object detection and other computer vision tasks. On the theoretical side, researchers in economics face the challenge that potential applications for image data and CNNs are often subtle and research questions often need to be cast into a computer vision problem at first. Yet, a number of recent studies already lead the way to overcome these challenges, resulting in high-profile, innovative research.

In this talk, we first provide an introduction to deep learning and convolutional neural networks. This is followed by an in-depth discussion of applications using CNNs in economic subfields, such as finance, risk management and macroeconomics. We conclude with an overview of unexplored or underexplored lines of future research and a summary of the major challenges regarding the use of CNNs in the field of economics.

Deep Physiological Models for Pain Intensity Recognition

Patrick Thiam[1], Friedhelm Schwenker[1] ([1]: Ulm University)

Standard feature engineering requires expert knowledge in the corresponding domain of application in order to develop a set of relevant descriptors for the task at hand. This characteristic hinders the generalization capabilities of machine learning approaches built upon so called handcrafted features. Deep learning approaches are characterized by the integration of feature engineering, feature selection and classification into a single optimization process. Such techniques have proven to be very successful in the domain of image and video classification. Moreover, deep learning approaches have

been able in several cases to outperform traditional approaches based on handcrafted features and also depict better generalisation capabilities.

In the following work, we explore deep learning approaches for the analysis of physiological signals. More precisely, deep learning architectures are designed and assessed for the combination of several physiological channels in order to perform an accurate classification of different levels of artificially induced pain intensities. Most of the previous works related to pain intensity classification based on physiological signals rely on a carefully designed set of handcrafted features in order to achieve a relatively good classification performance. Therefore, the current work aims at building relatively good pain intensity classification models without the need of domain specific expert knowledge for the generation of relevant features for the task at hand.

Keywords: Convolutional Neural Networks, Information Fusion, Signal Processing.

Churn Analysis Using Deep Learning: Methods and Application

Tobias Albrecht[1], Daniel Baier[1] ([1]: University of Bayreuth)

Customer Relationship Management (CRM) is becoming increasingly important in times of global markets and intense competition. In particular, due to the higher profitability of long-term customers for businesses, accurately predicting customer churn as an integral part of proactive customer churn management is becoming a fundamental issue for academics and practitioners alike. The main focus in that matter is on the search for precise, innovative forecasting methods. Deep learning, with increasing popularity especially in the field of automated information processing and decision-making, is one such method. Its practical application in the context of churn prediction however, is so far often limited by its lower practicability and interpretability in comparison with other popular methods in this area.

This study applies deep learning for churn prediction to customer data of a telecommunications provider and examines the model performances of various feedforward networks with different numbers of layers compared to random forests as one of the most effective and popular benchmarking methodologies. In addition to the prediction quality, it is shown to be crucial to include the practical applicability of deep learning in a churn context in the form of the methodological implementation as well as the interpretability of results in the analysis. For this purpose, a framework for the application of deep learning in customer churn analysis is derived from the requirements of the general churn prediction process and the methodical procedure when using supervised machine learning classifiers. The outcomes of the present study show very good prediction results in method comparison, especially for deep learning networks with increased model depth. In addition to the performance-related findings, the increased practicability of such models in churn analysis through the use of innovative methods in the areas of pre-processing and interpretation of results is to be emphasized. Inhibiting factors of the practical use of deep learning in churn context can be reduced so that a high model practicality can be achieved.

Marketing Research

Influence of Error Factors in Marketing Research: A Monte Carlo Method Based Analysis in the Retail Context

Michael Brusch[1], Ines Brusch[2], Eva Stüber[3] ([1]: Anhalt University of Applied Sciences; [2]: Brandenburg University of Technology Cottbus-Senftenberg; [3]: IFH Köln GmbH)

Marketing decisions are mainly supported by methods of market research due to their far-reaching consequences. Such methods are often based on empirically collected data and are evaluated using complex analysis methods. The goodness of data of the derived market research results plays an important role here. This quality of the data can be affected by several types of errors, which are distinguished in particular into systematic and random errors and is also influenced by the sample size (e.g., Cochran 1968). Accordingly, market researchers have to consider the goodness of data

and should know which factors will have which kind of influence. The problem of a vague goodness of data is particularly relevant in the case of new or difficult to describe offers such as (stand-alone or partial) services with their immateriality. Especially the increasing importance of services in several markets makes it necessary to consider this issue in market research projects.

In our paper, the influence of different factors of the goodness of empirical data in the vibrant retail context (e.g., Kushwaha/Venkatesh 2013) will be investigated within a Monte Carlo experiment (e.g., Fishman 1996). Therefore, a real empirical data set (n=1,500) of a survey regarding buying behavior in stationary and online shopping is used as "true" data and will be compared with "generated" data. The "generated" data are randomly disturbed and systematically varied alternatives of the "true" data. The systematical variations are, e.g., regarding the usage of a simplified response scale (like the traffic light system with three response options (red vs. yellow vs. green) instead of the original 5-point scale) or regarding the sample size (one vs. two vs. three thirds of the original sample). The two data sets can be compared with respect to their correlation, both regarding selected overall mean conformity values and regarding segment specific values. In addition to the disturbance of the data (e.g., based on normal distribution) sufficient replications will be implemented and allow influence estimations.

References

Cochran, W. G. (1968): Errors of Measurement in Statistics, in: Technometrics, 10 (4), 637-666.
Fishman, G. (1996): Monte Carlo, Concepts, Algorithms, and Applications; Springer, New York.
Kushwaha, T., Venkatesh S. (2013): Are Multichannel Customers Really More Valuable? The Moderating Role of Product Category Characteristics, in: Journal of Marketing, 77 (July), 67-85.

Handling Mechanisms of Missing Values Within Marketing Research Journals - A Literature-Based Analysis of Widely Used Missing Value Treatments Within High Ranked Marketing Journals and Open-Access Marketing Journals

Redhwan Amer[1], Malek Simon Grimm[1], Ralf Wagner[1] ([1]: University Kassel)

This study investigates the occurrence and treatment of missing values within highly ranked marketing research journals and marketing-related open-access journals. The study investigates how often missing values are reported, how missing values are handled, and which handling mechanisms are applied.

The most recent 100 studies of 10 relevant and highly ranked marketing journals were investigated based on a keyword supported text search; resulting in a sample of 1,000 journal contributions. The treatment of missing values was noted and categorized. In order to provide a comparison base, an identical investigation was conducted for relevant open access journals within the area of marketing research.

Missing values are frequently reported within the selected journals. However, most researchers tend to simply exclude subjects or entire cases if missing values occur; even though there are sufficient and fairly simple to use handling mechanisms such as listwise and pairwise deletion, mean-replacement, and/or data imputation. Higher ranked journals tend to report missing values more frequently. However, there seems to be no effect between journal credibility and missing value handling method.

This research includes only studies that have reported the occurrence of missing values. There might be further studies that have encountered and handled missing values but have not reported such occurrences. The real figure of studies with missing values remains therefore unknown.

This contribution shows that researchers tend to exclude missing values and neglect even simple to use handling mechanisms if missing values occur. Inappropriate or insufficient handling mechanisms might yield biased results.

This paper enriches the literature on missing values by comparing the attitude of researchers on handling missing values in open access and high ranked journals.

Keywords: Missing Values, Missing Information, Data Omission, Missing Value Handling.

Handled Missing Values in High-Ranked Journals: Empirical Study on Reasons and Impact" - Based on Literature Review Analysis

Redhwan Amer[1], Malek Simon Grimm[1], Ralf Wagner[1] ([1]: University Kassel)

This paper reviews and assesses recent empirical literature on the incidence of missing values, handling approaches and the impact on the quality of results in marketing research during the period from 2015 to 2018. Furthermore, the importance of exposing missing values in research is specifically outlined and mechanisms of handling missing values in high-ranked journals are provided.

The emphasis of the analysis is only on recent papers of marketing research in an attempt to reflect contemporary knowledge in marketing research with focus on missing values approaches. The paper proceeds with an overview of methods used to handle missing values in the context of marketing. The scarcity of papers which exposed the presence of missing values is illustrated. To evaluate the rate of recurrence of missing values, different analyses are conducted.

Outcomes of this research creates serious concern about the impact of exclusion of missing values on the quality of research. The findings come up with answers to the research questions by providing an objective assessment of the state of knowledge in the missing values. Researchers and practitioners may use the outcomes of this research to understand the impact that the handling of missing values has on the results of their research.

Imputation of missing values is crucial for the quality of research in marketing. Further studies of this issue in other disciplines represent a real need to consolidate the findings of this paper. The outcomes of this paper are related to a limited sample. Researchers are encouraged to conduct similar investigations. but with larger samples on varied respondents.

The selection and analysis of the sample is a tough task because some researchers tend to hide missing values. Due to the variety of research and the diversity of disciplines, a list of keywords to find the missing values, is created.

This paper brings the act of ignoring missing values into the academic debate. Furthermore, an exploration study of recent papers in high-ranked marketing journals and their outcomes are introduced.

Keywords: Missing Values, Missing Information, Data Omission, Missing Value Handling.

Social Network Analysis

Viral Marketing in the Event Industry: An Empirical Study to Determine Factors Influencing the Viral Distribution of Events

Thomas Reichstein[1], Ines Brusch[1], Rebecca Meier zu Ummeln[1] ([1]: BTU Cottbus-Senftenberg)

Digitalisation offers great potential for many areas, including the marketing of events. Newsletters and social media are already being used successfully by companies to draw attention to their events. Viral marketing as a marketing method can help to spread an event via customer recommendations and increase the coverage of events. Thus, viral marketing is an effective and mostly cost-effective form of advertising, which is increasingly used by companies, also in the field of event marketing.

Motives for the distribution of advertising campaigns have already been investigated by researchers such as Hennig-Thurau et al. (2004) in the early days of digital marketing methods. However, there are no studies that have investigated the impact of viral advertising campaigns and their response among social media users (Chu 2011). Especially in the event industry there is a lack of scientific research.

In this context, it is important to understand which factors influence the distribution of event ads. It is well known that images are an important factor in the distribution of content (Kourogi et al. 2015), but how should images be best designed? What emotions should these images convey? The question also arises as to whether external factors such as the headline have an influence and, if so, how strong this influence is. Based on the results of a survey of 154 people and the experiences of an expert, this paper should help to answer the questions asked and provide impulses for further research.

References

Chu, Shu-Chuan (2011): Viral Advertising in Social Media, in: Journal of Interactive Advertising, 12 (1), 30–43.

Hennig-Thurau, Thorsten; Gwinner, Kevin P.; Walsh, Gianfranco; Gremler, Dwayne D. (2004): Electronic word-of-mouth via consumer-opinion platforms: What Motivates Consumers to Articulate Themselves on the Internet? in: Journal of Interactive Marketing, 18 (1), 38–52.

Kourogi, Sawa; Fujishiro, Hiroyuki; Kimura, Akisato; Nishikawa, Hitoshi (2015): Identifying Attractive News Headlines for Social Media., in: Processings of the 24th ACM International on Conference on Information and Knowledge Management, 1859-1862.

The Text Mining and Dimension Reduction Method Application into Exploring the Isomorphic Pressures in the Corporate Communications on the Textual Tweets´ Data on a Sustainability Within the Energy Sector

Adriana Paliwoda-Matiolańska[1], Emilia Smolak-Lozano[2], Atsuho Nakayama[3] ([1]: Cracow University of Economics; [2]: University of Malaga; [3]: Tokyo Metropolitan University)

The study analyses the isomorphic pressures within the context of sustainability by exploring the Twitter communication in the energy sector. Recently, there can be observed an increasing focus on interactive and communicative construction of institution to understand how the organizations sustain the institutional pressures. The rhetorical commitments that create narrative dynamics in organisational communication are central to an institutional diffusion and change. Social Media, Twitter in particular, have been demonstrated as the new opportunity to explore the linguistic dimension in the corporate communications. We propose the use of Social Media linguistic data (tweets with their hashtags and keywords) and the triangulated method (text mining, web mining, and linguistic and content analysis) to examine the tweets´ trends in each company. Based on the institutional theory of organisational communication, the paper examines the relation between the idea of sustainability and isomorphism that leads to the adoption of similar models and attitudes among the organisations. It applies the text mining and correspondence methods within the R software. The energy sector tweets in English (from 2016) were treated by the text mining processes of the statistical linguistic analysis in the R tool. Text mining, involving the linguistic, statistical, and the machine learning techniques, reveals and visualizes the latent structures of the content in an unstructured or weakly structured text data in a given collection of documents. The method helps to represent the topic of a textual document containing a sample of tweets through the frequency study of the semantically significant terms used in these tweets. Document-term matrix has been calculated via text mining technique against the tweet data, then by aggregating it for each company, and representing a word frequency in each company. Since the matrix is sparse and large, it has been necessary to perform a dimensionality reduction analysis to uncover the underlying semantic structure. Dimensionality reduction methods such as: Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, and Non-Negative Matrix factorisation have been found to perform well for this task. Latent Semantic Analysis reduces the dimensionality of the document-term matrix by applying a singular value decomposition, and it then expresses the result in an intuitive and comprehensible form. In the Probabilistic Latent Semantic Analysis, a probabilistic framework is combined with the Latent Semantic Analysis. It has been shown that the Non-Negative Matrix Factorization and Probabilistic Latent Semantic Analysis alike optimise the same objective function, ensuring the equivalent use of both. The Non-Negative Matrix Factorization includes the positive coefficients in the linear combination. The computation is based on a simple iterative algorithm, which is particularly useful for applications involving a complicated linguistic tweets´ matrix. By the results of the analysis, we have clarified the tendency of words used by each company in their tweets, being able to determine the degree of homogeneity in the textual contents of the tweets. The results show the tendency among the energy companies to follow similar patterns in the Twitter communication on the sustainability. Therefore, we can observe the mechanisms that lead to isomorphism in the organisational communication.

An Integration of Time Series Model and Multi-Agent Simulation

Eiji Motohashi[1], Sotaro Katsumata[2], Akihiro Nishimoto[3] ([1]: Yokohama National University; [2]: Osaka University; [3]: Kwansei Gakuin University)

In this study, we integrate time series model and multi-agent simulation and propose an analytical framework for verifying the mechanism of consumers' interaction using purchase history data. Also, we propose a simulation method to optimize advertising plans using our model.

With the spread of UGC (User Generated Content) such as SNS and blogs, the influence of consumers' interaction on purchase behavior has increased. Thus, it has become important for companies to understand the mechanism of consumers' interaction in advertisement planning. On the other hand, since large amounts of data about individual customers are now accumulated, companies can capture precisely consumers' purchase behavior. Moreover, the high-performance cloud environment has been available, so companies have been able to analyze large amounts of data by using a model that requires long time for parameter estimation such as Bayes model.

The proposed model is flexible and complicated, but it can be expressed as a state space model. Since the model incorporates multi-agent simulation and non-gaussian distribution, the Kalman filter, which is the most popular estimation method for state space models, cannot be used. To estimate parameters, we use the particle filter, which is also called the sequential Monte Carlo method and applicable for all types of state space models. As our model includes multi-agent simulation, considering consumers' interaction, it can predict future sales after parameter estimation. In the empirical analysis, we applied the proposed model to actual purchase history data of super market and found that the model is useful to understand consumers' interaction and optimize advertising plans.

Future subjects of this study include the elaboration of the agent behavior, the validation of the parameter, the integrated promotion effect measurement and the application to other categories.

Statistical Learning

The Importance of Being Clustered: Uncluttering the Trends of Statistics from 1970 to 2015

Laura Anderlucci [1], Angela Montanari[1], Cinzia Viroli[1] ([1]: University of Bologna)

The recent history of statistics is retraced by analyzing all the papers published in five prestigious statistical journals since 1970, namely: Annals of Statistics, Biometrika, Journal of the American Statistical Association, Journal of the Royal Statistical Society, series B and Statistical Science. The aim is to construct a kind of ``taxonomy'' of the statistical papers by organizing and clustering them in main themes. In this sense being identified in a cluster means being important enough to be uncluttered in the vast and interconnected world of the statistical research. Since the main statistical research topics naturally are born, evolve or die during time, we also develop a dynamic clustering strategy, where a group in a time period is allowed to migrate or to merge into different groups in the following one. Results show that statistics is a very dynamic and evolving science, stimulated by the rise of new research questions and types of data.

Covariate Selection in Non-gaussian Linear Regression Models: A Bayesian Solution

Giuliano Galimberti[1], Saverio Ranciati[1], Gabriele Soffritti[1] ([1]: Università di Bologna)

Allowing the investigation of the effects of a set of covariates on a numerical outcome of interest, linear regression analysis is one of the most used tools in applied statistics. Most of the inferential procedures associated with linear models rely on the assumption that (i) the error terms follow a Gaussian distribution; (ii) the covariates included in the model contains actually affects the outcome.

These assumptions can be violated in many practical applications, due to the presence of heavy tails, skewness and/or multimodality in the error distribution. Furthermore, the set of relevant covariates may not be known in advance, but its identification can be part of the investigation process.

In this talk, departures from Gaussian distribution and covariate selection are addressed within a Bayesian framework. In particular, a class of linear regression models with errors distributed according to a mixture of Gaussian distributions is introduced, with the choice of relevant covariates embedded in the model specification. By introducing two layers of unobservable latent variables, a hierarchical formulation of such models is provided. A weakly informative modified g-prior for the regression coefficients is elicited, that is a conjugate prior for the likelihood of mixture model's component and induces a form of penalization, thus overcoming potential problems of overfitting. Conjugate prior distributions for the remaining parameters are also proposed, resulting in closed form conditional posterior distributions. Exploiting this formulation, parameter estimation and covariate selection are performed simultaneously, by sampling from the posterior distribution associated with the model. In particular, a Monte Carlo Markov Chain implementation of the sampling procedure is derived, consisting of Gibbs samplers steps based on full conditionals for the model parameters. Since the number of components is held fixed in this MCMC algorithm, the Deviance Information criterion is proposed as a tool to select the optimal number of components.

The proposed methodology is compared with other covariate selection techniques through an extensive simulation study, showing the impact on their performances due to different sources of deviation from the normality, along with the effects of other quantities such as the sample size and the number of relevant/candidate covariates. In particular, the results of this simulation study show that the proposed methodology seems effective in selecting the relevant covariates performs when the distribution of the error terms is characterised by heavy tails, skewness and/or multimodality.

Finally, an application on real data is described.

An Algebraic Estimator for Large Spectral Matrices

Matteo Farnè[1], Matteo Barigozzi[2] ([1]: University of Bologna; [2]: London School of Economics and Political Science)

We present a method to estimate a large p-dimensional spectral matrix assuming that the data follow a dynamic factor model with a sparse residual covariance matrix.

In specific, we apply a nuclear norm plus l_1 norm heuristics to any kernel input estimate at each frequency. We assume that the latent eigenvalues scale to $p^{\alpha}, \alpha \in [0,1]$, and the sparsity degree scales to p^{δ}, with $\delta \leq \frac{1}{2}$ and $\delta \leq \alpha$. We prove that the algebraic recovery of latent rank and sparsity patterns is guaranteed if the smallest latent eigenvalue λ_r and the minimum residual nonzero entry in absolute value min_s are large enough across frequencies. The same holds even allowing the eigenvalues of the factorial coefficients and the sparsity patterns of the residual coefficients to vary across lags.

The consistency of the input is derived via an appropriate weak dependence assumption both on factors and residuals in the sense of Wu and Zaffaroni (2017).

The recovery quality directly depends on the ratio $\frac{p^{\alpha}}{\sqrt{T}}$, where T is the sample length, and the magnitude of T is required to be $p^{\frac{3}{2}\delta}$ or larger. In a wide simulation study, we stress the crucial role of λ_r and min_s across frequencies, highlighting the conditions which cause our method to fail.

Fast and Robust Model Selection Based on Ranks

Wojciech Rejchel[1], Malgorzata Bogdan[2] ([1]: Nicolaus Copernicus University Torun; [2]: University of Wroclaw)

Rank LASSO is an efficient model selection strategy, which allows to identify important predictors under any unknown monotonic link function and any unknown distribution of the error term. However, similarly to the regular LASSO, the consistency of the rank LASSO holds only under a very restrictive

irrepresentable condition on the design matrix and the support of the vector of true regression coefficients. In this talk we will extend the scope of applications of rank LASSO by considering its thresholded and reweighted versions. We will present theoretical results on the consistency of these procedures under weak assumptions on the design matrix and the signal sparsity and demonstrate results of extensive simulation study illustrating the efficiency and usefulness of proposed methodology.

Statistics and Data Analysis

Interpretability of Prediction Intervals for Neural Net

Claus Weihs[1], Malte Jastrow[1] ([1]: University of Dortmund)

Generally, the unknown coefficients of neural nets are estimated by nonlinear least squares. Therefore, prediction intervals for the true value of the target feature exist at least asymptotically. Such intervals depend on the Hessian of the estimator of the unknown coefficients. Unfortunately, it is well-known that such Hessians are often ill-conditioned in practice because of the flatness of the tails of the activation function of the neural net (e.g. Saarinen et al., 1993). Therefore, prediction intervals can be extremely broad and little interpretable.

The paper discusses the interpretability for different activation functions in the regression and the classification case.

References

Saarinen, S., Bramley, R., Cybenko, G. (1993): Ill-conditioning in Neural Network Training Problems; SIAM J. Sci. Comput., 14(3), 693-714

Sense and Sensibility: Measuring the Predictive Significance of Classifiers

Ulrich Müller-Funk[1], Stefanie Weiß[2] ([1]: Westfälische Wilhelms-Universität Münster; [2]: Deutsche Telekom)

A myriad of metrics, often fancifully labelled, have been proposed for the assessment of classifiers. Accordingly, R/Python-packages on the subject offer an undifferentiated list of such indices. Typically, not only a single one is selected, but a bunch of them is made the basis for comparisons. A closer look at these metrics reveals, however, that they serve different purposes, express different attitudes - or are just meaningless. On the other hand, it is noticeable that only few textbooks on statistical learning or data mining incorporate metrics for evaluation. Authors who do so, provide a description of these indices, but hardly a justification.

In this contribution we shall confine ourselves to the predictive significance of binary classifiers, i.e. disregard its capability to determine class probabilities. We shall start out from the observation, that every real-valued Borel function of the predictors selected gives rise to a classifier that maximizes accuracy under some family of distribution functions (df) but minimizes it under another one. Consequences: 1) There is nothing like an "omnibus procedure" i.e. a classifier that always behaves reasonably well. 2) Benchmark data - representing some specific, unknown df - might be nice for illustrations but lack any meaning w.r. to evaluation. We point out, that predictive accuracy , moreover, is of little importance without stability -and sometimes interpretability.

The discussion of risk measures based on both the error probabilities is a major concern of the paper. We shall present a graphical representation - supporting the choice of a classifier out of the available tool-kit. The approach is related to precision and recall curves resp. the ROC. Accordingly, we shall scrutinize the meaning of these concepts. Some other metrics are touched upon as well.

On the Use of the Imprecise Dirichlet Model in the Case of Missing Data

Aziz Omar[1], Thomas Augustin[1] ([1]: LMU Munich)

In a multinomial setting, it is a custom to follow the Bayesian framework to estimate probabilities of the possible outcomes. While this has proven quite powerful when substantial prior information is present, the proper handling of complete prior ignorance is still understood as a big challenge. While different vague priors have been introduced to serve this purpose, Walley (1996, JRRSB) has introduced the imprecise Dirichlet model (IDM) as a substantially different approach to express prior ignorance in multinomial experiments. In the IDM prior ignorance is reflected by considering the set of all non-degenerate Dirichlet priors that have a common “strength” as a hyperparameter that expresses the velocity of learning from new observations. The employment of the IDM yields a set of posteriors. Thus, after obtaining the sample inference regarding the chance of a specific possible outcome is expressed as an interval-valued estimate representing the effect of the prior ignorance.

To use the full power of the IDM in applications, it is important to generalize it to situations of imperfect - potentially missing, coarsened or error-prone - data. However, the work by Piatti et al (2005, ISIPTA) warns us that some subtle, quite surprising difficulties may occur in such settings. They showed that an analysis of potentially misclassified data produces vacuous results, i.e. the resulting interval-valued estimate for the chance of every outcome ranges from 0 to 1.

In our work, we investigate whether the same vacuous interval-valued estimates would be the output in the case of partially observed data. We begin our study by the simple case of considering a binomial experiment and then examine the more general multinomial case. The final results of our work should have an impact on the practical use of the IDM in applications involving partially observed categorical data.

Stream Mining 1

End-to-End Motion Classification Using Smartwatch Sensor Data

Torben Windler[1], Junaid Ahmed Ghauri[1], Muhammad Usman Syed[1], Tamara Belostotskaya[1], Valerie Chikukwa[1], Rafael Rego Drumond[1], Lars Schmidt-Thieme[1] ([1]: University of Hildesheim)

Analysis of smart devices' sensor data for the classification of human activities has become increasingly targeted by industry and motion research. With the popularization of smartwatches, this data becomes available to everyone. The data from the accelerometers and gyroscopes are conventionally analyzed as a multivariate time series to obtain reliable information about the wearer's activity at a specific moment. Due to the particular sampling rate instabilities of each device, previous approaches mainly work with feature extraction methods to generalize the information independently of the gear, which requires a lot of time and expertise. To overcome this problem, we present an end-to-end model for activity classification based on convolutional neural networks of different dimensions without extensive feature extraction. The data preprocessing is not computationally intensive, and the model can deal with the irregularities of the data. By representing the input as twofold - both interpolated 1D time-series and encoded time-series as images with the help of Gramian Angular Summation Fields - the use of techniques from computer vision is enabled. In addition, an online prediction is possible, and the accuracy is comparable to feature extraction approaches. The model is validated with random 10-fold and leave-one-user-out cross-validation and shows improvement regarding the generalization of the task.

Statistical Analysis of Unauthorized Internet Access Log Data and its Interpretation

Hiroyuki Minami (Hokkaido University)

We have studied how to analyse unwilling network access logs. Our pre-analyses and empirical studies have suggested that we could classify the logs into some typical patterns and tried to develop a method to reveal them with aggregated statistical approaches, mainly symbolic data analysis (SDA).

In the previous conference, ECDA 2018, we introduced an application to analyse the logs with interval valued data whose elements consist of the number of the times about co-occurrence with IP address subsets (called CIDR; Classless Inter-Domain Routing). It requires plenty computer resources to count each event up. Moreover, the approach doesn't take types of transmission into consideration.

Internet-related data representation is familiar with SDA. For example, we can regard CIDR notation which represents a range of IP addresses (e.g. 192.168.0.0/24 stands for a range from 192.168.0.0 to 192.168.0.255; 256 addresses), simply as an interval in SDA.

A log record has a time-stamp, a pair of source IP address and a port number to establish virtual transmission circuit and a destination pair whose port number usually identifies a type of transmission. For instance, destination port number 80 means HTTP.

We have found that some records have a few specific source port numbers while an operating system usually assigns them by random integers. Thus, it is straightforward that the records would be brought by some unwilling software whose source port number is hard-coded.

Our motivation is to extract some specific patterns from the logs including the cases by the unwilling software. Many applications have been already developed to detect anomaly from them, but few are mainly based on statistics. To improve their quality, a mathematical viewpoint is a key idea since most unwilling actions are based on automatic algorithms, thus we could apply some statistical (and intensive) model to them. When we develop an intensive statistical analysis to the data, SDA, known as a typical aggregated data analysis method would be suitable.

In the study, we discuss how we reorganize the logs into the SDA world and derive a desirable output and interpretation of our real data.

Comparison of Decision Tree Algorithms for Data Streams with Concept Drift

Jennifer Neuhaus-Stern[1], Sarah Schnackenberg[1], Uwe Ligges[1] ([1]: TU Dortmund University)

The goal in classification is to build a model to predict one of a finite number of class labels for new observations. Traditional methods use a static data set to train this model, assuming that all observations come from a fixed but unknown distribution and that the data set contains all necessary information about this distribution. However, these assumptions are not justified for many real-world problems as they often appear as data streams. Important information can emerge over time or the underlying distribution is non-stationary and changes over time, what is called concept drift. In both cases, the model needs to be updated continuously. For some classification methods this can be done with little effort, for others like decision trees, however, completely new approaches must be developed.

In the last years, many algorithms for building decision trees on data streams with concept drift were developed, mainly based on the so-called Hoeffding trees, developed by Domingos and Hulten (2000) and their VFDT (Very Fast Decision Tree) algorithm. However, the question arises whether it is necessary to use such complex algorithms or if it suffices in most situations to extend classical algorithms such as the CART (Breiman et al. (1984)) with windowing techniques.

In our work, we compare the VFDT algorithm, its extension, the CVFDT algorithm (Hulten et al. (2001)) and the CART algorithm on sliding windows. We get the result that the classical algorithm can keep up with the current algorithms and even provides better results for some data situations.

References

Breiman, L., Friedman, J., Olshen, R., Stone C. (1984): Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks.

Domingos, P., Hulten, G. (2000): Mining High-Speed Data Streams. KDD. ACM, 71-80.

Hulten, G., Spencer, L., Domingos, P. (2001): Mining Time-Changing Data Streams. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD `01. San Francisco, California: ACM, 97-106.

Stream Mining 2

Topic Modelling for Summarizing Industrial Log Data

Benjamin Klöpper[1], Shunmuga Prabhu Siddharthan[1], Barbara Sprick[2], Marcel Dix[1] ([1]: ABB Corporate Research; [2]: SRH Hochschule Heidelberg)

Most industrial devices generate (semi-structured) log data, that documents events along with a time-stamp. For many tasks like root-cause-analysis, asset or process monitoring this log data is the starting point for human analysis. However, the data is ill fit for human analysis for various reasons: too many events, ordering by time not causality, and a high number of common and uninteresting events (e.g. program start/stop).

A learning from practical experience is that industrial log data and the applications requirements have the following characteristics: (1) interesting and important events happen very seldom resulting in heavily unbalanced data sets, (2) data can usually not be compared across many different individuals or installations, (3) labeled data for supervised learning is usually not available, and (4) the requirements of the industrial user regarding predictive machine learning performance is very high (~99.9% accuracy in our experience).

Summarization of logs should improve the understanding of the system that generates the data. However, the current approaches offer piecewise insights which lacks global structure and demands joint development between domain experts and data scientist. Most of the existing work on machine learning for log files focusses on supervised techniques, mining frequent patterns and Hidden Markov Model (HMM). Supervised techniques are unsuitable due to the lake of labeled data. Frequent episodes or patterns often fail to capture rare but important information like failures or produce a too large number of patterns. HMM implementation requires prior domain knowledge and understanding of system states in the first place.

We address the challenges associated with log data using an unsupervised learning method. The technique of topic modelling using Latent Dirichlet Allocation is utilized as a possible solution for analyzing industrial event log.

Topic modeling makes assumptions on the data. If the (preprocessed) data does not adhere to these assumptions, then the trained model does not deliver interesting results. To overcome this problem, the proposed solution is implemented in an iterative process: (1) A vocabulary is built from semi-structured log data (2) Collections of entries (documents) are extracted from the log data (based on temporal criteria like operation days or time between entries) (3) Creating topic models for different number of topics (4) The quality of topic models is assessed with the help of KPIs (5) The topics in the model are analyzed, visualized and labelled (6) The resulting topic model is evaluated (7) The vocabulary is refined (identification of ``stop words´´).

This approach supports users to monitor industrial systems by identifying topics in event log as episodes. The iterative process and the resulting topic model have been tested on a real-world industrial data set from robotics automation and evaluated using direct annotation of topic assignment by domain experts for about 300 documents (days of event log) with a good average score of 2.342 (range 0-3). An illustrative visualization and application demonstrates how topic modelling could be used to monitor hundreds of industrial robots in a factory.

The work has been supported by the EU ECSEL project Productive 4.0.

Modeling PLM Data to Optimize Car-Dealership Offerings

Dieter William Joenssen (HS Aalen)

Analyzing PLM (product lifecycle management) data is a task whose complexity grows with the product at hand. Drivers for this complexity are as diverse as the management challenges in creating the product data. Challenges include the temporal variability of product offering, technical solutions, production capabilities and human factors. Nonetheless, value may be gained by applying standard statistical methods, while accounting for peculiarities on the data side.

Arguably the most complex PLM data is available in the automotive industry (contending with space and aeronautic industries), where the products can be configured with a plethora of options. Here the application of statistical methods offer particular value, as human understanding undoubtedly will fail at discerning relevant interaction effects at higher dimensions. To showcase the application of these methods, we use a unique data set from a large OEM using a selection of PLM data: “Are car configurations from dealerships fundamentally different than those of consumers?”

The question is significant for management in multiple senses. First, certain markets do not allow for free consumer configuration. Thus, dealer offerings may be optimized using analogies from other markets. Secondly, dealer offerings may be optimized when contrasting against consumer configurations within the same market.

To answer the question, we use discrimination techniques on properly formatted data. Results yield a statistical model that can be utilized to benchmark configurations and recommend changes to selected options. Applied to existing data, the model indicates that different markets would benefit differently from optimization. Large markets with a tradition of dealership purchasing can benefit from cost saving measures (e.g. remove options from the base package). Small markets, in turn, may realize turnover potential (e.g. include more options in the dealer’s configuration). Limitations of the approaches used are discussed and a direction for further research is offered.

Cognitive Bayesian Classification Based on Quantum Theory

Ingo Schmitt (BTU Cottbus-Senftenberg)

Bayesian classification is a very prominent classification method based on conditional probabilities and the famous axiom from Bayes. Probability values are derived from training data and stored explicitly as values or implicitly by learnt parameters of a distribution function. The main idea of that work is to reformulate the Bayesian classification on concepts of quantum theory. By doing that, we strive for modeling scenarios with human interactions better than the classical method does.

In quantum theory probability values are obtained from quantum measurements on a normalized state vector of a vector space (see Gleasons theorem and Born rule). Here we focus on real and finite-dimensional vector spaces. Quantum theory combines very elegantly concepts from probability, geometry (linear algebra) and logic. A quantum measurement requires a projector p (Hermitian with the property p*p=p). Every projector corresponds to a vector subspace. For measurement the state vector is projected onto the vector subspace of a projector. Thus, the semantics of a probability measure is expressed by an event-specific projector to be learnt from training data. Learning projectors is performed by a gradient descent method on a well-defined loss function. Conditional probabilities are defined on multiplying two projectors. Both projectors P1 and P2 may commute, that is P1*P2 = P2*P1, which can be seen as P1 AND P2 = P2 AND P1. In that classical case the Bayesian axiom holds directly. Otherwise we obtain a so-called order effect of quantum measurement which is well-known for explaining many surprising effects in quantum physics. Furthermore, as Busemeyer and Bruza state, cognitive effects like conjunction and disjunction fallacies based on human interactions can be explained by non-commuting projectors. In that way, cognitive Bayesian classification generalizes the classical Bayesian classification by considering ordering effects. Of course, with means of a loss function it is possible to enforce the property of commuting projectors while learning them, but semantics may get lost. Otherwise, the degree of the property non-commutativity of two learnt projectors can be measured and can help to explain and to predict non-classical order effects between different class events.

The paper motivates the approach of using quantum theory for reformulating Bayesian classification. Furthermore, we develop the process of feature data encoding, learning of projectors and classifying

unclassified objects. The discussion and evaluation of small experiments will show the potential of that approach to consider non-classical effects from human interactions which go beyond the classical Bayesian classification approach. Further extensive studies are necessary in order to show the feasibility of that approach for a broad range of applications.

Structural Equation Models in Marketing 1

Technology Acceptance Model or Uses and Gratifications Approach: Which Approach Is Better Suited to Explain the Acceptance of Digital Voice Assistants?

Karolina Ewers[1], Daniel Baier[1] ([1]: University of Bayreuth)

Some studies predict, that voice-based communication in natural language will dominate human-computer-interaction in the future. Main arguments are the rapid progress in text processing as well as cloud computing speed and the improved convenience compared to button/icon clicking or syntax-based commands. Other studies are more skeptical and refer to the (still) low capabilities of chatbots to simulate humans. However, the studies up to now, which are explorative, model-based analyses that test the influencing acceptance factors, are rare.

In order to close this gap, this paper investigates the acceptance of a digital voice assistant for a selected use case basing on the technology acceptance model (TAM) as well as the uses and gratifications approach (UGA). The selected use case for this investigation is voiced-based navigation for a pedestrian when trying to better understand a more or less familiar urban area.

A within-subject design was used for comparing the two approaches (TAM and UGA): A student sample (n=173) was asked to use the installed voice-mode of their smartphone‘s digital voice assistant (all used Google Assistant) when walking around in the city where they study and to finish various tasks (e.g. find specified places, discover the city’s history and the achievements of its inhabitants). Their activities were tracked and after finishing their tasks they filled out a pre-tested questionnaire that followed the two approaches with acceptance as well as influencing factors and measurement models developed according to a TAM and UGA literature review.

A SmartPLS analysis of the collected data showed, in the context of UGA, a positive influence of interest in pastime and enjoyment on actual system use. There is also a positive influence of attitude towards using on behavioral intention to use and actual system use. Various barriers (for example: fear of distraction, fear of data misuse) do not play a significant role. Within the framework of the TAM, perceived usefulness has a positive influence on attitude towards using and behavioral intention to use. Furthermore, all TAM constructs have a positive effect on actual system use. A structural equation model employed in the current study demonstrates that users of digital voice assistants already rely on them mainly to gather brief information, navigate, ask about the weather and for travel tips. Customers use Google Assistant particularly if they cannot use their smartphones manually. In addition, most respondents will use Google Assistant in the future while driving a car or simply to receive short and basic information as quickly as possible. To answer the question, which of the two models (TAM or UGA) suits better, three requirements were looked closer at: the coefficient of determination of dependent variables, the information content and the adaption to the research context. Although both approaches are reliable acceptance theories, UGA provides specific information and a more exact understanding of the usage of digital voice assistants, whereas TAM constructs can be used quickly and easily. The findings suggest integrating both theoretical approaches to create an integrated model that predicts acceptance, usage and satisfaction.

Investigation of Student Satisfaction and Student Loyalty – A Hierarchical Latent Variable Model

Sarah Maria Wruck[1], Winfried J. Steiner[1] ([1]: Clausthal University of Technology)

During the last twenty years the competition in the local and global higher education market has intensified. Educational institution's competitive success is, e.g., measured by successful students or institution's high reputation in research and teaching. Hence, one of the core targets of universities is to attract and retain high-potential students. To meet the student's needs, universities have to gather sufficient information about how their students evaluate their study conditions and which factors influence and explain student loyalty. Nowadays, universities - similar to companies – thus have to assure their existence by implementing and controlling certain management ratios.

As discussed in the recent marketing literature, tertiary education is comparable to a service under certain conditions. Therefore, the management of higher education must be examined from a customer-orientated view, i.e., customer satisfaction constitutes a key research field. In this context, some researchers draw upon the European or American Customer Satisfaction Index (ECSI/ACSI). In turn, others attempt to replicate the SERVQUAL–, SERVPERF– or HEdPERF–scale if they try to explore the construct of "student satisfaction".

In order to substantially contribute to the explanation of the latent construct "student satisfaction", we conduct an empirical study at a German university. Within our questionnaire, we combine common constructs and items from different literature streams. Based on the data of round about 530 respondents, we further investigate the relationship between "student satisfaction" and "student loyalty" under consideration of additional drivers like the "image" of a university.

We use a hierarchical latent variable model for "student satisfaction". The developed model mainly consists of latent variables with an underlying formative measurement model. All first order latent variables, which form the third order construct "student satisfaction", are measured by formative multiple item scales. In contrast, the endogenous construct "student loyalty" is measured by several reflective items. Moreover, we try to model some moderators, e.g., "commitment" and "involvement" of students, to determine the relationship between "student satisfaction" and "student loyalty" even more precisely.

The results of our empirical study indicate that there are two second order domains which build the third order construct "student satisfaction". The second order construct "quality of study conditions", as one of the above mentioned domains, is composed of six first order latent constructs. Out of these first order constructs the construct named "courses" seems to have the widest influence on the construct "quality of study conditions". The results also reveal that the construct "student loyalty" is not only affected by "student satisfaction" but also by the constructs "image" and the location of the university.

Keywords: Student Satisfaction, Student Loyalty, Partial Least Squares Structural Equation Modeling, Formative Measurement Models, Hierarchical Component Model.

References

Abdullah, Firdaus (2006): The Development of HEdPERF: A New Measuring Instrument of Service Quality for the Higher Education Sector, International Journal of Consumer Studies, 30 (6), 569–581.

Hair, Joseph, F., Jr., Sarstedt, Marko, Ringle, Christian M., Gudergan, Siegfried P. (2018): Advanced Issues in Partial Least Squares Structural Equation Modeling, LA: Sage.

Discovering Kano's Model in Service Satisfaction Analysis Using Cubic Terms

Björn Stöcker[1], Aydin Nasseri[2] ([1]: BAUR Versand GmbH & Co KG; [2]: COGITARIS Gesellschaft für Marktforschung GmbH)

Ever since Noriaki Kano's research, we have known that the relationship between performance and customer satisfaction is not just linear. Depending on the performance, different customer requirements exist, which are visualized in the Kano Model with three curves. The model is intended to make clear that the increase in e. g. service fulfilment does not automatically generate more customer satisfaction and that a lack of service fulfilment does not automatically lead to dissatisfaction. These characters are represented in the model with two non-linear functions.

There are currently two leading methods for determining the specific characteristics of services. In the first method, each service is determined with two questions (functional / dysfunctional). The character is determined by means of a simple frequency count. In the second method Penalty Reward Analysis (PRA), the relationship between positive and negative fulfilment is calculated separately for customer satisfaction (using driver analysis) in order to capture both linear and non-linear relationships. However, the PRA has to struggle with some limitations: (I) the definition of high and low performer is not standardized, (II) linear correlations are again determined in the model itself, (III) by forming the dummy variables, the middle of the Likert response scale is ignored and (IV) the explanatory contribution is lost. It is also assumed that (V) a vertex can only occur in the middle of the scale.

In this article, we would like to present another method that models Kano characteristics for different services using a cubic term. Using the example of a service quality study of an online retailer, we were able to demonstrate that the relationships between service fulfilment and customer satisfaction can be very well estimated using cubic terms and that the character can be named according to Kano. We compare the results of the PRA, the more recent PRFA approach and the cubic terms and recommend how the cubic terms can be interpreted.

On the basis of two samples of an online retailer, we measure the satisfaction of the service quality. Two samples from 2011 and 2013 were analysed for this paper. The qualitative data sets collected via CATI (n= 480 and n=500), each by using the same standardised questionnaire.

How to Motivate a Reviewer? Adoption of CRM Strategies to Implement a Successful Relationship Between a Journal and a Reviewer

Victoria-Anne Schweigert[1], Andreas Geyer-Schulz[1] ([1]: Karlsruhe Institute of Technology)

There are four main participants (“clients”/”user group”) in the daily business process of a scientific journal – the readers, authors, reviewers and editors. Especially, the group of the reviewer is a very interesting group, because this person is expected to work for the journal without a typical reward like a salary. But without this well qualified researcher the scientific process is not possible. This contribution considers the questions:

1. How to motivate a reviewer to review a scientific article?
2. Is it possible to adopt assessed strategies from the customer relationship management (CRM) research to develop guidelines to motivate this user group?

This contribution consists of three parts, namely first a state-of-the-art overview about literature on reviewer motivation in different areas and second a short analysis of reviewer’s behaviour in the context of the journal Archives of Data Science. And finally, we will discuss trade-offs with the reviewer incentive systems.

In this part of the article we will check the applicability of the CRM concept of an internal customer to the review process.

Keywords: Reviewer Motivation, CRM, Scientific Publishing.

Structural Equation Models in Marketing 2

Value Creation in Fashion Retailing: Empirical Findings of a Jobs-To-Be-Done Framework Application

Franziska Kullak[1], Daniel Baier[1], Herbert Woratschek[1] ([1]: University of Bayreuth)

Currently, many online and offline fashion retailers are challenged by a large number of potential improvements enabled by new technologies along the ordering and delivery process: Augmented reality applications as well as (digital) shopping assistants improve the selection of appropriate clothing, (return) delivery systems the transportation, or (digital) service assistants the handling of inquiries and complaints. In order to select “best” alternatives for improvement out of many, some retailers

(e.g. Amazon, Zalando, BAUR or OTTO group) have installed a step-by-step filtering process that mainly consist of asking customers which alternatives they prefer and by implementing and testing most preferred ones. This filtering process – in online-shopping called site engineering process – resembles the new product development approach that traditional producers of consumer goods like, e.g., Procter&Gamble apply for a long time.

However, recently, this filtering process has been criticized as following a “Goods Dominant Logic" (GDL), since it only simulates customer centricity but in contrast is mainly firm centric by focusing on available alternatives (see, e.g., Bettencourt 2010, 2013; Bettencourt et al. 2014). Instead, it is proposed to change the perspective and focus on value creation for all actors (also for the customers), a “new” approach that has the potential to shed light on radically new alternatives has been called “Service-Dominant Logic" (SDL) (Vargo & Lusch 2004).

In this paper, we discuss how this SDL approach can be implemented in fashion retailing. The so-called Jobs-To-Be-Done (JTBD) framework is applied. Focal points of this approach are functional, personal or social “jobs” (needs, problems) customers want to be done (fulfilled, addressed) (Christensen et al. 2007). First, these jobs must be identified before better products or services should be developed. We apply JTBD to identify relevant jobs for online shops and in-store shopping and derive improvements. We conduct two qualitative studies, one with in-depth interviews with three focus groups (N=18) and one with fourteen face-to-face in-depth interviews for in-store shopping and online shopping.

The results viewed from a GDL perspective show that in the case of in-store shopping, personal and social jobs like taking time-out from daily duties or maintaining social contacts are of prime importance. The main potential for innovation in offline fashion retailing therefore lies in autonomous shopping, whilst restructuring of online shops, based on specific needs of customers, should be considered to enhance the overall shopping experience. Therefore, innovation potential may be exploited by focusing on the customer´s mood and its inquiry on the online shop. Online shops should not only be seen as a shopping platform but beyond that as entertainment medium.

Findings analysed from a SDL perspective show that innovative technologies like intelligent dressing rooms are one type of integrated resources, among others, to create value for the customer. However, our results underpin that social interactions also drive the shopping experience. Therefore, omni-channel management should not only focus on the development of technological innovations, but also on the integration of different resources in the purchase process to increase customer experience as overarching goal.

SEM-Tree Hybrid Models in the Preferences Analysis of the Members of Polish Households

Adam Sagan[1], Mariusz Lapczynski[1] ([1]: Cracow University of Economics)

The aim of the paper is to identify the dimensions of the resource allocation strategies of the members of Polish households. These dimensions were identified on the basis of nationwide empirical data gathered on a representative sample of 1100 respondents nested in 410 households.

SEM-Tree hybrid models are used in the analysis of the results, which combine the confirmatory structural equation models with exploratory and predictive classification and regression trees. This allows to apply structural modeling for the study of heterogeneous populations and to assess the hierarchical impact of exogenous predictors on the identification of segments with separate and unique model structural parameters.

The approach combines the advantages of a model approach (at the stage of constructing hypotheses on structural relationships and specifications of measurement models) and exploration-based data (at the stage of recursive division of the sample).

In the analysis mixture SEM models and SEM-Tree hybrids will be developed and compared.

Keywords: Customers' Attitudes, Structural Equation Models, Decision Trees, Hybrid Models.

References

Brandmaier A.M., Oertzen T., McArdle J.J., Lindenberger U. (2013a): Structural Equation Model Trees, Psychological Methods, vol. 18, no. 1, 71–86.

Brandmaier A.M., Oertzen T., McArdle J.J., Lindenberger U. (2013b): Exploratory Data Mining with Structural Equation Model Trees, [w:] McArdle J.J, Ritschard G. (red.), Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences, Routledge, New York, s. 96-127.

Marcoulides G.A., Ing M. (2012): Automated Structural Equation Modeling Strategies, in: Hoyle R.H. (ed.), Handbook of Structural Equation Modeling, Guilford Press, New York.

Steinberg D., Cardell N.S. (1998): The Hybrid Cart-logit Model in Classification and Data Mining, Eighth Annual Advanced Research Techniques Forum, American Marketing Association, Salford Systems, s. 1-7.

Change Management and Acceptance of CRM Systems: An Analysis by Means of Partial Least Squares

Christian Wehowski[1], Heike Papenhoff[1], Karsten Lübke[1] ([1]: FOM University of Applied Sciences)

User Acceptance is known to be one of the factors that are linked to a successful Customer Relationship Management (CRM) System implementation (Avlonitis & Panagopoulos 2005). In order to improve this aspects of Change Management should be considered (Becker et al. 2009; Greve & Albers 2006; Avlonitis & Panagopoulos 2005). This study investigates the contribution of Change Management instruments improving the acceptance of CRM Systems and by this the success of the CRM implementation in a firm (Chen & Popovich, 2003).

Therefore we investigated the link of Change Management aspects like Communication (Bueno & Salmeron, 2008), Training (Kohnke & Müller, 2010; Amoako-Gyampah & Salam, 2003) and Participation (Kohnke & Müller, 2010) with the Unified Theory of Acceptance and Use of Technology: Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions and to the end Behavioral Intention (Venkatesh et al., 2003, 2008, 2012) by a Structural Equation Modeling using Partial Least Squares.

With a sample of n=103 users of CRM Systems the PLS Analysis indicates a positive link between Participation and Training on Behavioral Intention. We also discuss model and measurement diagnostics as well as managerial implications suggested by the more detailed analysis.

3 Index of Authors

ARCHIVES OF DATA SCIENCE
SERIES A

www.ArchivesofDataScience.org