Book of Abstracts

ECDA 2018

EUROPEAN
CONFERENCE ON
DATA ANALYSIS
Paderborn | Germany

04 - 06 July

Multidisciplinary Facets of Data Sience

Programme Chairs

Hans Kestler	University of Ulm, Germany
Adalbert Wilhelm	Jacobs University Bremen, Germany

Programme Committe

Stefan van Aelst	KU Leuven, Belgium
Casper Albers	University of Groningen, Netherlands
Martin Atzmüller	Tilburg University, Netherlands
Thomas Augustin	LMU Munich, Germany
Rolf Biehler	Paderborn University, Germany
Daniel Baier	University of Bayreuth, Germany
Bernd Bischl	LMU Munich, Germany
Ulf Brefeld	Leuphana University of Lüneburg, Germany
Claudio Conversano	University of Cagliari, Italy
Reinhold Decker	Bielefeld University, Germany
Sebastian Destercke	University of Technology of Compiègne (UTC), France
Florian Dumpert	University of Bayreuth, Germany
Ralph Ewerth	TIB, Leibniz Universität Hannover, Germany
Mohsen Farid	University of Derby, United Kingdom
Peter Flach	University of Bristol, United Kingdom
Johannes Fürnkranz	TU Darmstadt, Germany
Michaela Geierhos	Paderborn University, Germany
Andreas Geyer-Schulz	KIT Karlsruhe, Germany
Daniel Guhl	HU Berlin, Germany
Barbara Hammer	Bielefeld University, Germany
Dominik Heider	University of Marburg, Germany
Christian Hennig	University College London, United Kingdom
Tadashi Imaizumi	Tama University, Japan
Salvatore Ingrassia	Catania University, Italy
Krzysztof Jajuga	Wrocław University, Poland
Wolfgang Konen	TH Köln, Germany
Georg Krempl	Utrecht University, Netherlands
Koji Kurihara	Okayama University, Japan
Berthold Lausen	University of Essex, United Kingdom
Xiaohui Liu	Brunel University, United Kingdom
Volker Lohweg	Ostwestfalen-Lippe University of Applied Sciences, Germany
Eneldo Loza Mencia	TU Darmstadt, Germany
Felix Mohr	Paderborn University, Germany
Angela Montanari	University of Bologna, Italy
Emmanuel Müller	University of Potsdam, Germany
Fionn Murtagh	University of Huddersfield, United Kingdom
Mohamed Nadif	Paris Descartes University, France
Axel Ngonga	Paderborn University, Germany
Oliver Niggemann	Ostwestfalen-Lippe University of Applied Sciences, Germany
Friederike Paetz	University of Clausthal, Germany
Józef Pociecha	Cracow University of Economics, Poland
Niel le Roux	Stellenbosch University, South Africa
Lars Schmidt-Thieme	University of Hildesheim, Germany
Frank Scholze	KIT Karlsruhe, Germany
Carsten Schulte	Paderborn University, Germany
Jerzy Stefanowski	Poznan University of Technology, Poland
Kevin Tierney	Bielefeld University, Germany
Alfred Ultsch	University of Marburg, Germany
Maurizio Vichi	Sapienza University of Rome, Italy
Henning Wachsmuth	Paderborn University, Germany
Marcel Wever	Paderborn University, Germany

EMPOLIS
INFORMATION MANAGEMENT

Springer

SICP
Software Innovation Campus Paderborn

PC²

SFB901
ON - THE - FLY COMPUTING

DFG Deutsche Forschungsgemeinschaft

HEINZ NIXDORF INSTITUT
UNIVERSITÄT PADERBORN

Contents

1 Plenary Talks

Disentangling the thoughts: Latest news in computational argumentation

Iryna Gurevych

In this talk, I will present a bunch of papers on argument mining (co-)authored by the UKP Lab in Darmstadt. The papers have appeared in NAACL, TACL and related venues in 2018. In the first part, I will talk about large-scale argument search, classification and reasoning. In the second part, the focus will be on mitigating high annotation costs for argument annotation. Specifically, we tackle small-data scenarios for novel argument tasks, less-resourced languages or web-scale argument analysis tasks such as detecting fallacies. The talk presents the results of ongoing projects in Computational Argumentation at the Technische Universität Darmstadt: Argumentation Analysis for the Web (ArguAna), Decision Support by Means of Automatically Extracting Natural Language Arguments from Big Data (ArgumenText) .

Analysis of data from educational achievement tests with generalized linear mixed models

Johannes Hartig

Generalized linear mixed models (GLMM) allow modeling clustered non-normal data with both fixed and random effects. GLMMs include specific models that are traditionally taught as separate topics in educational science and psychology. For instance, hierarchical linear models (HLMs) and many models from item response theory (IRT) are special cases of GLMMs. Thus, GLMMs provide a flexible framework to combine separate lines of research and address new research questions. The talk will illustrate the potential of GLMMs for educational research with empirical applications in the analysis of data from educational assessments. This data has a hierarchical structure with item responses nested in students nested in classrooms and / or schools. Research questions can focus on variables on all levels. Examples include effects of item level predictors, effects of the item position within an assessment, and effects of school level predictors. Challenges regarding the analysis of data sets from international large scale assessments will be discussed.

Finite mixtures for simultaneous clustering and reduction of matrix value observations

Roberto Rocci

Finite mixture models are often used to classify two-way (units by variables) data. However, two issues arise: model complexity and recovering the true clustering structure. Indeed, a huge number of variables and/or occasions implies a large number of model parameters, while the existence of noise dimensions could mask the true cluster structure. The approach adopted in the present work is to reduce the number of model parameters by identifying a sub-space containing the information needed to classify the observations nesting a PCA-like reparameterization into the classification model. This also helps in identifying and discarding noise dimensions that could coincide with some observed variables. The aforementioned problems become more serious in the case of three-way (units by variables by occasions) data where the observations are matrices rather than vectors. Our approach is extended to this case by nesting a three-way PCA-like reparameterization (named Tucker2) into the classification model. This allows us to reduce the number of parameters identifying noise dimensions for the variables and/or the occasions. The effectiveness of the proposals is assessed through a simulation study and an application to real data.

Kernel-based methods in machine learning

Andreas Christmann

Machine learning and big data analysis play an important role in current research in statistics, computer science, computational biology, engineering, and in many other research areas. Although there exist a huge number of different approaches in machine learning, almost of all them have two goals in common: universal consistency and applicability in real life situations where the knowledge on the data generating process and on the data quality is limited or non-existent. In the first part of the talk, I will give a short overview of some currently used machine learning approaches, with special emphasis on kernel based methods. In the second part, some recent results on universal consistency, learning rates, statistical robustness, and stability of kernel based methods will be given. Some results will also be given for the bootstrap approximations of such kernel based methods.

Can we automate data science?

Luc de Raedt

Inspired by recent successes towards automating highly complex jobs like automatic programming and scientific experimentation, I want to automate the task of the data scientist when developing intelligent systems. In this talk, I shall introduce some of the involved challenges of and some possible approaches and tools for automating data science. More specifically, I shall discuss how automated data wrangling approaches can be used for pre-processing and how both predictive and descriptive models can in principle be combined to automatically complete spreadsheets and relational databases. Special attention will be given towards the induction of constraints in spreadsheets and in an operations research context.

Gaussian process emulation of computer models with massive output

James Berger

Often computer models yield massive output; e.g., a weather model will yield the predicted temperature over a huge grid of points in space and time. Emulation of a computer model is the process of finding an approximation to the computer model that is much faster to run than the computer model itself (which can often take hours or days for a single run). Many successful emulation approaches are statistical in nature, but these have only rarely attempted to deal with massive computer model output; some approaches that have been tried include utilization of multivariate emulators, modeling of the output (e.g., through some basis representation, including PCA), and construction of parallel emulators at each grid point, with the methodology typically based on use of Gaussian processes to construct the approximations. These approaches will be reviewed, with the startling computational simplicity with which the last approach can be implemented being highlighted and its remarkable success being illustrated and explained; in particular the surprising fact that one can ignore spatial structure in the massive output is explained. All results will be illustrated with a computer model of volcanic pyroclastic flow, the goal being the prediction of hazard probabilities near active volcanoes.

Maximum likelihood estimation from coarse data: what do we maximise?

Inés Couso

The term ”coarse data” encompasses different types of incomplete data where the (partial) information about the outcomes of a random experiment can be expressed in terms of subsets of the sample space. Maximum likelihood estimation (MLE) is a point-valued estimation procedure widely used in different areas of Statistics and Data Analysis. It searches for the parameter/vector of parameters that maximises the probability of occurrence of the dataset, and is assymptotically optimal under some conditions. We present and compare several extensions of this procedure to the case of coarse data, independently proposed by different authors during the last decades. We highlight the importance of modelling the so-called “coarsening process” that transforms the true outcome of the experiment into an incomplete observation and show that it can be modelled by means of a family of conditional probability distributions. We show some specific areas of statistics where such a process is very often overlooked, and provide examples illustrating how ignoring this coarsening process may produce misleading estimations. We discuss the conditions under which it can be safely ignored.

Transfer learning and learning with concept drift

Barbara Hammer

One of the main assumptions of classical machine learning is that data are generated by a stationary concept. This, however, is violated in practical applications e.g. in the context of life long learning, for the task of system personalisation, or whenever sensor degradation or non-stationary environments cause a fundamental change of the observed signals. Within the talk, we will give an overview about recent developments in the field of learning with concept drift, and we will address two particular challenges in more detail: (1) How to cope with a fundamental change of the data representation which is caused e.g. by a misplacement or exchange of sensors? (2) How to deal with drifting concepts which change either rapidly or smoothly over time, e.g. caused by a non-stationary environment? We will present novel intuitive distance-based classification approaches which can tackle such settings by means of suitable metric learning and brain-inspired adaptive memory concepts, respectively, and we will demonstrate their performance in different application domains ranging from computer vision to the control of protheses.

Analyzing molecular tumor profiles for precision oncology

Nico Beerenwinkel

Molecular profiling of tumor biopsies plays an increasingly important role not only in cancer research, but also in the clinical management of cancer patients. Multi-omics approaches hold the promise of improving diagnostics, prognostics, and personalized treatment. To deliver on this promise of precision oncology, appropriate bioinformatics methods for managing, integrating and analyzing large and complex data are necessary. I will discuss some of these computational challenges in the context of the molecular tumor board and the specific bioinformatics support that it requires, from the primary analysis of raw molecular profiling data to the automatic generation of a clinical report and its delivery to decision-making clinical oncologists.

2 Sessions

Advances in Recursive Partitioning and Related Methods

Visual Pruning for Informative Prediction Trees

Roberta Siciliano, Antonio D'Ambrosio, Carmela Iorio, Giuseppe Pandolfo

In the framework of supervised statistical learning, this paper provides a one-step procedure of pruning and decision tree selection. Main issue is an inedited tree graph where the edge lengths linking two nodes are geometrically identified such to be informative of the recursive tree data partitioning to be used for selection of the prediction tree. Our method is a valid alternative to CART cost-complexity pruning and selection reducing enormously the computational cost. This is particularly important in all cases where tree interpretability is to be taken seriously. In addition, this approach can be fruitfully used when dealing with non-standard data sets such as preference rankings, symbolic data, web data, circular data, etc. For all these examples, accurate decision tree-based rules using ensemble methods such as boosting, bagging or random forest are prohibitive.

Amodio, S., D'Ambrosio, A., & Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. European Journal of Operational Research, 249(2), 667–676.

Aria, M., D'Ambrosio, A., Iorio, C., Siciliano, R., and Cozza, V. (2018) Dynamic recursive tree-based partitioning for malignant melanoma identification in skin lesion dermoscopic images, Statistical Papers, 1-17.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth International Group.

D'Ambrosio, A., Amodio, S., & Iorio, C. (2015). Two algorithms for finding optimal solutions of the Kemeny rank aggregation problem for full rankings. Electronic Journal of Applied Statistical Analysis, 8(2), 198–213.

D'Ambrosio, A., Heiser, W. J. (2016). A Recursive Partitioning Method for the Prediction of Preference Rankings Based Upon Kemeny Distances, Psychometrika, vol. 81, no. 3, 774–794.

Iorio, C., Aria, M., D'Ambrosio, A., Siciliano, R. (2018). Informative Trees by Visual Pruning, Submitted.

Johansson, U., Boström, H., Löfström, T. (2018). Interpretable regression trees using conformal prediction, Expert Systems With Applications, 97, 394-404.

Siciliano, R., & Mola, F. (2000). Multivariate data analysis and modelling through classification and regression trees. Computational Statistics and Data Analysis, 32, 285–301.

Siciliano, R., D'Ambrosio, A., Aria, M., Amodio, S. (2017). Analysis of Web Visit Histories, Part II: Predicting Navigation by Nested STUMP Regression Trees, Journal of Classification, Volume 34, Issue 3, pp 473–493.

Semisupervised Clustering through Recursive Partitioning and Complex Networks

Claudio Conversano, Giulia Contu, Luca Frigau, Francesco Mola

Recently, in literature new clustering methods that take advantages of complex networks algorithms have been developed. The use of complex networks approaches allows obtaining higher accuracy than traditional clustering methods. These new kinds of approaches consist in defining the distances between objects by classical metrics and performing complex network algorithms to partition the data, considering the distances as links for the derivation of the network.

We propose a new approach in which the distances are defined by machine learning methods. Particularly, we consider tree-based methods such as Classification and Regression Tree (CART) and Random Forest (RF). This approach results in a semi-supervised clustering method in which the communities are identified using community detection algorithms (traditional methods, divisive methods, modularity methods and other methods).

We illustrate the advantages of our approach through a simulation experiment and some real data examples. One of them concerns a the performance metrics and the content characteristics of 137 websites of UNESCO sites located in Italy, France and Spain.

Boosted Decision Trees for Behaviour Mining of Concurrent Programs Combinated with Genetic Algorithms

Hana Pluháčková, Tomáš Vojnar, Bohuslav Křena

Testing of concurrent programs is difficult since the scheduling non-determinism requires one to test a huge number of different thread interleavings. Moreover, repeated test executions that are performed in the same environment will typically examine similar interleavings only. One possible way how to deal with this problem is the noise injection technique which helps to increase the number of thread interleavings examined during repeated test executions. However, for noise injection to be efficient, one has to choose suitable noise injection heuristics from among the many existing ones as well as to suitably choose values of their various parameters. This problem is sometimes denoted as the test and noise configuration search problem (TNCS problem), and it is not easy to solve. There exist genetic algorithms which focus on finding test and noise settings that allow one to cover a high number of distinct interleavings (especially those which are rare) and provide stable (i.e. repeatable) test results at the same time. However, genetic algorithms are often quite time consuming. Our work introduces a method of solving the TNCS problem based on a data mining approach, aiming at a higher efficiency than the genetic algorithms. The approach is, in particular, based on using boosted decision trees in the context of noise-based testing to identify which test and noise parameters are the most influential for a given program and a given testing goal and which values (or ranges of values) of these parameters are suitable for meeting the given goal. Results from this process can also be combined with the genetic solution of the TNCS problem by providing an initial solution for the genetic approach, saving some time compared with starting the genetic algorithm with random test and noise configurations.

A Framework for Measuring Stability of Recursive Partitioning Methods

Michel Phillip, Thomas Rusch, Carolin Strobl, Kurt Hornik

Stability is a major requirement to draw reliable conclusions when interpreting results from recursive partitioning methods. In this paper, we present a general framework for assessing and comparing the stability of results, that can be used in applications as well as in simulation and benchmark studies. Typical uses of the framework in would be to compare the stability of results generated by different candidate algorithms for a data set at hand or to assess the stability of existing or newly developed algorithms in benchmark studies. We further use the framework to demonstrate that stability is a property of both the algorithm and the data-generating process. In particular, we demonstrate that algorithms like recursive partitioning which are generally considered unstable can produce stable results when the functional form of the relationship between the predictors and the response matches the algorithm. Code to perform the stability analyses is provided in the form of an R package.

Distributional Regression Forests for Probabilistic Modeling and Forecasting

Lisa Schlosser, Torsten Hothorn, Heidi Seibold, Achim Zeileis

To obtain a probabilistic model for a dependent variable based on some set of explanatory variables, a distributional approach is often adopted where the parameter(s) of the distribution are linked to regressors. In many classical models this only captures the location/expectation of the distribution but over the last decade there has been increasing interest in distributional regression approaches modeling all parameters including location, scale, and shape. Notably, the GAMLSS framework allows to establish generalized additive models using this approach. However, in situations where variable selection is required and/or there are non-smooth dependencies or interactions (especially unknown or of high-order), it is challenging to establish a good GAMLSS. A more natural alternative would be the application of regression trees or random forests but, so far, no general distributional framework is available for these methods. The R package ”disttree” is based on ”partykit” and provides regression trees and forests for classical continuous distributions from the GAMLSS framework as well as their censored or truncated counterparts. Furthermore, package ”trtf” takes the idea one step further and implements trees and forests based on a novel and very flexible parametric family of distributions characterized by their transformation function.

Treatment Subgroup Interactions and Personalized Treatment Effects

Heidi Seibold, Achim Zeileis, Torsten Hothorn

Established statistical methods for the analysis of randomised experiments with two or more treatment groups estimate a universal (i.e. constant) treatment effect that applies to all subjects observed and - even more importantly - to all future subjects. Common use cases are: clinical trials where patients are assigned to standard of care vs. new treatment; behavioral experiments in psychology or economics where individual responses are assessed under different experimental settings; or A/B testing where users see different versions of a website.

We propose model-based trees and model-based random forests as a way to relax the assumption that all subjects have the same treatment effect and estimate stratified or personalised treatment effects that depend on characteristics of the subject, e.g., the biomarkers of a patient; personal traits of inviduals; history of a website user.

The R package ”model4you” (based on infrastructure from the ”partykit” package) provides a simple user interface that allows the user to define the model estimating the overall treatment effect and then partition this model. The model is easy to define, since it is the basic treatment model, which - in the case of a clinical trial - is defined the study protocol. The stratified and personalised models can be visualised and are easy to interpret.

Algorithm Selection/Configuration and Machine Learning

Multi-Objective Selection of Algorithm Portfolios over Multiple Data Sets

Daniel Horn, Rosa Pink

Many (machine learning) algorithms depend on multiple parameters and thus require problem-specific hyperparameter tuning. In some situations a single measure may not be enough to estimate the performance of an algorithm and multiple measures should be considered. For example in the large data situation, not only accuracy but also runtime of algorithms is important. In such multi-objective tuning situations, different hyperparameter settings will lead to different trade-offs between conflicting objectives. For instance, the number of trees in a random forest increases its runtime, but is also very likely to increase its accuracy. The set of all optimal trade-offs has to be computed in the tuning process and is called the Pareto front.

Selecting the best algorithm in a single objective setting is easy: We can simply choose algorithms that optimize the single performance measure. In the multi-objective case, however, we have to compare Pareto fronts between conflicting performance measures in order to identify best algorithms. It is very unlikely that a single algorithm is best for all possible trade-offs between the objectives. Hence, a portfolio must be selected that provides optimal algorithms for all trade-offs. In recent years we published a method to select such portfolios.

Back then, the method was only able to select portfolios based on algorithm performances on a single data set. However, in most practical applications we are interested in the global performance of algorithms, considering multiple data sets. Therefore, we extend our former work. Our new aim is to select a portfolio that works well for all considered data sets. Naturally, there may be data sets that are too heterogeneous to find a common portfolio. These situations should also be detected.

In this work we introduce our extended version and validate it using artificial data sets. Moreover, we present a practical application from the field of machine learning: We compare different approximative SVM solvers for large data sets by their accuracy and runtime.

Predicting Rankings of Classification Algorithms in AutoML

Helena Graf, Marcel Wever, Felix Mohr, Eyke Hüllermeier

In recent years the demand for machine learning functionality increased quite dramatically, leading to the (partial) automation of engineering machine learning applications (AutoML). Current AutoML approaches select an optimal sequence of machine learning algorithms (ML pipeline), while simultaneously optimizing the parameters of the respective algorithms. The selection process of these tools heavily relies on executing candidate pipelines and evaluate them, e.g., by means of cross-validation. Moreover, these approaches usually propose a single best solution to the user. Sometimes also a set of solutions can be returned but in this case there is no relation among the returned pipelines.

In this work, we consider the AutoML problem from a meta learning perspective. Instead of actually executing candidate pipelines, we use a machine learning model to predict a ranking of classifiers based on properties of that data (aka. meta-features). To this end, we consider two approaches of implementing these models. (A) For each classifier, we fit a regression model to predict its performance (e.g., predictive accuracy) for a dataset described by the meta-features. To train the regression models, we use the previously observed performance for other datasets. Subsequently, the predictions are aggregated to a ranking. (B) Instead of exact performance values, a preference model is fitted directly where a training example consists of the meta-features of a dataset and the label is a ranking of the classifiers.

We compare the two alternatives to each other in an experimental study and show that the regression-based approach with RandomForests proves beneficial. Furthermore, we show that some of the approaches outperform a naïve ranking based on the frequency of best performances.

Challenges of Meta-Learning on a Distributed Machine Learning Platform

Christian Geißler

Applying state-of-the-art algorithm selection (AS) and hyperparameter optimization (HPO) for large classification tasks is still a challenge. Although service platforms for machine learning offer ecosystems around those tasks, their goals differ from the users intention regarding the computational efficiency of the offered methods (since they sell computational power), interoperability and flexibility (vendor lock-in regarding storage and analysis service solutions). In other use cases, the data is too sensitive or there are other reasons like bandwidth limits or limits on the response time that require on-site distributed machine learning, capable of processing large amounts of data.

One of the well-known frameworks to realize such a system is the apache spark framework with its machine learning library Spark ML. It offers distributed basic machine learning implementations, but lacks many of the meta methods that are available in otherwise commonly used frameworks like weka or scikit learn and especially advanced HPO and AS methods.

The BMBF-funded coda research project addresses these issues by researching, if, which and how advanced methods for hyperparameter optimization and algorithm selection can be implemented for a distributed processing platform like apache spark. Within coda, a gpu-powered distributed data analytics system is being build in order to investigate new algorithmic solutions as well as their application on co-located research projects. This talk will explain the overall project goals, setup, challenges and approaches.

ML-Plan: Automated Machine Learning for Multi-Class and Multi-Label Classification

Felix Mohr, Marcel Wever, Eyke Hüllermeier

Automated machine learning (AutoML) seeks to automatically select, compose, and parametrize machine learning algorithms, so as to achieve optimal performance on a given task (data set). In this presentation, we demonstrate how the task of automated machine learning for multi-class and multi-label classification can be approached using an AI technique referred to as hierarchical planning.

Planning with hierarchical task networks (HTN) is a very natural approach to solve the AutoML problem, because the basis of the HTN problem description is the grammar-like structure that describes how machine learning pipelines can be constructed recursively. Hierarchical planning is a technique from the field of AI planning that can be used to describe a planning problem with background knowledge that describes the possibilities of how to achieve a certain goal. In contrast to classical planning, goals are described in form of so called task networks. Each task either corresponds to a basic operation that can be directly executed, or it is complex, which means that it needs to be decomposed into simpler sub-tasks. The planning process is to recursively decompose the tasks of the initially given task network until all remaining tasks correspond to existing operations. For AutoML, the initial task network is simply the task to create a machine learning pipeline. This goal can be achieved by first setting up a preprocessor an then a classifier or to just setup a classifier and ignore preprocessing. The preprocessor and classifier in turn need first to be chosen and then to be parametrized. Since classifiers can be meta-classifiers like AdaBoost or ensembles, which are configured with other classifiers, the configuration of the machine learning pipeline may recurse. In the HTN planning formalism, this is naturally covered by the task networks of the methods that can be used to decompose a task. For example, there is a method to decompose the classifier selection task using AdaBoost, which induces a new sub task not only to choose the parameters of the algorithm but also to configure a further classifier, the base classifier used for AdaBoost.

In contrast to other existing approaches, which are mainly built on local search techniques, ML-Plan conducts a global best-first search. Solving an HTN problem usually means to conduct a graph search, i.e., so find a path between the root node and a goal node of a graph induced by the HTN problem. ML-Plan conducts such a search. Similar to Monte Carlo tree search, it estimates for the value of nodes by randomly completing paths to goal nodes and evaluating the randomized completions, which correspond to executable pipelines, on the given data.

We have shown that ML-Plan is competitive and often superior to all other state-of-the-art tools for Auto-ML. We evaluated ML-Plan against Auto-WEKA, auto-sklearn, and TPOT. For a timeout of 1h, we see that ML-Plan is almost never beaten by any of those tools but often achieves better performance. For a timeout of 1d, the other tools manage to catch up, but ML-Plan is still competitive.

Applications 1

HMM with Non-Emitting States for Map Matching

Wannes Meert, Mathias Verbeke

When collecting GPS traces, the GPS points are often inaccurate and require cleaning before further analysis can be performed. Therefore, a common preprocessing step is map matching, i.e. mapping the GPS points to a sequence of actual roads. Simply mapping the points to the nearest street, however, will return inaccurate results.

A popular approach is to use a Hidden Markov Model and Viterbi decoding, with the actual road segments as hidden states and the GPS points as observations. This approach was popularized by Microsoft (Newson and Krumm, 2009) and implemented in various services such as Mapbox's Map Matching API for OpenStreetMap.

A typical problem for HMMs is that for every state transition (e.g. moving from one street to another) an observation is 'absorbed'. In settings where measurements are far apart, thus less frequent than transitioning segments, this model fails. Two common proposed solutions are: interpolation, and interweaving Viterbi with a search algorithm. Neither is ideal. The first typically requires a high sample rate with more GPS points than segments and will generate too many observations. In the second, the HMM model is not used for all computations which makes it impossible to introduce a more sophisticated transition function.

We propose a new map matching approach using HMMs with non-emitting states, inspired by Profile HMMs in bioinformatics. These are states that are not associated with an observation. They thus allow for dynamic interpolation based on the route. This model, however, loses the property that the Viterbi algorithm is executed on a lattice bounded by the number of observations. To cope with this, we propose a smart pruning strategy such that only non-emitting states with the highest relevance are visited.

We evaluate on two applications, tracking vehicles and runners, and show how we can now learn the transition probabilities from history.

P. Newson and J. Krumm (2009). Hidden markov map matching through noise and sparseness. In Proc. of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 336–343

Automating Time Series Feature Engineering for Activity Recognition from Synchronized Inertial Measurement Units

Andreas W. Kempa-Liehr, Jonty Oram, Thor Bezier

The ubiquitous availability of wearable sensors is responsible for driving the Internet-of-Things, but is also making an impact on sport sciences and precision medicine. While human activity recognition from smartphone data or other types of inertial measurement units (IMU) has evolved to one of the most prominent daily life examples of machine learning, the underlying process of time series feature engineering still seems to be a time-consuming process. This inhibits the development of IMU-based machine learning applications in sport science and precision medicine. This contribution discusses the automation of time series feature engineering on the basis of the FRESH algorithm (FeatuRe Extraction based on Scalable Hypothesis tests) to identify statistically significant features from synchronized IMU sensors (IMeasureU Ltd, NZ). By identifying time series characteristics in an early stage of the data science process, our approach closes feedback loops with domain experts and fosters the development of domain specific features. The automated time series feature engineering process for human activity recognition will be discussed on the basis of the Python package tsfresh, which implements the application programming interface of standard machine learning libraries like scikit-learn and has been adapted by more than 2600 data scientists since its publication in October 2016.

The Moderating and Mediating Role of Meaning of Work – a PLS Path Analysis

Joachim Schwarz, Heiko Weckmüller

Empirical research on meaning of work forms an important part in the field of human resource management. There exist several studies which analyse the relationship between meaning of work and a variety of different causes and effects (see e.g. Dick and Stegman, 2015; Höge and Schnell, 2012). Based on the Job Characteristics Model (Hackman and Oldham, 1976), meaning of work is considered as mediator between job characteristics and a variety of different outcomes (see e.g. May et al., 2004; Albrecht and Su, 2012; Schnell et al., 2013). Other studies use meaning at work as moderator (see e.g. Harris et al., 2007). Despite the numerous studies, there still remain several unsolved issues. So it is quite common to test mediating and moderating effects separately and not in a common model. Further, the intention to quit is only seldomly used as target variable. And finally, the collected survey data sets are usually small and not representative.

Based on a representative survey from Germany, the DGB Index Good Work 2014 (DGB: Deutscher Gewerkschaftsbund – German Trade Union Confederation), we examine the moderating and mediating role of meaning of work for the relationship between several social, personal, environmental and motivational job characteristics and the intention to quit as a manifest binary target variable. The DGB Index Good Work 2014 contains about 4,000 responses.

First, due to the novelty of the data within the scientific environment, we evaluated the data quality. Then, coping with the complexity of the model, consisting of more than 70 latent variables, all formatively measured, many of them one indicator constructs, we fitted a PLS path model which can both handle formatively measured constructs and a binary target variable.

S. L. Albrecht, M. J. Su (2012). Job resources and employee engagement in a Chinese context: the mediating role of job meaningfulness, felt obligation and positive mood. International Journal of Business and Emerging Markets, 4(4): 277-292.

R. v. Dick, S. Stegmann (2015). Sinnvolle Arbeit ist identitätsstiftend – Zur Bedeutung der sozialen Identifikation als Wirkmechanismus zwischen Bedeutsamkeit der Aufgabe und Arbeitseinstellungen. Arbeit, 24(1-2): 49-66.

J. R. Hackman, G. R. Oldham (1976). Motivation through the design of work: Test of a theory. Organizational Behavior and Human Performance, 16(2): 250-279.

K. J. Harris, K. M. Kacmar, S. Zivnuska (2007). An investigation of abusive supervision as a predictor of performance and the meaning of work as a moderator of the relationship. The Leadership Quarterly, 18(3): 252-263.

T. Höge, T. Schnell (2012). Kein Arbeitsengagement ohne Sinnerfüllung. Eine Studie zum Zusammenhang von Work Engagement, Sinnerfüllung und Tätigkeitsmerkmalen. Wirtschaftspsychologie, 1: 91-99.

D. R. May, R. L. Gilson, L. M. Harter (2004). The psychological conditions of meaningfulness, safety and availability and the engagement of the human spirit at work. Journal of Occupational and Organizational Psychology, 77(1): 11-37.

T. Schnell, T. Höge, E. Pollet (2013). Predicting meaning in work: Theory, data, implications. The Journal of Positive Psychology, 8(6): 543-554.

The Effect of Ambient Light Conditions on Road Safety

Valentin Schiele, Christian Bünnings

Still more than 25,000 people die on European roads every year, a disproportionately large number of them during night. While certain causes of road accidents like drunk driving or speeding are well studied and addressed by policy, only little is known about the role of ambient light conditions for road safety. To estimate effects of darkness on road accident counts, we rely on a large administrative data set from Great Britain covering more than 3 million accidents. Our identification strategy exploits three sources of variation in darkness for a given hour of the day: i. within region variation that is due to variation in sunrise and sunset time throughout the year; ii. between region variation that is due to earlier sunrise / sunset in eastern compared to western regions; and iii. between region variation that is due to longer (shorter) daytime during summer (winter) in northern compared to southern regions. Our main results suggest that darkness increases the number of road accidents by around 7%. We find especially large effects for the number of accidents involving pedestrians and older people, indicating that the effect operates through an increased risk of poor night vision among older drivers.

Applications 2

Sociohistorical Recommendations for the Dewey Decimal Classification Editorial Policy Committee for the Reclassification of Pentecostalism

Adam Stewart

Pentecostalism is very likely the fastest-growing form of Christianity in the world today (Anderson, 2013, 1–10). Pentecostalism—especially in Africa, Asia, and South America—often exerts a major influence on a number of important economic, social, and political issues at both the local and national levels (Freston, 2001). The religion's both tremendous growth and significant influence on society, have caused Pentecostalism to become the focus of intense, global academic study (Anderson, et al., 2010).

Despite the numerical size of Pentecostalism and the vast literature that Pentecostalism has provoked, accessing the full range of materials published on the subject of Pentecostalism can be difficult. This challenge is largely explained by the fact that Pentecostalism is frequently classified using an Americentric and ahistorical understanding of the religion, which obscures or marginalizes works that do not neatly fit into the predominant understanding of Pentecostalism currently in operation within the United States of America (Stewart, 2010, 17–37; 2012, 43–48; 2014, 151–72; 2015, 23–30; 2017, 243–50).

In this paper, I do three things. First, I summarize the evolution of suggestions given to cataloguers for classifying Pentecostalism in the Dewey Decimal Classification System since the time the religion is first mentioned in the fifteenth edition published in 1951 until the most recent twenty-third edition published in 2011. Second, I describe the various definitional and historical errors regarding the description of Pentecostalism, which, I argue, derive from an insular, American understanding of the religion. Finally, I propose some recommendations for the reclassification of Pentecostalism, which would both correct existing outmoded understandings of the religion as well as improve access to works published on Pentecostalism, especially by authors originating from outside of the United States of America.

Anderson, A. H. 2013. To the ends of the earth: Pentecostalism and the transformation of world Christianity. New York: Oxford University Press.

Anderson, A., M. Bergunder, A. Droogers, and C. van der Laan. eds. 2010. Studying global pentecostalism: Theory and methods. Berkeley: University of California Press.

Freston, P. 2001. Evangelicals and politics in Africa, Asia, and Latin America. New York: Cambridge University Press.

Stewart, A. 2010. "A Canadian Azusa? The implications of the Hebden mission for pentecostal historiography." In Winds from the north: Canadian contributions to the pentecostal movement, edited by Michael Wilkinson and Peter Althouse, 17–37. Leiden: Brill Academic Publishers.

———. 2012. "Azusa Street mission and revival." In Handbook of pentecostal Christianity, edited by Adam Stewart, 43–48. DeKalb: Northern Illinois University Press.

———. 2014. "From monogenesis to polygenesis in pentecostal origins: An examination of the evidence from the Azusa Street, Hebden, and Mukti missions." PentecoStudies: An Interdisciplinary Journal for Research on the Pentecostal and Charismatic Movements 13(2):151–72.

———. 2015. The new Canadian pentecostals. Waterloo: Wilfrid Laurier University Press.

———. 2017. "A subject analysis of pentecostalism in the Dewey Decimal Classification system." Biblioteka 21:243–50.

Multidimensional Comparative Ranking of the European Union Countries in the Area of Sustainable Development

Marcin Pełka, Tomasz Bartłomowicz

The subject of the study is a comparative analysis of the level of sustainable development of the European Union countries. The basis of the study was the data collected in the Eurostat database, which the European Commission uses to monitor the implementation of the objectives of the EU Sustainable Development Strategy. In order to determine similarities and differences in the level of development of the European Union countries in the study area, the methods of multidimensional comparative analysis (linear ordering and classification methods) were used. The obtained results allowed to identify similar as well as different countries because of the designated indicators. It enabled the creation of a ranking of the EU countries based on indicators of Sustainable Development. In the calculations the environment and packages of R program were used.

Keywords: ranking of the European Union countries, multidimensional statistical analysis, Eurostat.

1. Bell S., Morse S. (2008), Sustainability Indicators: Measuring the immeasurable (2nd ed). London: Earthscan.

2. EUROSTAT: Sustainable Development Indicators, http://ec.europa.eu/eurostat/web/sdi, Accessed: June 2014.

3, Hair J.F., Anderson R.E., Tatham R.L., Black W.C. (1995), Multivariate Data Analysis with Readings, Englewood Cliffs, Prentice-Hall T.

4. Phillis Y.A., Grigoroudis E., Kouikoglou V.S. (2011), Sustainability Ranking and Improvement of Countries, Ecological Economics, no. 70.

Radiocarbon Dating of the Turin Shroud: New Evidence from Raw Data

Tristan Casabianca, Benedetto Torrisi, Giuseppe Pernagallo, Emanuela Marinelli

In 1988, three laboratories performed a radiocarbon analysis of the Shroud of Turin, a controversial linen cloth believed by some to be the burial cloth of Jesus of Nazareth. The results, which were centralized by the British Museum and published in Damon et al. (1989), provided “conclusive evidence” of the mediaeval origin of the artefact. However, the raw data were never released by the institutions. In 2017, in response to a legal request, all raw data kept by the British Museum were made accessible.

The aim of this paper is to analyse the raw data and to measure their convergence with the official radiocarbon dates published in Damon et al. (1989). The conclusions of parametric tests, non-parametric tests, and OxCal (a software commonly used by radiocarbonists), strongly suggest that homogeneity is lacking in the data and that the procedure should be reconsidered.

J. Christen, and S. Pérez (2009). A New Robust Statistical Model for Radiocarbon Data. Radiocarbon, 51: 1047-1059.

P. E. Damon, D. J. Donahue, B. H. Gore, A. L. Hatheway, A. J. T. Jull et al. (1989). Radiocarbon Dating of the Shroud of Turin. Nature, 337: 611-615.

C. B. Ramsey, and S. Lee (2013). Recent and Planned Developments in the Program OxCal. Radiocarbon, 55: 720-730.

M. Riani, A. C. Atkinson, G. Fanti, F. Crosilla (2013). Regression Analysis with Partially Labelled Regressors: Carbon Dating of the Shroud of Turin. Statistics and Computing, 23: 551-561.

E. M. Scott, G. T. Cook and P. Naysmith (2007). Error and Uncertainty in Radiocarbon Measurements. Radiocarbon, 49: 427-440.

R. Van Haelst (2002). Radiocarbon Dating of The Shroud of Turin: a critical review of the Nature report (authored by Damon et al.) with a complete unbiased statistical analysis. (Available from: http://www.sindone.info/VHAELST6.PDF)

G. K. Ward, S. R. Wilson (1978). Procedures for Comparing and Combining Radiocarbon Age Determinations: A Critique. Archaeometry, 20(1): 19-31.

Big Data and Complex Network Analytics

A New Approach to Measuring Distances in Dense Graphs

Fatimah Almulhim, Peter Thwaites, Charles Taylor

The problem of computing distances and shortest paths between vertices in graphs is one of the fundamental issues in graph theory. It is of great importance in many different applications, for example, transportation, and social network analysis. However, efficient shortest distance algorithms are still desired in many disciplines. Basically, the majority of dense graphs have ties between the shortest distances. Therefore, we consider a different approach and introduce a new measure to solve all-pairs shortest paths for undirected and unweighted graphs. This measures the shortest distance between any two vertices by considering the length and the number of all possible paths between them. The main aim of this new approach is to break the ties between equal shortest paths SP, which can be obtained by the Breadth-first search algorithm (BFS), and distinguish meaningfully between these equal distances. Moreover, using the new measure in clustering produces higher quality results compared with SP. In our study, we apply two different clustering techniques: hierarchical clustering and K-means clustering, with four different graph models, and for a various number of clusters. We compare the results using a modularity function to check the quality of our clustering results

The Impact of Graph Symmetry on Clustering

Fabian Ball, Andreas Geyer-Schulz

We have empirically shown that real-world graphs contain symmetries with a probability larger than 70% [1]. MacArthur et al. [2] have also investigated symmetry in real-world graphs and identified most of the found symmetries to be small appendices to the graph rather than symmetries which affect large parts of the whole graph.

This contribution analyzes if existing automorphisms (self-mappings of the node set of the graph that describe the symmetry) affect the modularity optimal graph clustering solution by exchanging nodes between clusters. To achieve this, we connect the empirical findings of MacArthur et al. and theoretical implications of the modularity clustering resolution limit of Fortunato and Barthélemy [3] with our own analysis to show that (i) existing symmetry normally affects only small parts of the graph and, therefore, (ii) the clustering solution is unlikely to be affected by the symmetries but still with reasonable probabilty.

[1] F. Ball and A. Geyer-Schulz, “How Symmetric Are Real-World Graphs? A Large-Scale Study,” Symmetry, vol. 10, no. 1, p. 17, Jan. 2018.

[2] B. D. MacArthur, R. J. Sánchez-García, and J. W. Anderson, “Symmetry in complex networks,” Discrete Applied Mathematics, vol. 156, no. 18, pp. 3525–3531, Nov. 2008.

[3] S. Fortunato and M. Barthélemy, “Resolution limit in community detection,” Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 1, pp. 36 – 41, 2007.

Scalable Knowledge Graph Exploration for Sentiment Classification

Gezim Sejdiu, Ali Denno, Mohamad Denno, Hajira Jabeen, Jens Lehmann

In the recent years, the automation of a range of processes, advancement of smart services, and internet-based access has lead to information explosion over the web. As a result, data and information can be searched from numerous sources (e.g knowledge collections, databases, articles, blogs, social media, etc). Search engines and question answering systems are capable of locating such information and answering a wide range of different queries starting from daily life questions and ending with complex questions that might relate to sensitive topics such as military, armed forces, etc. Such systems and the services they offer might be used in a suspicious or non-ethical way to extract sensitive information. As an example, users can possibly ask questions that are related to terrorist or destructive activities. The sensitive and protected information can also be modified and altered without the creators or the public knowing about it, especially that we now have open source search engines and systems that can provide a variety of information. Furthermore, users of these systems might use innocent looking questions to extract sensitive information about some organizations. The organizations providing these services do not usually wish to answer such questions or give out such sensitive information, and thus our system can help to apply measures to prevent this situation.

In this paper, we introduce ARCANA, a novel approach that is capable of restricting the provided information by detecting the nature of a request. ARCANA uses natural language processing and machine learning techniques to extract useful information from the Linked Data by extracting a set of categories and their relations. These set of words are then used to predict the response as malicious with respect to the given query. We demonstrate empirically that our approach is capable to answer effectively 85% of the test questions correctly and is always able to improve by enriching further the database.

Comparing Partitions of the Petersen Graph

Andreas Geyer-Schulz, Fabian Ball

The Petersen graph has been of long term interest to many graph theorists because of its appearance as counterexample in many places. In this contribution we use the Petersen Graph – because of its transitivity and its large automorphism group – as a show piece for invariant partition comparison measures. We show that we can decompose distances between partitions of the Petersen graph in an (invariant) structural part and a (variable) part caused by an automorphism. In addition, we study the effects caused by subgroups of the automorphism group and their interpretation.

Graph-Theoretic Network Analysis of the Interactions Between Patients, Physicians and Prescribed Drugs

Reinhard Schuster, Timo Emcke

A Morbidity Related Group (MRG) is the drug group at the third level of the international Anatomic Therapeutic Chemical Classification System (ATC) with the highest costs in a quarter with respect to a certain physician-patient combination in the outpatient sector. This drug group strongly relates to the morbidity of the patient leading to the term MRG and a unique patient classification. Thereby we get fractions of patients in 256 drug groups in form of a prescription vector for each physician. Each physician belongs to a specialist group primarily determined by admission law. If we calculate Manhattan distances of physicians and look for n clusters with a value n roughly representing the number of specialist groups, we get poor results compared with admission groups which are unstable with respect to different methods and quarters.

In 2017 the outpatient prescription data of the northern German state of Schleswig-Holstein include around 8.6 million patients and 2,800 physicians. Each patient is counted per quarter and physician. If physicians cooperate in one office, they are counted for each specialist group only once. These physicians are considered as vertices of a graph. For two physicians we consider the Manhattan distance of their prescription vector. Vertex x is connected to Vertex y by a directed edge if y belongs to the k smallest Manhattan distances regarding x. From an application point of view the cubic case k=3 is of special interest. We get a related undirected graph. With respect to each pair of physicians the combination of all drug groups (3rd level of ATC) with a positive vector component for at least one physician, has more than 10 billion elements. Hence effective algorithms are needed.

For k=3 we get three connected graph components. The smallest component has 6 vertices belonging to physicians (within three admission groups) that focus on drug substitution therapy. The medium size component with 106 vertices belongs to ophthalmologists. It has a graph diameter of 12 and a Hamiltonian cycle. The large component has 2,676 vertices, 6,848 edges and a graph diameter of 38. Otolaryngologists and pediatrician are well separated in contrast to a cluster analysis with respect to the used Manhattan metric.

Internists, nephrologists, gastroenterologists and some minor groups are connected to GPs by almost linear paths. The same is true for oncology and specialists like gynecologists.

Physicians are typical, if all adjacent edges are connected to vertices of the same specialization. There is a subgraph near typical physicians, which has a local minimum with respect to the Cheeger isoperimetric subgraph value. The global determination of the minimum of the isoperimetric problem is NP-complete in general.

The resulting graph allows for stable refinements and some classification changes with respect to the administrative classification. The MRG system can be applied to economic considerations leading to results which are sensitive with respect to the specialization of physicians. Thereby the graph theory can help to improve drug economic evaluations.

Traffic Flow Analysis Using 12 Years of Data

Ian Marsh

Predicting traffic congestion is an important tool for users, authorities, fleet management companies, and road planners. Typically road traffic authorities know the long-term demands on road sections, however, the shorter-term prediction is a research problem. That said, algorithms, data, and processing have advanced to a point where new road analysis can be done, which is very much the topic of this contribution.

This paper applies big data practices on 12 years of data from the Swedish road traffic authority to predict traffic flow congestion around the city of Stockholm. We present a simple transparent model to detect traffic build up in space and time, and hence find the end of any queue.

An important insight in this work is the use of traffic density as a new measure for congestion detection, rather than the average velocity or flow values. We present processing techniques and execution times as well as video visualizations of the queue buildups around Stockholm.

Bioinformatics and Biostatistics

Ensemble Feature Selection for Regression Problems

Ursula Neumann, Dominik Heider

Feature selection (FS) can be used to detect and select a relevant subset from a set features for the construction of a machine learning or statistical models. In this process, as much of irrelevant and redundant information as possible should be identified and removed from the data. Thus, FS methods cannot only be used to identify relevant features, but also to reduce noise. Based on the assumption that combining several weak FS algorithms obtains more reliable results than individual FS approaches, ensemble feature selection gained a high level of attention in the recent years (Neumann et al., 2016). This is due to the fact that single FS methods are prone to be biased depending on the data quality and distribution. Datasets are often imbalanced and heterogeneous. Thus, different methods have different biases and benefits according to the type of features, the degree of imbalance, and the size of the dataset. In a former study, we implemented an ensemble feature selection (EFS) for binary classification. In the current study, we introduce the extension to regression problems. Classification is the task of predicting a discrete class label, whereas regression is the problem of predicting a continuous quantity output. Some FS algorithms can be used for both classification and regression with small adaptions, such as the random forests importance estimations, however, others are only applicable for feature selection in classification problems. In the current study, we implemented and evaluated different FS methods for quantitative feature ranking in regression problems. Moreover, the best-performing methods were then combined into a regression EFS approach, implemented into the R package EFS (Neumann et al., 2017), and evaluated using several real-world datasets.

U. Neumann , M. Riemenschneider, J.-P. Sowa,T. Baars, J. Kälsch, A. Canbay, D. Heider (2016). Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. BioData Mining, 9(1):36.

U. Neumann, N. Genze, D. Heider (2017). EFS: An Ensemble Feature Selection Tool implemented as R-package and Web-Application. BioData Mining, 10:21.

Group-Wise Feature Selection with Stacked Domain Learning

Wouter van Loon, Marjolein Fokkema, Botond Szabo, Mark de Rooij

More and more often, researchers are confronted with multi-domain data, i.e., data for which the total feature set can be divided into meaningful subsets. In health research, for example, multi-domain data can occur when data are collected from multiple sources (e.g. medical imaging, genomics), or when different feature sets are computed from a single source (e.g. different MRI modalities). Combining data from multiple domains can potentially lead to more accurate disease status classification. However, focusing solely on classification accuracy does not contribute to a better understanding of the domain structure. Of interest is which domains (sources or feature sets) are important for accurate prediction of medical outcomes. We present an ensemble learning-based framework for selecting important domains called Stacked Domain Learning (SDL). The ensemble consists of domain-specific models of which the predictions are combined using a meta-learner. SDL is a highly flexible method which can be tuned to fit the research question at hand. Commonly used ensemble methods typically sacrifice interpretability to obtain more stable results. To ensure interpretability of the final stacked model, we consider a special case of SDL where all base- and meta-learners are (possibly regularized) logistic regression functions. To perform domain selection, we put L1 and non-negativity constraints on the parameters of the meta-learner. In a simulation study we show that non-negativity constraints are crucial in maintaining a low false positive rate, and provide some theory motivating this behavior. We additionally show that compared to competing non-ensemble methods such as the Group Lasso, our method is more accurate and generally faster.

Learning the Topology of Latent Signaling Networks from High Dimensional Transcriptional Intervention Effects

Zahra Sadat Hajseyed Nasrollah, Achim Tresch, Holger Fröhlich

Data based learning of the topology of molecular networks, e.g. via Dynamic Bayesian Networks (DBNs) has a long tradition in Bioinformatics. The majority of methods take gene expression as a proxy for protein expression in that context, which is principally problematic. Further, most methods rely on observational data, which complicates the aim of causal network reconstruction. Nested Effects Models (NEMs – Markowetz et al., 2005) have been proposed to overcome some of these issues by distinguishing between a latent (i.e. unobservable) signaling network structure and observable transcriptional downstream effects to model targeted interventions of the network.

The goal of this project is to develop a more principled and flexible approach for learning the topology of a dynamical system that is only observable through transcriptional responses to combinatorial perturbations applied to the system. More specifically, we focus on the situation in which the latent dynamical system (i.e. signaling network) can be described as a network of binary state variables with logistic activation functions. We show how candidate networks can be scored efficiently in this case and how topology learning can be performed via Markov Chain Monte Carlo (MCMC).

As a first step, we extensively tested our approach by applying it to several known network motifs (Feed Forward, Feed Backward, Bifan, Diamond, Protein Cascading) over a wide range of possible settings (e.g. different number of observations, time points). Moreover, we evaluated our method with synthetic data generated from ODE systems taken from the literature. As a next step we evaluate our method with data from the DREAM 8 challenge. In future work, we also plan to extend our method to incorporate multi-omics data and apply it to patient samples to identify disease related networks.

de.NBI Cloud - Compute Power for your Project

Peter Belmann

In life sciences today, the handling, analysis and storage of enormous amounts of data is a challenging issue. For example, new sequencing and imaging technologies result in the generation of large scale genomic and image data. Hence, an appropriate IT infrastructure is necessary to perform analyses with such large datasets and to ensure secure data access and storage. The fully academic de.NBI Cloud consists of compute centers in Bielefeld, Heidelberg, Giessen, Freiburg and Tübingen and provides more than 10000 compute cores, 110 TB RAM and 4 PB of storage capacity. It offers a solution to enable integrative analyses for the entire life sciences community and the efficient use of data in research. To a large extent, the de.NBI Cloud will close the gap of missing computational resources for researchers in Germany. The de.NBI Cloud operates the major service levels. (1) Infrastructure as a Service (IaaS) - suited for power users requiring full control of the computing environment. (2) Platform as a Service (PaaS) - provision a fully operational infrastructure and frameworks for the deployment of bioinformatics workflows (3) Software as a Service (SaaS) - access to preconfigured, state-of-the-art bioinformatics pipelines and analysis tools. Through a cloud federation concept, all five de.NBI sites are integrated into a single cloud platform. The central access point to the de.NBI cloud platform is represented by the de.NBI Cloud portal. It is a crucial part of the federated cloud concept and will offer an easy entry point for de.NBI Cloud users and an important management tool for cloud and project administrators. In collaboration with Elixir the de.NBI Cloud portal manages the authorisation of users and offers Single Sign On to all cloud centers.

Towards Enabling Virtual Clinical Studies with Longitudinal Bayesian Network Modeling

Meemansa Sood, Akrishta Sahay, Reagon Karki, Martin Hofmann Apitius, Holger Fröhlich

The unclear etiology of diseases like Alzheimer's Disease (AD) and Parkinson's Disease (PD) still possesses a big challenge in their early diagnosis and treatment. Because of the failure of several clinical trials around established hypotheses in these diseases, lately, they are regarded as a complex and multi-factorial disorders. This has acquainted the scientific community with the need of agglomerating and understanding data at clinical (demographics, cognitive tests and functional tests), molecular (omics), genomic (SNPs/copy number variants) and neuro-imaging (brain volumes) level. Two such multi-scale and heterogeneous data sets, namely Alzheimer's Disease Neuroimaging Initiative (ADNI) data and Parkinson's Progression Markers Initiative (PPMI), have been used intensively for research purposes. These studies have the potential to understand and model diseases in a longitudinal manner. In this work we showed that it is possible to describe the entire course of a multivariate clinical trial with the help of a Bayesian Network (BN), which specifically takes into account that typically a considerable amount of patients drops out of a study before completion. As BNs are generative models our approach allows for simulating virtual patient cohorts, which has the potential to address principle concerns about data privacy of medical data as well as limitations of sample size. We performed rigorous comparisons of virtual vs real patients to demonstrate the principle feasibility of our approach. Moreover, we showed the possibility to simulate counterfactual interventions into a clinical study cohort. The idea of virtual patients proposed by this work holds promise to facilitate virtual clinical trials as well as increase the statistical power of trials.

Gaining New Knowledge on the Cell biological Processes of Cancer by Interpretable Machine Learning

Alfred Ultsch

Biological processes on the cellular level are to a large extent steered by gene products. One process, for example, the production and release of exosomes into the cellular microenvironment, is influenced by many genes. On the other hand, one gene may be involved in many cell biological processes. This work describes how a large collection of genes identified to be relevant for a certain type of cancer can be translated into interpretable knowledge on cell biological processes. Overrepresentation analysis (ORA) [1] translates the gene collection to knowledge on cell biological processes [2]. Using a scientifically acceptable p-value limit for the ORA translates the n (typically 500-1500) genes into a knowledge representation in form of a directed acyclical graph (DAG) of precisely defined knowledge items (terms). However, due to the many-to-many relationship from genes to processes, this results in conspicuous large DAGs which possess a number of terms in the order of n, even after a rigorous correction for multiple testing. In this work the steps to separate relevant from irrelevant terms are described. Fuzzy set [3] and information theory [4] are used for the weighting of the knowledge representing terms. Computed ABC Analysis [5] is used to decide which terms are relevant. Functional Abstraction [6] is used to present the knowledge in a human interpretable form. The resulting knowledge is called focused knowledge. The success of the approach is demonstrated on several types of cancer, such as Melanoma, AML, APL the cancer consensus genes [7] and the so called hallmarks of cancer [8, 9].

[1] Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Müller, R., Meese, E. and Lenhof, H.P., 2007. GeneTrail—advanced gene set enrichment analysis. Nucleic acids research, 35(web), pp.W186-W192.

[2] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. and Harris, M.A., 2000. Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), p.25.

[3] Dubois, D., Hüllermeier, E. and Prade, H., 2006. A systematic approach to the assessment of fuzzy association rules. Data Mining and Knowledge Discovery, 13(2), pp.167-192.

[4] Ultsch, A., 2005. Pareto density estimation: A density estimation for knowledge discovery. In Innovations in classification, data science, and information systems (pp. 91-100). Springer, Berlin, Heidelberg.

Clustering

Ensemble Clustering for Symbolic Data for Green Growth Analysis

Marcin Pełka

It has been twenty years since the first Rio Summit and the all countries around the world are facing two challenges. The first one is the expanding economic opportunities for all the context of a growing global population. The second one is environmental pressures, that if left unaddressed could undermine the abilities to seize the opportunities. Green growth is where these two challenges meet. It means fostering economic growth and development while ensuring that natural assets continue to provide the resources and environmental services on which our well-being relies. Green growth is a very important problem that is described by many authors (see for example Ekins 2002, Lorek & Spangenberg 2014) The proposed paper presents a green growth ensemble clustering of the OCED countries with the application of symbolic data analysis. The results are compared with single clustering using adjusted Rand index.

Keywords: symbolic data analysis, ensemble clustering, green growth

1. Bock, H.-H., Diday E. (Eds.), (2000), Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Berlin-Heidelberg.

2. Ekins, P., (2002), Economic growth and environmental sustainability: the prospects for green growth. Routledge, New York.

3. Lorek, S., Spangenberg, J.H., (2014), Sustainable consumption within a sustainable economy–beyond green growth and green economies. Journal of cleaner production, 63, 33-44.

4. Platform, G.G.K., (2013), Moving towards a common approach on green growth indicators. A Green Growth Knowledge Platform Scoping Paper, 46.

The Performance of Tube Distance in Clustering Tasks

Andrzej Sokołowski, Małgorzata Markowska, Sabina Denkowska

The tube distance has been introduced by Sokołowski et. al (2018). It is based on euclidean distance, and the adjustment is being made, taking into account points lying in-between those involved in distance calculations, within a given radius from the line connecting these two points. Distributions of tube distance has been studied under one-group population and under clustered structure. In this paper we want to compare the performance of the tube distance in clustering tasks, with the application of agglomerative methods and k-means. Results are compared with those obtained using classical euclidean distance, squared euclidean distance and Manhattan distance. Artificial, simulated data sets will be used as well as real applications, including well known sets (like Fisher iris data) and Eurostat data on EU NUTS 2 regions.

Two-Step Clustering of Micro Panel Data

Maria Stachova, Lukas Sobisek

Micro panel data contains statistical objects that are periodically observed over time. Compared to macro panel, in micro panel data the number of repeated measurements is significantly smaller and number of objects is significantly larger. Because supply of the clustering methods for these type of data is limited, the main goal of this contribution is to devise its own two-step approach. In the first step, the panel data are transformed into a static form using a set of proposed characteristics of dynamics. These characteristics represent different features of time course of the observed variables. In the second step, the elements are clustered by conventional spatial clustering techniques (agglomerative clustering and the K-means partitioning). The clustering is based on a dissimilarity matrix of the values of clustering variables calculated in the first step. Based on the simulation studies, the best combination of selected characteristics of dynamics and conventional clustering techniques is then chosen.

A Proposal of a New PAM-Like Clustering Algorithm for Symbolic Data

Marcin Pełka, Andrzej Dudek

In general terms, clustering methods seek to organize certain sets of objects (items) into clusters in the way allowing objects from the same cluster be more similar to each other than to objects from other clusters. Usually such similarity is measured by some distance measure (e.g. Euclidean, Manhattan, etc.).

Also in symbolic data analysis, where objects can be described by various variables (interval-valued, histogram, multi-valued, etc.), clustering methods usually use some dissimilarity measure to cluster objects into groups (see e.g. Verde 2004; Billard & Diday 2006).

The proposed paper presents a proposal of a new clustering algorithm for symbolic data that uses PAM-like approach (Kaufman & Rousseeuw 1990) where instead of distance measure the generality degree measure is used to build clusters of object that share similar variable properties.

Keywords: symbolic data analysis, clustering, PAM

1. Billard, L., & Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley.

2. Brito, P. (2002). Hierarchical and pyramidal clustering for symbolic data. Journal of the Japanese Society of Computational Statistics, 15(2), 231-244.

3. Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis, 68-125.

4. Verde, R. (2004). Clustering methods in symbolic data analysis. In Classification, clustering, and data mining applications (pp. 299-317). Springer, Berlin, Heidelberg.

On the Selection Uncertainty in Parametric Clustering

Alessandro Casa, Luca Scrucca, Giovanna Menardi

The selection of a specific model from a given set of alternatives is a crucial step in any data analysis. Usually this step preceeds inference, which is carried on without taking into account the selection procedure, hence considering the chosen model as fixed. However, model selection is itself data dependent and, as such, possesses its own variability. Drawing inference without accounting for the selection of the model is equivalent to neglect a source of uncertainty and lead to anti-conservative statements. An alternative to rely inference on a single model is to use a weighted average of the estimates resulting from each of the models under consideration. This leads to the class of model averaging estimators. In the context of parametric clustering the model selection step involves the choice of the number of components of the mixture model (i.e. the number of clusters), the parametrization of the model and possibly even the choice of the component densities. The aforementioned issues are particularly important since the "single best model" paradigm is the predominant one in the framework; several models are fitted and, among them, the best one is selected according to information criteria such as the BIC, and used to draw the subsequent conclusions. Doing so all the fitted models but the best one are thrown away even when they are possibly providing reasonable partitions of the data. In the process some useful information could be lost, especially if the values of the information criterion for the discarded models are close to the one of the selected model. In this work the ideas on which stacking is based on are adapted in an unsupervised framework (Smyth and Wolpert, 1999) in order to obtain a viable solution to workaround the aforementioned issues. The proposed methodology is tested both on simulated and real data and its performances are compared with some competitive methods (Wei and McNicholas, 2015, Russel et al, 2015).

Russell, N., Murphy T. B., and Raftery A. E. (2015). Bayesian model averaging in model-based clustering and density estimation, arXiv preprint arXiv:1506.09035.

Smyth, P. and Wolpert, D. (1999). Linearly combining density estimators via stacking, Machine Learning, 36, 59–83.

Wei, Y. and McNicholas, P. D. (2015). Mixture model averaging for clustering, Advances in Data Analysis and Classification, 9(2), 197–217.

Comparison and Benchmarking of Cluster Analysis Methods

The Threats and Traps in Benchmarking of Cluster Analysis Methods

Andrzej Dudek, Marcin Pełka, Marek Walesiak

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

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

Due to those challenges measuring the quality of clustering method is complicated multi-criterial procedure taking into account various cluster shapes, the stability of partitions, the computational complexity of methods and finally the goal to be achieved through clustering procedure.

The paper describes the common pitfalls that even experienced researchers may encounter in benchmarking of cluster analysis methods such as relying only on normal distribution clusters or not to measure the stability of partitioning.

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

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

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

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

Some Thoughts on Simulation Studies to Compare Clustering Methods

Christian Hennig

Simulation studies are often used to compare different clustering methods, be it with the aim of promoting a new method, or for investigating the quality of existing methods from a neutral point of view. Having been involved in a number of such studies (e.g., Coretto and Hennig 2016, Amorim and Hennig 2015, Anderlucci and Hennig 2014) , I will go through a number of aspects of designing and running such studies, including the definition and measurement of clustering quality, the choice of models to generate data from, aggregation and visualisation of results, and also limits of what we can learn from such studies. Some of this work was produced in collaboration with the IFCS Cluster Benchmarking Task Force (Van Mechelen et al. 2018).

L. Anderlucci and C. Hennig (2014) Clustering of categorical data: a comparison of a model-based and a distance-based approach. Communications in Statistics - Theory and Methods 3, 704-721.

R. C. Amorim and C. Hennig (2015) Recovering the number of clusters in data sets with noise features using feature rescaling factors. Information Sciences 324, 126-145.

P. Coretto and C. Hennig (2016) Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. Journal of the American Statistical Association 111, 1648-1659.

I. Van Mechelen, A.-L. Boulesteix, R. Dangl, N..Dean, I. Guyon, C. Hennig, F..Leisch, and D. Steinley (2018) Benchmarking in cluster analysis: A white paper. Submitted.

Towards Automatic Assessment of Clustering Quality

Andrey Filchenkov, Sergey Muravyov

Clustering is a generally stated problem that is grouping objects with respect to their similarity. Since grouping is a simplification that helps one to handle with complex situations and environments, no surprise that cluster analysis is used in most of domains, such as geography, medicine, chemistry, economics, sociology and many others.

A challenging sequence of such generality is that clustering is an ill-defined problem. This nourishes plethora of mathematical formalizations, with which clustering is considered as an optimization problem with a broad choice of what can be used as its target function. These target functions measure how well objects are assigned to groups called clusters. Nowadays, there are two types of such criteria: external and internal.

Evaluation with external measures is based on information that was not used for clustering, such as known class labels and external benchmarks. We must point out that this approach has theoretical restrictions. They most important one comes from the fact that objects may be equally well grouped in a several different ways, while using a single labeling does not allow to work properly with all other partitions. Usage of external measures thus may be considered as at least questionable.

Evaluation with internal measures is based only on data partition. These metrics usually assign the best score to an algorithm that returns partition with high similarity within a cluster and low similarity between clusters. There are plenty of different internal metrics nowadays, and they appear at a very high rate. We will refer to these metrics as cluster validity indices (CVI).

There are several works that are devoted to the comparison of different CVIs properties, but all of them state different CVI to be the best on certain types datasets that have structure of well-separated hyper spheres. Thus, we can conclude that there is no perfect CVI to this moment.

In this work, we study how well CVIs can reflect the clustering quality. We use assessments of assessors for cluster partition adequacy as the ground truth and explain, why this is the only measure that can be used in this quality. We compare different clustering validity indices and show that none of them can be the universal, reflecting quality for each cluster partition. To do so, we introduce four quality measures for CVI evaluation. Also, we suggest a novel meta-learning based approach for predicting the best CVI for a given dataset.

Towards Evidence-Based Computational Statistics: Lessons from Clinical Research on the Role and Design of Real-Data Benchmark Studies

Anne-Laure Boulesteix, Rory Wilson, Alexander Hapfelmeier

The goal of medical research is to develop interventions that are in some sense superior, with respect to patient outcome, to interventions currently in use. Similarly, the goal of research in methodological computational statistics is to develop data analysis tools that are themselves superior to the existing tools. The methodology of the evaluation of medical interventions continues to be discussed extensively in the literature and it is now well accepted that medicine should be at least partly “evidence-based”. Although we statisticians are convinced of the importance of unbiased, well-thought-out study designs and evidence-based approaches in the context of clinical research, we tend to ignore these principles when designing our own studies for evaluating statistical methods in the context of our methodological research.

In this paper, we draw an analogy between clinical trials and real-data-based benchmarking experiments in methodological statistical science, with datasets playing the role of patients and methods playing the role of medical interventions. Through this analogy, we suggest directions for improvement in the design and interpretation of studies which use real data to evaluate statistical methods, in particular with respect to dataset inclusion criteria and the reduction of various forms of bias. More generally, we discuss the concept of "evidence-based" statistical research, its limitations and its impact on the design and interpretation of real-data-based benchmark experiments.

We suggest that benchmark studies—a method of assessment of statistical methods using real-world datasets—might benefit from adopting (some) concepts from evidence-based medicine towards the goal of more evidence-based statistical research.

Estimating the Quality of an Optimal Treatment Regime

Aniek Sies, Iven Van Mechelen

When multiple treatment alternatives are available for a disease, an obvious question is which alternative is most effective for which patient. One may address this question by searching for optimal treatment regimes that specify for each individual the preferable treatment alternative based on that individual's baseline characteristics. Optimal treatment regime estimation comes down to a clustering problem with a well-defined, pragmatic optimality criterion (viz., maximizing the expected outcome if the regime were used for treatment assignment of all patients in the population under study). When an optimal treatment regime has been estimated, its quality (in terms of the true value of the optimality criterion) is of obvious interest. Obtaining a good and reliable estimate of this quantity is a key challenge for which so far no satisfactory solution is available. In this presentation, we consider for this purpose several estimators of the expected outcome in conjunction with several resampling methods. The latter have been evaluated before within the context of statistical learning to estimate the prediction error of estimated prediction rules. Yet, the results of these evaluations were equivocal, with different best performing methods in different studies, and with near-zero and even negative correlations between true and estimated prediction errors. Moreover, for different reasons it is not straightforward to extrapolate the findings of these studies to the context of optimal treatment regimes. To address these issues, we set up a new and comprehensive benchmarking study with simulated data. In this study, combinations of different estimators with a .632+ bootstrap resampling method performed best. In addition, the study shed a surprising new light on the previously reported problematic correlations between true and estimated prediction errors in the area of statistical learning.

External Validity Indices and Cluster Size Imbalance

Matthijs Warrens

To evaluate the performance of a clustering method researchers typically assess the agreement between a reference standard partition that purports to represent the true cluster structure of the objects, and a trial partition produced by the method that is being evaluated. High agreement between the two partitions indicates good recovery of the true cluster structure. Agreement between the partitions can be assessed with so-called external validity indices.

Researchers tend to use and report validity indices that quantify agreement between two partitions for all clusters simultaneously. External validity indices are usually divided into three approaches, namely, 1) indices based on counting pairs of objects, 2) indices based concepts from information theory, and 3) indices based on matching sets. Commonly used examples based on counting pairs are the Rand index and the Hubert-Arabie adjusted Rand index. A commonly used information theoretic measure is the normalized mutual information.

Many external validity indices are sensitive to imbalanced cluster sizes. This usually means that larger clusters have a bigger impact on the index value than smaller clusters. Normalization of indices has been proposed as a solution to the effect of cluster size imbalance.

In this presentation we present algebraic analyses of commonly used external validity indices. Various indices turn out to be weighted means (variously defined) of indices that reflect agreement on individual clusters. Furthermore, normalization of indices does not solve sensitivity to cluster size imbalance. Moreover, the algebraic analyses reveal that some indices are quite robust to cluster size imbalance. The latter indices tend to reflect agreement on all clusters more evenly, instead of reflecting agreement on the larger clusters only. The analyses are used to quantify robustness to cluster size imbalance. The quantification results in an ordering of the various external validity indices with respect to robustness to cluster size imbalance.

Benchmarking Cluster Analysis Methods using PDE-Optimized Violin Plots

Michael Christoph Thrun, Felix Pape, Alfred Ultsch

Conventional clustering algorithms could be limited in their clustering ability in the presence clusters defined by a combination of distance- and density-based structures [1-6]. A benchmark set of several artificial datasets with predefined cluster structures (FCPS) [7] is used to thoroughly inquire into this property of data sets. This approach allows the comparison of conventional clustering algorithms with simultaneous restriction to searching for distance and density-based structures. Seven conventional algorithms (k-means, model-based clustering, affinity propagation clustering, spectral clustering, partition around medoids, ward and single-linkage) with clustering criteria and Databionic swarm with no global objective function [8] are compared using a Monte Carlo approach. The best of all permutation of labels with the highest accuracy is selected in every trial because algorithms arbitrarily define the labels. Then, the probability density of each algorithm is compared with every other for each dataset by using violin plots [9]. Pareto density estimation (PDE) is used to improve the violin, or the so-called bean plot. PDE is particularly suitable for the discovery of structures in continuous data, hinting at the presence of distinct groups of data [10]. The PDE-optimized violin plots indicate that the clustering algorithms sometimes produce different results of highly varied quality depending on a trial. Only if the predefined data structures happen to meet the particular criterion, the clusters are recovered by conventional clustering algorithms. Using insights into graph theory, our benchmarking suggests that the investigated clustering criteria impose specific structures on the data. PDE-optimized violin plot is available on CRAN in the R package “DataVisualizations”.

1. Duda, R.O., P.E. Hart, and D.G. Stork, Pattern Classification. Second Edition ed. A Wiley-Interscience Publication, Ney York, USA: John Wiley & Sons, 2001.

2. Everitt, B.S., S. Landau, and M. Leese, Cluster analysis. Fourth Edition ed, London: Arnold, 2001.

3. Handl, J., J. Knowles, and D.B. Kell, Computational cluster validation in post-genomic data analysis. Bioinformatics, 21(15): p. 3201-3212, 2005.

4. Theodoridis, S. and K. Koutroumbas, Pattern Recognition. Fourth Edition ed., Canada: Elsevier. 961, 2009.

5. Ultsch, A. and J. Lötsch, Machine-learned cluster identification in high-dimensional data. Journal of Biomedical Informatics, 66(C): p. 95-104 2017.

6. Jain, A.K. and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall Advanced Reference Series : Computer Science, ed. B. Marttine. Vol. 3, Englewood Cliffs, New Jersey, USA: Prentice Hall College Div,1988.

7. Ultsch, A. Clustering wih SOM: U* C. in Proceedings of the 5th Workshop on Self-Organizing Maps. 2005.

8. Thrun, M.C., Projection Based Clustering through Self-Organization and Swarm Intelligence, Doctoral dissertation, Heidelberg: Springer, 2018.

9. Hintze, J.L. and R.D. Nelson, Violin plots: a box plot-density trace synergism. The American Statistician, 52(2): p. 181-184, 1998.

10. Ultsch, A., Pareto density estimation: A density estimation for knowledge discovery, in Innovations in classification, data science, and information systems, D. Baier and K.D. Werrnecke, Editors, Springer: Berlin, Germany. p. 91-100, 2005.

Computational Social Science

Minorities in Social Networks

Claudia Wagner

Networks are the infrastructure of our social and professional life and also of modern information systems where billions of documents and entities are interlinked. However, not all nodes are equal. Often we observe attributes (e.g. gender or ethnicity) that define the group membership of a node. This assignment may lead to the emergence of minority groups. In this talk I will explore the role of minorities in social networks and discuss a simple mechanism that partly explains why they are often less popular and successful.

Community Analysis based on linguistic characteristics in Social Networks

Kristi Lubonja, Mirco Schönfeld

Online social networks have become more important over the last years. Therefore, the problem of analyzing them for purposes of marketing, politics or understanding of social relationships has gained more attention. A wide-spread interest in analyzing social networks is to detect and distinguish different communities that use them. In this paper, we attempt to distinguish different communities in a popular social network named Reddit by exploiting the comments in it and classifying them according to the community they belong in. Rather than representing the comments by using the popular bag-of-words model, which is not very suitable for short text, we use a set of descriptive features. The selected set of attributes aims to measure complexity, readability, subjectivity, and other similar traits of the comment text and use them as a distinguishing tool for communities. Popular and widely used classification algorithms are then used for the classification process in order to evaluate the selected set of features. Our results support a conclusion of a differing use of language in different communities.

Reputation through Observation: Active Lurkers in Online Communities

Clemens Niemeyer, Mirco Schönfeld

Lurkers becoming an active part of an online community is a seldom phenomenon. By definition, lurkers are users that silently observe, consume, and become accustomed to a community without interacting themself. At some point in time, a small fraction of lurkers decides to start commenting, interacting, or taking part in a community in some other way. In this paper we investigate the implications of lurking for the interactions of such newly-active users. Based on an analysis of users and user-generated comments from the well-known Online Social Network Reddit, we propose a framework for tracking linguistic development as well as for comprehending users' reputation inside such communities. This allows for new insights into the effects of lurking not only as a matter of linguistic adaption of community habits but also as a matter of reputation former lurkers are able to gain.

(Automated) Text Analysis of German Online Participation Projects with an Interdisciplinary Approach from Computer Science and Communication and Media Studies

Matthias Liebeck, Katharina Esau

In the last couple of years, more and more cities in Germany have offered their citizens online-based discussion platforms on which the citizens are able to participate in decision-making processes, for instance on how the budget of the next fiscal period should be allocated. In this talk, we highlight how we combined the strengths of the two disciplines computer science and communication and media studies in order to analyze text content from such online participation projects. We focused on two online participation projects in the political domain: (i) the "Tempelhofer Feld" project which revolved around the future use of a former airport in Berlin; (ii) the "Braunkohle" project which focused on the future of lignite mining in Germany.

With our talk, we want to share our good experiences from the last three years and encourage more researchers to work in interdisciplinary teams. We talk about how our research began and how our first steps of interdisciplinary research looked like. Then, we deliver insight into our annotation processes, by detailing how we developed our codebook for our manual annotation with a web-based annotation tool. Afterwards, we showcase technical solutions that we developed to correct and visualize our annotated datasets.

We annotated a variety of variables that are interesting for both disciplines. From a computer science perspective, we focused on argument components. From a social sciences perspective, we are interested in variables for discourse analysis (e.g., emotions and narrations) and how they influence the discussion dynamic between the users. Our annotations allowed us to evaluate the two machine learning tasks of automatically identifying argumentative content and classifying argument components. We present some of our results based on our work (Liebeck et. al., 2016) in the evolving research field of argument mining.

We finish our talk with an outline of our future work in discourse analysis and a summary of advantages and problems of working in interdisciplinary teams.

M. Liebeck, K. Esau, S. Conrad (2016). What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld. Proceedings of the 3rd Workshop on Argumentation Mining, pp. 144-153

A Practical Approach to Tackling Fake News

Martin Potthast

The fake news hype has given rise to a frenzy of activity, research and otherwise, into the question of how to deal with online news whose authors disregard journalistic standards and ideals. This pertains particularly to news that aim at misleading their readers by giving only an extremely one-sided (hyperpartisan) account of a given event.

We hypothesize that hyperpartisan news exhibit writing style characteristics that render them distinguishable from more nuanced reporting, and our contribution is a first corresponding investigation. For this purpose, 1,627 articles from 9 political publishers, three each from the mainstream, the hyperpartisan left-wing, and the hyperpartisan right-wing, have been fact-checked by professional journalists: from the 299 identified fake news articles 97% are also hyperpartisan. We show how via a style analysis hyperpartisan news can be distinguished from mainstream news (F1=0.78), and satire from both (F1=0.81). As one may have expected, style-based fake news detection did not work (F1=0.46), showing that stylometry is not a silver bullet for fake news in general. Our analysis also reveals that the style of left-wing and right-wing news share significantly more similarities than does either of them with mainstream news. This result is robust: it was obtained with three different modeling approaches, one of which employs the ”Unmasking” technology in a novel way to assess style similarity. Applications of our results include prescreening for semi-automatic fake news detection.

Discourse Analysis as an Information Retrieval Problem

Tim Gollub, Henning Schmidgen, Benno Stein

Presented is our ongoing research on information retrieval technologies for the analysis of discourses.

In the social sciences, discourse analysis meanwhile is an established research paradigm. Based on a corpus of relevant documents, discourse analysis aims to uncover the social formations, practices, and regulations that sway the structure and the dynamics of a discourse.

To support the systematic analysis of discourses, we propose an analytical layer on top of a keyword-based fulltext search engine that indexes a discourse corpus. The analytical layer communicates with the search engine via ordinary search queries and result lists, while itself providing extended search and analysis functionality to the user via a so called discourse query language.

Beyond keywords, the discourse query language supports the use of semantic units in search queries, namely concepts, facets, and frames. For query processing, an extensible semantic knowledge base is used to translate semantic units into keyword-based queries that are ’understood’ by the underlying search engine. Users have the possibility to add and actively train semantic units that are specific to the underlying discourse or their research question.

The discourse query language further facilitates analytical operations on search results. In addition to ’find’ as default request operation, discourse related operations such as ’characterize’, ’compare’, ’contrast’, ’cluster’, and ’connect’ are available. Each operation refers to a data analysis task that is performed on the semantic bipartite graph that spans between the relevant documents of a discourse query and the semantic units contained in them.

By building our data analysis technology on top of a search engine, we pick up on technology that is exceptionally well accepted in humanities research. We speculate that the exploitation of this established point of contact between computer science and the humanities maximizes the chances of a successful uptake of our technology in the humanities.

Consumer Preferences and Marketing Analytics

Identifying Nested Preference Structures in Choice-Based Conjoint Analysis: A Simulation Study

Nils Goeken, Peter Kurz, Winfried J. Steiner

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. The MNL model, however, suffers from the Independence of Irrelevant Alternatives property (IIA), which implies proportional substitution patterns across alternatives. A further limitation of the basic MNL model is its inability of addressing market heterogeneity. The recent development of Hierarchical Bayesian (HB) estimation methods for MNL models enables the estimation of part-worth utilities at the individual level, thereby also relaxing the IIA property at the aggregate level. A further relaxation of the IIA property can be achieved by applying Bayesian nested multinomial logit (NMNL) models. By allowing for different degrees of similarity between subsets of alternatives (nests), market structures for the intuitive modeling of consumers’ choice behavior can be depicted.

The aim of this research is to compare the Bayesian MNL and the Bayesian NMNL model in a detailed simulation study (N=1280). To the best of our knowledge, no simulation study has yet systematically explored the performance of Bayesian individual NMNL models. Using statistical criteria for parameter recovery, goodness-of-fit and predictive accuracy we evaluate the performance of both types of models under varying sample sizes, different degrees of similarity (log-sum coefficients), different levels of heterogeneity, varying number of alternatives within a choice task and different numbers of parameters at the individual level. To assess the generalizability of the results and to investigate the influence of various bimodal preference structures we further conduct a sensitivity analysis (N=160).

Our results show that the choice of the model has only a moderate influence on our measures of performance. Despite the unimodal prior assumption of the MNL, we can see that both models are able to identify and to predict bimodal preference structures with correlated alternatives. In the presence of a nested structure with high correlations (small log-sum coefficients) between alternatives the NMNL model shows slight advantages compared to the MNL model in terms of parameter recovery ($p < .001$). Furthermore, small log-sum coefficients affect the fit (effect sizes > 0.14) and the prediction accuracy (effect sizes > 0.06) positively.

Advertising for a Scientific Publishing Service in Social Media Networks: Effects on Reach and Prominence

Victoria-Anne Schweigert, Andreas Geyer-Schulz

This contribution consists of two parts, namely first the description of web activity measurement methods compatible with the German Data Protection Law and second the presentation of an exploratory analysis of the data collected. The first part documents the infrastructure setup and the data model followed by a discussion of measurement problems resulting from legal restrictions. The second part concentrates on empirical insights on user reaction patterns on different platforms and their effects on reach and prominence based on first experiments.

Keywords: Social Media Marketing, Advertising, User Reaction, Scientific Publishing

On the Effect of HB Covariance Matrix Prior Settings: A Simulation Study

Maren Hein, Peter Kurz, Winfried J. Steiner

The key strength of HB is its ability to provide individual part-worth estimates even if relatively few observations per respondent are available. In order to estimate individual-level part-worths, HB inference requires prior beliefs about the unknown parameters. These prior beliefs are represented by a probability distribution. According to the Bayes theorem, Bayesian estimators combine the prior information about the part-worths with information contained in the choice data to arrive at the posterior distributions of the parameters. In the Hierarchical Bayes model, the prior is specified in a two-stages process. The multivariate normal distribution is typically used as probability distribution (first-stage prior) with a vector of means and a covariance matrix that captures the extent of heterogeneity across individuals. In a second step, HB also captures uncertainty on the parameters of the prior distribution by treating them as random variables. Therefore, to estimate the population mean and the covariance matrix of the prior distribution, hyperprior distributions (second-stage priors) have to be specified. In the frequently used ”normal model” it is common to assume that the prior on the population mean is normal and the prior on the covariance matrix is inverse Wishart distributed.

In this context the objective of our study is to substantially contribute to the question how HB covariance matrix prior settings (i.e. the prior variance and the prior degrees of freedom) affect the performance of HB-CBC models. We therefore provide an extensive simulation study based on synthetic choice-based conjoint data. We investigate the influence of the HB priors by systematically varying experimental factors, such as the number of respondents, the number of choice tasks and the number of parameters to be estimated. Overall, the statistical performance of HB is evaluated under experimentally varying conditions based on seven experimental factors using criteria for goodness-of-fit, parameter recovery and predictive accuracy. Moreover, analyses of variance (ANOVAs) are conducted to assess the impact of the experimental factors on the measures of performance. In addition, a sensitivity analysis is proposed to test even more extreme prior variance levels.

The results indicate that the prior degrees of freedom play a negligible role as there is not any noticeable impact on the performance of HB when varying that factor. For increasing prior variance levels overfitting problems occur with respect to parameter recovery and model fit. The most striking finding is that the predictive performance of HB-CBC is not markedly affected by an increase of the prior variance.

Dynamic Structural Equation Models of Momentary Assessments in Consumer Research

Adam Sagan

Development of modern techniques of momentary assessment of consumer behavior and experience sampling like smartphones, smartglasses, smartwatches, activity trackers, day reconstruction diaries, electronically activated recording, allows obtaining an intensive longitudinal data (ILD) on the level of individual consumers (Hektner, Schmidt, Csikszentmihalyi, 2012). These methods help to avoid the retrospective errors in the process of data gathering like personal heuristics, recency, salience and mood-congruent memory effects and increase the ecological validity of the data. The aim of the paper is to explain the regularities in momentary consumer behavior, emotional states, life events and everyday consumer experiences on the basis of dynamic structural equation models (DSEM). Selected DSEM models will be estimated and compared that represent the generalized approach to ILD that involve time series analysis, latent variable models, multilevel analysis and time - varying effect modelling (Asparouchov, Hamaker, Muthen 2017).

1. Asparouchov, T., Hamaker, E., L., Muthen, B. (2017). Dynamic Structural Equation Models, “Structural Equation Modeling: A Multidisciplinary Journal”, 00, s. 1 – 30

2. Hektner, J. M., Schmidt, J. A., Csikszentmihalyi. (2012). M. Experience Sampling Method: Measuring the Quality of Everyday Life. Thousand Oaks, CA: Sage, 2012

Lexicographic Preferences in Customer Review Data Following a Criterion-Based Approach

Michael Bräuning

So-called non-compensatory strategies, including lexicographic orders, have already received some attention in literature especially in marketing sciences. This work builds on the studies regarding lexicographic orders and focuses on the field of preference learning. The main contribution results from introducing a criterion-based lexicographic preference structure: several attributes can be mapped, by aggregating or grouping, to one criterion. As part of a case study, criterion-based lexicographic preference lists (CBLP-lists) are applied to customer review data such as those provided by Amazon or TripAdvisor. The resulting CBLP-lists are finally assessed in general and the proposed criteria in particular.

Making a lexicographic decision between two alternatives means that a decision is made at the first attribute the alternatives differ on where attributes are ranked according to their importance. Succeeding attributes are not taken into account. There is a major concern that this strategy might be overly restrictive compared to compensatory structures. This issue is addressed by relaxing the underlying assumptions. Instead of assessing one attribute at a time, several attributes can be mapped to one criterion and assessed together. A customer might be more interested in a museum’s average rating score rather than the museum’s total number of rating scores equal to three for example.

Pairwise preferences are derived from Amazon and TripAdvisor review data for learning and assessing CBLP-lists. Moreover, some mappings from attributes to criteria comprising the median or mean are defined and considered during the learning process. My first research findings indicate that customers put more weight on aggregated attributes compared to individual attributes. The case study is complemented by a comparison to conventional models like ordered logit or probit and the more sophisticated SVMrank algorithm.

Data Analysis Models in Economics and Business

Data Mining Models in the Evaluation of the Importance of Financial Indicators for Firms’ Financial Condition Assessment

Józef Pociecha, Barbara Pawełek

In theory and practice, firms financial condition assessment uses different methods based on financial indicators. Data Mining models are useful tools for assessing the financial condition in the context of bankruptcy risk. Among them we distinguish ”black box” models and ”white box” models. The computational complexity of the ”black box” models impedes identification factors supporting the assessment of firms financial condition. Searching for financial indicators to warn the possible bankruptcy is very important for the practice. The aim of the paper is to present the results of empirical investigations on the evaluation of selected financial indicators in the assessment of firms financial condition, in the context of their threat for bankruptcy. The added value of the work is the application of Sensitivity Analysis to the evaluation of the significance financial indicators in the assessment of the financial condition of companies and the deepening methods of firms financial analysis. The analysis uses financial data of companies in the industrial processing sector of Poland in the years 2005-2008. The study included such methods as: classification trees, bagging and boosting mathods, randon forests, k-nearest neighbours, support vector machines, neural networks and naïve Bayes classifier. Evaluation of the significance selected financial indicators in the assessment of companies financial condition was carried out using three algorithms Sensitivity Analysis (SA): Data-based SA, Monte-Carlo SA, Cluster-based SA. Calculations were performed in R using packages ’rminer’.

Key words: Data Mining; financial condition of company; Sensitivity Analysis; financial indicator.

Cortez P., Embrechts M.J. (2013), Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models, “Information Sciences”, Elsevier, Vol. 225, p. 1-17, doi: 10.1016/j.ins.2012.10.039.

Pawełek B., Pociecha J. (2017), Sensitivity Analysis in Corporate Bankruptcy Prediction, [in:] Book of Short Papers, eds. F. Greselin, F. Mola, M. Zenga, CLADAG2017, pp. 295-300.

Foreign Trade Effects on Regional Growth in Ukraine

Victor Shevchuk

The effects of foreign trade in general and foreign output in particular upon regional growth in Ukraine are estimated with the dynamic Arellano-Bond estimator. Annual dataset of the 2000-2015 period is used. It is found that an increase in foreign income, as measured by GDP growth in the Eurozone and several neighbouring countries (Hungary, Poland, Russia, Romania, Slovakia), contributes to regional growth in Ukraine, with both exports and imports being increased. However, there are asymmetric effects of exports and imports upon regional growth. There is a strong standard positive impact of exports upon regional growth, with a negative effect of lagged imports. Both regional growth and foreign trade are subject to gravity effects, being negatively correlated with a distance from the Western border of Ukraine. Among other results, there is a restrictionary effect of exchange rate depreciation upon regional growth. Although a weaker local currency is contributing to exports, there is a corresponding increase in imports either. As expected, higher investments in physical capital are a factor of both regional growth and foreign trade dynamics.

An Approach towards a Decentralized and Forecast-based Energy Trading Model

Moritz Mönning, Gerrit Schumann

Increasing decentralization of an energy production, which is primary driven by the increasing number of prosumers, who are also produce energy, and at the same time a decreasing number of pure consumers, means that the current energy market model needs to be revised. Meantime, more and more smart meters are installed in households. These meters record consumption data in real time and could enable an individual forecasts for each participating household. And usually, such forecasts are more accurate than a common standardized energy load profile. Due to that, it is possible to create an individualized forecast model for each household, such that the forecast is going to use measuring values from a household’s own production and consumption.

These forecasts provide the foundation for a prototypical decentralized energy trading model, which offers an allocation of decentralized producers and consumers taking into consideration regionality and transparency. The current work demonstrates the data analytics part of the energy trading model and shows a preliminary result how a building of a production and consumption predictions for the proposed energy trading model could be achieved.

In order to generate consumption forecasts, time series forecasting methods enriched with external data sources were used. Since recognitions of a seasonal habits in the historical data has a particularly large influence on the prediction quality, a consideration of multiple seasonalities is essential. At the beginning of the data analysis, the univariate methods such as Holt-Winters were applied, and later on, the multivariate methods such as Random Forest and Long Short-Term Memory were applied on the same datasets. At the end, the multivariate methods, which were using external data sources, enabled a higher precision over the univariate methods.

In contrast to the consumption of the energy in a household, the production of the energy is essentially influenced by the local weather conditions (e.g. photovoltaic panels). In order to be able to adequately consider these high influential factors, differentiated approaches were used. Thus, approaches of supervised machine learning are used for the production model. The methods used include Conditional Random Forest and Neural Networks. The goal was to derive a set of classification rules that were able to provide an optimal combination between weather factors and a household's individual production rate to be used in future forecast models.

Due to the increasing energy demand, individualized energy consumption and production profiles, the energy trading model combined with the data analytics approaches, such as introduced in this work, are going to be core stones of the future energy market.

Context-Sensitive Performance Benchmarking of a Portfolio of Industrial Assets

Alessandro Murgia, Elena Tsiporkova, Mathias Verbeke, Tom Tourwe

The key elements of Industry 4.0 are the sensorization, connection and continuous monitoring of assets. These assets generate data streams often recorded as time series. This data is then used for preventive maintenance and root-cause analysis. A common approach to exploit it is to extract various indicators in the form of time-sensitive production features as input for training machine learning models. Time-sensitive features do not necessarily take into account the operational context in which the different assets are operating. Consequently, they lead to models that are too specific for an individual asset and are difficult to generalise to a fleet of similar assets. This work presents a new method to (i) characterize the operational context of an asset, (ii) quantify the impact on its performance and (iii) identify subsets of similarly behaving assets. The operational contexts of the asset are extracted by exploring the streaming data associated to exogenous factors (e.g. temperature, humidity) in order to derive relatively homogeneous windows of operating and environmental conditions. Each context is described by a 'hypercube' of exogenous factor ranges. Subsequently, diverse performance indicators of the assets are extracted for each hypercube and used to identify similar assets according to their performance or to evaluate the influence of the operational contexts on a single asset's performance. This methodology allows to objectively benchmark asset performance across the fleet in a context-sensitive fashion. The discrimination and sensitivity potential of the method is evaluated on real-world data from a fleet of wind turbines. It is illustrated how the asset performance (energy production) is connected to the operational context (weather conditions). Moreover, it is investigated how the same operational context impacts the performance of the different assets in the fleet and how groups of wind turbines moving synchronously though different operational contexts can be determined.

Automatic Monitoring System for the Competency Gap Evaluation at the Russian and Polish Labour Market

Sergey Belov, Ivan Kadochnikov, Paweł Lula, Renata Oczkowska, Katarzyna Wójcik, Petr Zrelov

Comparative analysis of competency gap in Russia and Poland is the main goal of this presentation. To achieve the goal the monitoring system was proposed which can analyze educational programs and job requirements. The description of study programs and job offers obtained from different web sites were used as the main sources of information. To automate the processing as much as possible, we used machine learning technologies for semantic parsing. Creation of semantic models is one of the well-known key problems of natural language processing. In our research we used several methods for identification main semantic concepts: a) keywords- and phrases-based approach, b) Latent Semantic Analysis method, c) Latent Dirichlet Allocation method, d) ontology-based analysis. In the presentation we are going to show results concerning comparative analysis in the IT area. We would like to focus on: a) comparative analysis of main competences existing in study programs in IT field (we are going to investigate study programs realized at the Plekhanov Russian University of Economics in Moscow and Cracow University of Economics), b) comparative analysis of competences required by employers from the IT area in Russia and in Poland, c) estimation of competence gaps for these two countries.

Data Analysis in Finance 1

Mixture Models in Competing Risks Analysis. Application to Credit Risk Assessment

Ewa Wycinka, Tomasz Jurkiewicz

In survival analysis, time to event is a study object. Competing risks are considered if there are more than one types of events. Competing risks issue can be expressed as a bivariate random variable (T, C) in which T is a continuous variable representing time of the first event and $C = j (j = 0, 1, \dots, p)$ is a discrete variable denoting type of the first event $(j = 1, \dots, p)$. Marginal and conditional distributions of the bivariate random variable can be expressed in relation to joint distribution as $P(T = t|C = j) = (P(T = t, C = j))/(P(C = j))$ and $P(C = j|T = t) = (P(T = t, C = j))/(P(T = t))$, where $P(C = j)$ specifies the marginal distribution of type of the first

event and $P(T = t)$ is for the marginal distribution of time of the first event. Transformations of above formulas allow to decompose the joint distribution in a mixture model as
$P(T = t, C = j) = P(C = j)P(T = t|C = j)$
or
$P(T = t, C = j) = P(T = t)P(C = j|T = t)$.
The first decomposition is known as a horizontal approach, the other one as a vertical approach.

In this paper we analyse the consequences of applying each of above approaches for regression modelling. In horizontal approach, the multinomial logistic regression is used for the distribution of types of events and Cox PH model for the conditional distribution of times to the given type of event. Expectation-Maximization (EM) algorithm is usually applied to infer the missing cause of failure for censored observations. In vertical approach, a Cox PH model is used for the overall event time and a multinomial logit model for the type of event, given the event time. Parameters of both components of a vertical mixture model are estimated separately. In both approaches, each component model can consist a different set of covariates what gives a flexibility of modelling.

As an illustration of the issue we use mixture models in credit risk assessment where default and early repayment are considered to be competing events. Empirical research is conducted on the sample of credits granted by one of the Polish financial institutions. Covariates in regression models are characteristics of creditors and credits which are usually used in credit scoring.

Residual Based Consistent Bubble Detection

Leopold Sögner

We present a consistent monitoring procedure with the goal to detect bubbles or market dislocations in the presence of multiple cointegrating relationships. We show that after applying the between transform we obtain a model that allows to apply the monitoring tool developed in Wagner and Wied (2017). Then, modified least squares estimators are used to estimate the model parameter θ. An expanding window detector is used to test the null hypothesis of r cointegrating relationships and $\theta = \theta_0$ against the alternative of less cointegrating relationships or $\theta \neq \theta_0$. The monitoring procedure is applied to investigate possible dislocations in the Covered Interest Rate Parity.

Forecasting Non-Negative Financial Processes Using Different Parametric and Semi-Parametric ACD-Type Models

Sarah Forstinger, Yuanhua Feng, Christian Peitz

In this paper we propose the combination of non-parametric trend estimation methods with automatic bandwidth selection and bootstrap methods for a semi-parametric prediction of different high-frequent financial data. The advantage of this extension is on the one hand the wide field of application of the presented methods and on the other hand the non-necessity of previously determined distribution assumptions. The selected models are the ACD model and its logarithmic extension, which are applied and compared, both parametrically and semiparametrically. In addition, Kalman predictions according to the ARMA model are used and extended semiparametrically, which roughly corresponds to the prediction according to the Semi-LogACD model under the conditional normal distribution assumption. Their empirical application to various financial market data and the evaluation of forecasts according to various criteria for the assessment of point and interval forecasts shows that semiparametric methods are clearly preferable to parametric methods.

Taxonomy of Risk on Metal Market

Dominik Krężołek, Grażyna Trzpiot

The aim of this paper is to work with risk which can be observed when we deal on metal market. Usually definition of risk includes two dimensions: the probability of occurrence and the associated consequences of a set of hazardous scenarios. We try to add next dimension - the source of a risk: level of turnover (volume volatility) and price (return volatility). We can categorize risks along a multidimensional ranking, based on a comparative evaluation of the consequences, probability, and source of a given risk. Another dimension is the chosen risk measures, in the meaning on risk model. The empirical part is connected with metal market.

Data Analysis in Finance 2

Joint Input and Predictive Model Parameters Selection for Financial Forecasting

Iaroslav Shcherbatyi, Wolfgang Maass

With ever growing amounts of data collected by companies in an attempt to gain a competitive edge by leveraging data analytics, a challenge arises to extract useful signal correlated with values of interest from data wide and big. Such challenge is also pertinent to tasks of automated financial forecasting, where historical data is limited in size, and abundance of data can be used as input to forecasting model, hence making input arbitrary "wide". Based on a study of financial forecasting for 10 different commodity futures, we test different approaches for joint selection of relevant historical market data and predictive model parameters. We discuss an algorithm that can efficiently perform such selection. We evaluate such algorithm empirically and show that analysis of inputs selected by the algorithm can provide deeper insights into the relationship between target value and inputs. Our results also indicate that there is no single "best" model for financial forecasting, albeit we find that recurrent neural networks perform best for forecasting of most commodities. Outlined approach and best practices can be directly transferred to other corporate forecasting tasks.

A Box-Cox Semiparametric Multiplicative Error Model

Xuehai Zhang, Yuanhua Feng

A general class of SemiMEM (semiparametric multiplicative error) models is proposed by introducing a scale function into a MEM class model to analyze the non-negative observations, such as the mean duration, the absolute returns, the trading volume and the trading numbers. The estimation of the scale function is not limited by any parametric models specification and the moments condition is also reduced via the Box-Cox transformation. For the purpose, an equivalent scale function is applied in the local linear approach and it can be converted to the scale function by multiple a constant under any weak moment condition. The equivalent scale function estimation and the bandwidth, c_f and the power transformation parameters selection are proposed based on the iterative plug-in (IPI) algorithms. In the power transformation selection, the maximum likelihood estimation (MLE), the normality test (Jarque-Bera test, JB; Shapiro–Wilk test, SW) and the the quantile-quantile regression (QQr) are employed and simulation algorithms for the confidence interval of estimated power transformation parameter by the MLE, SW, JB and QQr are also developed by the block bootstrap method. The algorithms fit the selected real data well.

Model Risk of Selected Systemic Risk Measures for Polish Banking Industry

Katarzyna Kuziak, Krzysztof Piontek

In this paper systemic risk will be meant as a risk of breakdown or major dysfunction in financial system. Some researches use the term to include the potential insolvency of a major player in or component of the financial system. There are many approaches to measure systemic risk (Hansen 2013), but the major distinction includes financial soundness indicators and advanced systemic risk models. Advanced models are based on the statistical multivariate distributions or on the stochastic processes (Benoit, Colliard, Hurlin i Pérignon 2015). The development of these models is observed after the financial crisis of 2007-2008, especially after 2010. In the paper Conditional Value-at-Risk (CoVaR) proposed by Adrian and Brunnermeier (2009, 2011, 2016) and Marginal Expected Shortfall (MES) proposed by Acharya et al. (2010) and Brownlees, Engle (2012) will be used to assess systemic risk. Two methods conditional (DCC-GARCH) and unconditional (quantile regression) will be used for estimating. As a proxy of a financial system Author will propose construction of an index which should be free of disadvantages as opposed to stock market index. Empirical analysis for Polish financial system will contain: model risk analysis, comparison of results for selected measures of systemic risk and estimation methods, evaluation of results (backtesting and financial condition assessment of financial institutions).

Acharya. V.V., Pedersen L.H., Philippon T. and Richardson M.P. (2010) Measuring Systemic Risk. Available at SSRN: http://ssrn.com/abstract=1573171, referred on (10/04/2018).

Adrian T., Brunnermeier M. K. (2009) CoVaR. Federal Reserve Bank of New York Staff Reports. no. 348.

Adrian T., Brunnermeier M. K. (2011) CoVaR. Available at https://www.princeton. edu/ markus/research/papers/CoVaR, referred on (10/04/2018).

Adrian T., Brunnermeier M. K. (2016) CoVaR. American Economic Review, 106, 7 p. 1705-1741. Benoit S., Colliard

J.-E., Hurlin C., Perignon, C. (2015): Where the risks lie: A survey on systemic risk, HEC Paris Research Paper, FIN-2015-1088.

Brownlees, C.T., Engle, R.F. (2012) Volatility, Correlation and Tails for Systemic Risk Measurement. Available at https://bfi.uchicago.edu/sites/default/files/research/SSRN-id1611229.pdf, referred on (10/04/2018).

Hansen L.P. (2013) Challenges in Identifying and Measuring Systemic Risk. Available at SSRN: http://ssrn.com/abstract=2158946 , referred on (10/04/2018).

The Effects of the Regulatory Capital Requirements of Basel III on the Cost of Capital of Banks – an Empirical Analysis.

Florian Naunheim, Matthias Gehrke, Jeffrey Heidemann

As a response to the financial crises beginning in 2007, the Basel Committee on Banking Supervision amended the Basel III framework to increase the overall loss absorbency of the banking sector. This paper contributes to the ongoing discussion on higher capital requirements as it analyses the relation of higher capital requirements to the weighted average costs of capital (WACC) as well as to the cost of equity. Based on a theoretical framework from corporate finance, a linear model is estimated using year-over-year differences for OLS and fixed-effects estimations. The sample is constructed using observation for about 680 banks from 22 jurisdictions covering the years 2003 to 2016 extracted from Thomson Reuters Datastream. The results show a significant, positive relation between the capital ratio and the WACC. However, the effect is reduced when only bank data from the period after the amendment of Basel III were used. If we assume a constant linear relationship between WACC and capital ratios as well as constant surpluses above the regulatory minimum capital requirement our study indicates an increase of around 14 basis points of WACC for the whole sample and an increase of around 69 basis points even for a subsample of the largest banks by reason of the change in capital requirements. While the results for WACC are in line with comparable studies, no significant relation between the capital ratio and the cost of equity using year-over-year differences could be revealed.

Data Analysis in Psychology and Mental Health

Depression Diagnosis using Deep Convolutional Neural Networks

Mofassir ul Islam Arif, Maurício Camargo, Jan Forkel, Guilherme Holdack, Rafael Rêgo Drumond, Nicolas Schilling, Tilman Hensch, Ulrich Hegerl, Lars Schmidt-Thieme

Depression is a prevalent psychiatric disorder that impacts the quality of life of 300 million people around the world. The complex nature of depression manifestations in patients and the lack of technological advances in the diagnosis process has left a lot of room for improvement in this particular domain. At present, the diagnosis is mainly made by physicians during a conversation comprising the exploration of the symptoms and the diagnostic criteria for depression. Recently, the Electroencephalography(EEG) has regained interest as a promising approach to provide bio-markers which are of clinical value in the diagnostic process and for response prediction to therapy In the present landscape, even the addition of EEG data has resulted in a semi-automated manner, where the expert still has to heavily modify the raw data. This adds an inherent bias to the process based on the expert and incurs costs as well as time to the process of diagnosis.

In this paper, we present a fast, effective and automated method that is able to quickly determine if the patient has depression while still maintaining a high accuracy of diagnosis. Our approach is built on using raw EEG-data, performing frequency domain preprocessing in order to split the data into its different frequency domains and to create EEG ’images’. These images are then treated by a convolutional neural network, which is a novel approach in this area. Experimental results have shown to provide outstanding results and to work without the need for feature engineering or any human interaction, which is a core strength of the model we are proposing.

Mental Health: Analytical Focus and Contextualization for Deriving Mental Capital

Fionn Murtagh

The contextualizing of large and complex data sources is crucial (Murtagh and Farid, 2017). Associated with analytical focus can be the addressing of bias in social media and other data sources, and associated with contextualization is Big Data calibration (Murtagh, 2018). The growing importance of Data Science for mental capital has, among the convergence of disciplines, a great deal that is covered in Arezzo (2016) and Qassem et al. (2014). In this work, our main objective is the evaluation of national mental health survey data. While specific findings and outcomes are the major objectives here, relating to definition and properties of mental capital, and a further objective is as follows: to plan with metadata and ontology for further, future and rewarding integration with other data sources, both nationally and globally.

M.F. Arezzo, ”Is social capital associated with health? Evidence from a study on the elderly Italians”, Statistica Applicata, Italian Journal of Applied Statistics, 28(1), 7-23, 2016.

F. Murtagh and M. Farid, ”Contextualizing Geometric Data Analysis and related data analytics: A virtual microscope for Big Data analytics”, Journal of Interdisciplinary Methodologies and Issues in Sciences, vol. 3, 2017. Published article: https://arxiv.org/pdf/1611.09948.pdf

F. Murtagh, ”Data mining and Big Data analytics: Exploiting resolution scale, addressing bias, Having analytical focus”, International Journal of Computer and Software Engineering, 3:127, 2018.

T. Qassem, G. Tadros, P. Moore and F. Zhafa, ”Emerging technologies for monitoring behavioural and psychological symptoms of mental health”, Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (Guangdong, China, November 8-10, 2014). IEEE Computer Society, pp. 308-315, 2014.

Gaussian Process Panel Modeling – Kernel-Based Analysis of Longitudinal Data

Julian Karch, Andreas Brandmaier, Manuel Voelkle

Longitudinal/panel data obtained from multiple individuals measured at multiple time points are crucial for psychological research. To analyze panel data, a variety of modeling approaches such as hierarchical linear modeling or linear structural equation modeling are available. These traditional parametric approaches are restricted to a relatively strong set of assumptions, which are often not met in practice. To address this problem, we present a novel, flexible modeling approach for panel data. Its central feature is the utilization of parametric restrictions on Gaussian processes in the form of parameterized mean and kernel functions for model specification; an idea that has been popularized by the nonlinear regression method Gaussian process regression. Consequently, we term our approach Gaussian Process Panel Modeling (GPPM). While substantially more complex models can be specified with GPPM, traditional inference procedures remain valid for GPPM; such as likelihood-ratio test-based confidence intervals. As a consequence, GPPM subsumes most common modeling approaches for panel data, like linear structural equation modeling and state-space modeling, as special cases within one common model specification language. Additionally, it possesses multiple advantages over the existing methods: It facilitates specification of a much broader set of assumptions, including a large class of nonlinear models, enables continuous-time modeling, entails a natural mechanism for person-specific forecasting, and offers a modular system that enables specifying models by selecting from and combining simple base models.

Linking Data and Psychological Theory with Process-based Models and Bayesian Data Analysis

Alexander Krüger, Jan Tünnermann, Ingrid Scharlau

Data analysis is important for all empirical sciences, but rather different, depending on how data are assessed, analyzed, and understood. In this contribution, we focus on the practice of data analysis in Psychology, especially on often unrecognized assumptions which are an implicit part of the procedures adopted. We furthermore show how process-based models can amend most of these problems.

In Experimental Psychology, data sets are usually highly structured, but alarmingly, approaches to turn them into theoretical advances are often poorly structured. Models provide an epistemically valuable link between data and theory. However, current models in statistical analyses within Psychology are applied mechanistically and are seldom recognized as models. They connect theory and data by providing a decision procedure for a specific yes-no question. Meandering through the vast theoretical possibilities in this piece-meal manner is inefficient. This has led many researchers to reason qualitatively when developing theories, disconnectedly from data and statistics. This facilitates independent, local, and idiosyncratic causal stories that are difficult to integrate into a coherent theory.

We combine Bayesian inferential statistics with mathematical process modeling. Using probabilistic programming, the resulting process-based graphical Bayesian models are combined with MCMC methods. Instead of yes-no decisions, the resulting models describe quantities that are theoretically relevant, predict individual and population behavior, and allow for quantitative comparisons with models of other theories. Challenges we discuss include a higher degree of subjectivity in scientific data analysis and less definite rules for acceptance and rejection of hypotheses.

This approach is not limited to Experimental Psychology. Instead, it provides an interdisciplinary perspective on data analysis as an integral part of the scientific method. With an example from visual attention research, we show how this model-based link between data and theory leads to theoretical progress and new insights.

Probabilistic Time Series Clustering by Vector Autoregressive Metric

Anja Ernst, Casper Albers, Marieke Timmerman

In recent years psychological research has profited increasingly from the inclusion of temporal dynamics into its theories and assessment, for instance with regard to the development of emotions and psychopathology over time and across places. Especially the ideographic approach, focusing on the dynamics of a single individual to fully express individual characteristics, received a lot of attention. These ideographic models usually rely on the vector autoregressive (VAR) model. Often, researchers are interested in generalizing to a population of individuals, rather than being interested in the single individuals per se. As dynamics can be rather heterogeneous across individuals, one needs sophisticated tools to express their essential similarities and differences across individuals. A way to proceed is to identify subgroups of people who are characterized by qualitatively similar dynamics. Recently, dynamic clustering methods have been proposed to discern groups of individuals who exhibit homogeneous development over tyime. So far, these methods assume equal generating processes for individuals of a cluster. To avoid this, I will present a probabilistic clustering approach based on the Gaussian finite mixture model that clusters on individuals' VAR coefficients, thereby allowing for individual deviations within clusters. I will contrast the proposed method to the current gold-standard time series clustering procedure within psychology, on the basis of an extensive simulation study. The performance will be illustrated using an empirical data set from the "How nuts are the Dutch?", a large scale online data platform measuring the mental health of Dutch citizens. The models are applied to ecological momentary assessment data for N = 366 individuals with external measures of depression and anxiety.

Health Shocks and Cognitive Decline in Older Ages

Hendrik Schmitz

Cognitive decline in age is a well-known phenomenon which has attracted increasing attention in the economic literature. Recently, it has been shown that cognitive abilities affect retirement savings and investment decisions and are decisive factors of wealth trajectories pre and post-retirement. While these results highlight the role of cognitive abilities for (financial) wellbeing in older ages, only little is known about potential determinants that speed up or slow down cognitive decline.

In this paper we shed light on one specific but very common and potentially important determinant of cognitive decline in older ages, namely health shocks. More specifically, we study whether the effect of a health shock on cognitive abilities differs by certain characteristics. These characteristics that can be affected, at least potentially, by policy makers include individual and institutional characteristics, such as schooling or health insurance. Better education might not only increase cognitive abilities but also make individuals more robust towards threats against them. Moreover, higher levels of health capital but also institutional factors such as health insurance might make health shocks less severe.

Our analysis is based on data from the Survey of Health Ageing, and Retirement (SHARE), a large representative micro data set covering more than 120.000 individuals above the age of 50 years from 27 European countries and Israel. In addition to a large set of socioeconomic variables, SHARE provides information from several tests that are used as measures of cognitive ability. Our two objective measures for health shocks are based on the information about the onset of a serious illness and the development of grip strength between two waves. The effect of health shocks on cognitive decline is estimated using a first-differences specification that eliminates time-constant unobserved effects that affect both cognitive decline as well as the likelihood to suffer from a health shock. In order to allow for the effect of health shocks on cognitive abilities to depend on e.g. education in a second step, we include interactions between health shocks and potential mediators (e.g. education) in a second step.

First results suggest that all kinds of health shocks adversely affect cognitive abilities. Furthermore, in line with our expectations we find some evidence that individuals with high levels of education might be less prone to adverse effects of health shocks on cognitive functioning, a promising finding in times of population ageing that shows that cognitive decline in older ages is preventable at least to some extent.

The Default Mode Revolution and Computational Psychoanalysis: Implications for Future Health Research.

Miloš Borozan, Rosapia Lauro Grotto

Neuroscientists have long considered that brain activation is mainly connected to the achievement of specific cognitive tasks. However, the serendipitous observation of persistent, coherent activity in rest states, when the brain was thought to be involved in producing unstructured background noise activity, has led to the introduction of the concept of the default mode of brain functioning. This default mode is underpinned by the extensive functional network of brain regions, called the Default Mode Network (DMN) and it is activated when the person is engaged in undirected mental activities such as day-dreaming or mind-wandering. The activity of the DMN decreases when the brain is engaged in conscious, focused attention tasks and it has therefore been interpreted as a neural system in a mutually exclusive dynamics with respect to the conscious system of attention. Since the seminal work of Raichle and collaborators in 2001, there have been more than 3000 papers that explored the rest state of the brain with functional magnetic resonance imaging (fMRI), in connection both with normal function and psychopathology. Within the so called “D.M.N. revolution” a new paradigm in the interpretation of mental disease is emerging: the focus is shifting from the analysis of specific brain areas to the study of the dynamic characteristics of the global, organized/disorganized functional state of the brain. Results within this approach have been already achieved for the disorders of the autistic spectrum, schizophrenia and Alzheimer’s disease. Psychodynamic psychology and psychoanalysis, with their emphasis on the integration of the “Self” as a core aspect of mental function, provide a possible theoretical and clinical framework to read these results. Furthermore, psychoanalytic psychotherapy techniques, such as free associations, appear to follow a logic very similar to the one underlying the DMN function. In order to further explore this connection, specific tools to empirically investigate the domain of psychoanalytic psychotherapy in severe mental health problems should be developed. Here we will discuss the theoretical framework of computational psychoanalysis as a possible formal approach to the modeling of mental function. Computational psychoanalysis is a set of theoretical models and computational tools that was derived in the last two decades in an attempt to formalize the principles of Freudian psychoanalysis and in particular the Bi-Logic model of the Chilean analyst Ignatio Matte Blanco. The need of a unifying formal paradigm in order to promote, orient and constrain data collection and data analysis in the domain of mental health research will be discussed on the base of some examples from recent attempts developed in the framework of computational psychoanalysis.

Data Mining and Knowledge Discovery

Swarm Data Mining for Energy Harvesting in the Boundary Layer of the Atmosphere

Alfred Ultsch

Flights can be largely extended by harvesting potential energy provided by solar induced uplifting effects in the boundary layer of the atmosphere. The main hotspots for such energy provision are thermals and atmospheric gravity waves. Although this harvesting of energy enables an enormous energy efficient extension of flights [1] only few data on the meteorological fine structure of these hotspots are publically available [2, 3]. The area wide introduction of GPS based positioning of aircrafts induced a swarm behaviour in glider flights worldwide. “Food” in form of uplifting energy is sought by the participating gliders while transmitting their position, velocity and vertical speed using the FLARM system [4]. The resulting fight patterns can be analysed for the fine structure of thermals [5, 6]. Current developments in small single-board computers [7] together with the low prices of high precision sensors, originally developed for smart-phones or automotive applications allow the realization of low cost equipment for measuring and data logging of meteorological and flight related parameters. Typical parameters are: ambient temperature, pressure, humidity and 3D accelerations. In addition to the flight pattern analysis this allows the meteorological fine analysis of interesting atmospheric structures. First results of data mining in such data sets in aride climate (Texas, USA) in semi continental climate (Innsbruck, Austria) and in lee waves on the European continent (Pyrenees) are presented.

Adaptation of Boosting Algorithm for Classification in Imbalanced Datasets

Aouatef Mahani, Ahmed Riadh Baba Ali

The problem of extraction of classification rules from imbalanced datasets is a fundamental problem in machine learning, and it has received much attention. A dataset is considered as imbalanced if it contains majority instances and minority ones. This kind of datasets appears in several domains such as fraud detection, medical diagnosis and intrusion detection in networks. It is worth noting that classical rule based classification algorithms have a bias towards the majority classes. In fact, they tend to discover the rules with high values of accuracy and coverage. Consequently, these rules are usually specific to the majority instances, whereas specific rules that predict minority instances are usually ignored or treated as noise. Consequently, minority instances are more often misclassified than those from other classes and classifiers generally perform poorly, because they are designed to minimize the global error rate. The existing approaches for dealing with class imbalance problem are divided into three categories. The first category is considered as data level, it applies sampling methods to balance an imbalanced dataset, this is done in general by deleting majority instances or by duplicating minority ones. However, the second category is at algorithmic level in which the specific algorithms are proposed or some existing algorithms are modified to adapt them to deal with this problem. Final, the third category is at a hybrid level. In this paper, we will propose a based rule approach to extract classification rules that represent majority and minority instances. Our approach applies a modifying boosting algorithm in which we will use a classifier based on genetic algorithm. Since, boosting algorithm alone cannot solve the class imbalance problem. Therefore, we will adapt it by manipulating the weights of instances and the strategy vote for boosting. Also, we will use an appropriate fitness function of genetic algorithm. Our approach will be tested on several imbalanced datasets using different measures of performances and we will compare it with some existing approaches.

Sparsity-Inducing Fuzzy Subspace Clustering

Arthur Guillon, Marie-Jeanne Lesot, Christophe Marsala

Subspace clustering (Agrawal et al., 1998, Vidal 2010) is a data mining task which generalizes clustering by allowing clusters to reside in different subspaces of the original space. These subspaces are not known beforehand and the relevant dimensions of the clusters are thus to be recovered by the algorithm.

This paper studies a fuzzy variant of this task, in which the original features of the data are weighted according to their importance in the subpsaces, as originally proposed by Keller and Klawonn, 2000. However, standard Fuzzy Subspace Clustering (FSC) algorithms may then identify dimensions that are mostly irrelevant to describe the clusters and assign them weights that are small, but non-zero, polluting the description of the subspaces. A solution to this problem is to introduce sparsity in the solutions (Borgelt, 2008): by setting their weights to 0, a FSC algorithm will remove irrelevant features, identifying the intrisinc dimensionality of each cluster. It can be noted that sparsity has also been applied to non-fuzzy subspace clustering (Jing et al. 2007).

This paper proposes an original sparsity-inducing regularization term to weight the features, based on an ℓ_0 regularization term. Although the optimization of such problems is computationally hard in the general case and usually dictates the use of surrogates such as ℓ_1 norm, the settings of this paper allow the use of a polynomial-time algorithm which explores the space of solutions using elementary geometric tools. It is based on the use of recent techniques from convex optimization theory, namely proximal optimization, and relies on the proposition of a new proximal operator.

The paper discusses the proposed cost function, in particular in comparison with state-of-the art subspace clustering mathods with sparsity constraints, and reformulates it in a geometric framework which leads to the proposition of the proximal operator. It gives a theoretical proof of its correctness and derives a new FSC algorithm, named PROSECCO, that optimises the cost function using both proximate and alternate gradient descent. An experimental study comparing PROSECCO to the state-of-the-art in sparse fuzzy subspace clustering, both on artificial and real datasets, shows the relevance of the proposed approach, for recovering the clusters, the dimensionality of their associated subspaces as well as their dimensions.

Process Mining on Machine Event Logs for Profiling Abnormal Behavior and Root Cause Analysis

Jonas Maeyens, Annemie Vorstermans, Mathias Verbeke

Besides an increasing amount of time series sensor data that is collected from industrial machinery, machine events (e.g., statuses, failures, warnings) are also logged. Based on the collected data, the behaviour of machines can be analysed, e.g., for predictive maintenance purposes.

Traditionally, techniques from pattern mining are employed to analyse event log data. Whereas these techniques have proven successful in several domains (including the analysis of machine data), the main downside is that these techniques consider parts of the machine event flow, instead of the process as a whole.

Process mining is a set of techniques in the field of process management that until know have primarily been used to analyse business processes, for example to optimize the enterprise resource planning based on the analysis of transaction logs.

In this research, the feasibility of tailoring and using process mining techniques for the analysis of event data from machine logs is explored. A novel methodology, build on process mining, for profiling abnormal machine behavior will be proposed. Firstly, a process model is constructed from the event logs of the healthy machines. This model can subsequently be used as a benchmark to compare process models of other machines with by means of conformance checking. This comparison results in a set of conformance scores, related to the structure of the model and other more complex aspects such as the differences in duration of particular traces, the time spent in individual events, the relative path frequency, etc. These differences can also be visualised (interactively) on the process model itself, which serves as the basis for root cause analysis.

The proposed approach is evaluated on an industrial dataset from the renewable energy domain, more specifically event logs of a fleet of inverters from several solar plants.

Patient Similarity Analysis for Personalized Health Prediction Models

Araek Tashkandi, Nicole Sarna, Lena Wiese

Advanced analysis of electronic medical records (EMRs) supports medical decision-making, which brings valuable outcomes for personalized medicine. The scoring models such as SAPS and SOFA and the predictive models such as Logistic regression are approaches for analyzing medical data for patient health prediction. For improving the prediction performance, a personalized predictive model is implemented. Patient similarity analysis for personalized medicine helps to find a cohort of patients similar to an index patient. To predict the health of an index patient only the data of the cohort patients are used for the predictive model.

Data explosion and computational burden are produced when calculating pairwise patient similarity analysis of big data sets. For various applications of patient similarity analysis, computational performance should be one of the significant evaluation criteria. Moreover, accuracy is another crucial requirement. Therefore in this paper, we implement patient similarity analysis with predictive models that aim to achieve a balance between analysis performance and accuracy. The efficiency of the analysis is considered by implementing two approaches for patient similarity analysis (in-database analysis vs. data analysis tool). The approach that brings high analysis performance will then be chosen to implement the predictive models for personalized health prediction. In order to gain high prediction accuracy, different predictive models are tested.

Dimension Reduction and Visualisation for Classification 1

Forward Stagewise Linear Regression for Ensemble Methods

Daniel Uys

In supervised learning, the forward stagewise regression algorithm is considered a more constrained version of forward stepwise regression. In its turn, the forward stagewise regression algorithm can be refined to produce the incremental forward stagewise regression model. In the latter model, the idea of slow learning is introduce where the residual vector and the appropriate regression coefficient are updated in very small steps at each iteration. Ensemble methods combine a large number of simpler base learners to form a collective model that can be used for prediction. Learning methods such as Bagging, Random Forests, Boosting, Stacking and Regression Splines amongst others, can all be regarded as ensemble methods. In these methods the linear model is expressed as a linear combination of these simpler base learners, where the coefficients of the base learners are to be estimated by least squares. Since a large number of base learners is typically involved, the residual sum of squares of the linear combination of base learners has to be penalised by, for example, the lasso penalty. However, the large number of base learners also complicates the minimisation of the coefficients in the penalised residual sum of squares criterion. By using the iterative forward stagewise linear regression algorithm for ensemble methods, which includes the idea of slow learning and closely approximate the lasso, estimators of the coefficients of the base learners can be obtained. In the talk the performance of various ensemble methods is evaluated. This is done by applying the forward stagewise linear regression algorithm for ensemble methods to simulated, as well as real life datasets.

Sum Score as Latent Variable for Sparse Multivariate Binary Data

Vartan Choulakian, Jacques Allard

We consider data sets where p binary items, such as true-false questions in an arithmetic test or items in health-related questionnaire, describe similar complementary aspects of an underlying latent unobserved variable. The main question that we ask is: Can we use sum score statistic as latent variable to summarize the multivariate binary data? Our answer to this question is positive by following Benzécri's platform: First we build a table W by Negac coding of the data, then apply taxicab correspondence analysis (TCA) to W. The major result states the following: The person scores of the first dimension of TCA of W is an affine function of the sum score of items calculated on an optimal taxicab associated subset of binary variables. We show the usefulness of the method for sparse multivariate binary data.

Visualising Incomplete Data with Subset Multiple Correspondence Analysis

Johané Nienkemper-Swanepoel, Niël Le Roux, Sugnet Gardner-Lubbe

Subset multiple correspondence analysis (sMCA) enables an isolated analysis of a chosen subset while preserving the scaffolding of the original data set. Multivariate categorical data sets are frequently represented in a coded dummy matrix, referred to as an indicator matrix. Additional category levels can be created for the indicator matrix to account for the unobserved information. Two approaches are employed to handle the incomplete observations: for single active handling a missing category level per variable is created, whereas multiple active handling creates a unique category level per missing variable response per sample. By extending the indicator matrix no observed information is forfeited and separate analyses can be performed on the complete and incomplete data subsets.

sMCA biplots are used for the visual exploration of the subsets. Configurations of the incomplete subsets enable the recognition of non-response patterns which could aid in the selection of suitable imputation methods by determining the particular mechanism of missingness. The missing at random (MAR) missing data mechanism (MDM) refers to missing responses that are dependent on the observed information and is expected to be identified by patterns and groupings occurring in the incomplete sMCA biplot. The missing completely at random (MCAR) MDM states that all observations have the same probability of not being captured which could be identified by a random cloud of points in the incomplete sMCA biplot.

Clustering techniques are explored to identify the MDM in different scenarios. However, most available clustering methods are not suitable for direct application on dimensionally reduced data. A comparison of different clustering approaches will be presented from a simulation study which will be used as a guideline to identify the MDM in a real application.

Model Selection for Projected Divisive Clustering

David Hofmeyr, Nicos Pavlidis

Model selection in unsupervised learning is extremely challenging. The absence of a ground truth precludes general methods like cross validation in their classical forms. The problem of determining model degrees of freedom, used in, among other things, reliable estimation of residual variance for parametric models, has in many cases been inadequately explored. Internal reliability estimates, like cluster stability metrics, although widely applicable tend to rely on computationally expensive bootstrap procedures. All these problems are exacerbated in high dimensional applications, where an appropriate embedding of the data in a lower dimensional space is required. If the stability of the embedding is not accounted for when performing model selection, then the selection procedure will invariably overfit.

This work looks at a recently developed framework for assessing the stability of projection pursuit solutions, through the divergence of the model fit. In the specific case of homoscedastic Gaussian errors, and under suitable smoothness assumptions, the divergence provides an unbiased estimate of the model degrees of freedom [1]. This has been shown to hold for principal component models as well as many other spectral matrix functions [2, 3]. Even in the absence of such assumptions, the divergence provides an intuitive interpretation in the context of model complexity and sensitivity, as a complex model will be highly sensitive to slight variations in the random components of the data. If this is the case the divergence will be large, and hence the degrees of freedom consumed is likely to be high.

Specific attention will be paid to cluster motivated projection indices, and to how this approach can be applied to the problem of model selection in projected divisive clustering.

Unravelling Black Box Machine Learning Technique Predictions using Biplots

Adriaan Rowan, Sugnet Gardner-Lubbe, Francesca Little

Following the development of new mathematical techniques, the improvement of computer processing power and the increased availability of possible explanatory variables, the financial industry is moving toward the use of new machine learning methods, such as neural networks and away from older methods such as generalised linear models, to assist in the decision-making process. i.e. whom to grant loans to.

The goal of talk is to show biplots methods can be used to visualise the various input factors and the output of the machine learning black box. This will assist in quantifying and understanding the black machine learning model by visualising:

- How the model's decision probabilities vary by different variables along the biplot axes and therefore identify the most significant variables influencing the model.
- The difference in the predictions probabilities and therefore the model decision made by linear and non-linear model for different combinations of input values.
- The distance for an individual case to the decision boundary and which variable needs the smallest amount of change to impact the decision made for the case. This will help identify the individual inputs that has the most significant impact on the decision made for the case.
- Find outliers in the training and test data that needs to be further investigated for accuracy and for their impact on the calibration of the model.

Because the results are all visualised in 2 dimensions, they are easy to explain and intuitive to understand. This makes them suitable for decision makers to quickly review the results and justify the decisions made based on the output of the black box model.

Detecting Disease Subtypes by Means of Cluster Independent Component Analysis (C-ICA) of Multi-Subject Brain Data

Jeffrey Durieux, Tom F. Wilderjans

An emerging challenge in the study of brain diseases and mental disorders, like dementia and depression, consists of revealing systematic differences and similarities between subgroups of patients in functional connectivity (FC), that is, coordinated activity across brain regions. As such, existing subtypes of the disease may be characterized in terms of FC and disease subtypes may get detected which transcend the current diagnostic boundaries.

In order to obtain FC, researchers often collect resting-state functional Magnetic Resonance Imaging (rs-fMRI) data and analyze this data with Independent Component Analysis (ICA). ICA is a data reduction technique that decomposes a rs-fMRI dataset into a set of FC patterns and a mixing matrix that contains time courses of activation. The extracted FC patterns can then be visualized as a three-dimensional brain image, indicating regions that show synchronized activity.

Analyzing the brain data of each patient separately with ICA has as major drawback that each patient will be characterized by different FC, which makes it difficult to detect the systematic differences and similarities in FC between (groups of) patients. Therefore, we propose Cluster Independent Component Analysis (C-ICA). The goal of this method is to cluster the patients into homogenous groups based on the similarities and differences in their FC. As such, patients allocated to the same cluster are assumed to have similar FC, whereas patients belonging to different clusters will be described by different FC. This allows a data-driven detection of disease/disorder subtypes based on different FC.

In this presentation, the C-ICA model is proposed, along with an alternating least squares type of algorithm to estimate its parameters. Further, the results of an extensive simulation study to evaluate the performance of the novel C-ICA method are presented. Lastly, the use of C-ICA is illustrated on empirical brain data.

Dimension Reduction and Visualisation for Classification 2

Supervised Feature Selection and Global Sensitivity Analysis

Hana Sulieman, Ayman Alzaatreh

The problem of variable (feature) selection in predictive modelling has received considerable attention during the past 10 years in both statistics and machine learning literatures. The aim of feature selection is to identify the subset of features that provides a reliable and robust model for a given classification or target variable. Feature selection is a key mechanism to reduce curse of dimensionality of feature space.

In this presentation, we discuss supervised feature selection based on global sensitivity information of the target variable to its input features. Two approaches to the assessment of global sensitivity of the target variable are will be presented: Sobol sensitivity measure developed by I.M. Sobol in 1990 and profile-based sensitivity coefficient developed by Sulieman et.al. (2001, 2004).

Sobol sensitivity measure is a variance-based sensitivity technique that decomposes the target variable variance into summands of variances of features in an increasing dimensionality. It evaluates the contribution of each feature and their interactions to the overall variance of the target variable.

Sulieman et.al. (2001, 2004) proposed profile-based nonlinear sensitivity measure based on the total derivative of the target variable with respect to the inputs. While it is inherently local because it is derivative based measure, it provides a broader picture of the sensitivity condition in the presence of parameter co-dependencies and model nonlinearity. The general framework for the empirical computations of Sobol sensitivity indices and the application of profile-based sensitivity measure will be presented. Benchmark data sets used in machine learning literature will be used to demonstrate the implementation of the two global sensitivity measures and their differences will be discussed.

A Biplot based on a Principal Surface

Raeesa Ganey, Sugnet Gardner-Lubbe

Principal surfaces are smooth two-dimensional surfaces that pass through the middle of a pdimensional data set. They minimize the distance from the data points, and provide a non-linear summary of the data. The surfaces are non-parametric and their shape is suggested by the data. The formation of a surface is found using an iterative procedure which starts with a linear summary, typically with a principal component plane. Each successive iteration is a local average of the pdimensional points, where an average is based on a projection of a point onto the surface of the previous iteration. Biplots are considered as extensions of the ordinary scatterplot by providing for more than three variables. When the difference between data points are measured using a Euclidean-embeddable dissimilarity function, observations and the associated variables can be displayed on a non-linear biplot. A non-linear biplot is predictive if information on variables is added in such a way that it allows the values of the variables to be estimated for points in the biplot. Prediction trajectories, which tend to be non-linear are created on the biplot to allow information about variables to be estimated. The goal is to extend the idea of nonlinear biplot methodology onto principal surfaces. The ultimate emphasizing will be on high dimensional data where the nonlinear biplot based on a principal surface will allow for visualization of samples and the predictive variable trajectories. Keywords: Biplots; Principal surfaces; Nonparametric principal components; Multidimensional scaling

The Alpha-Procedure and Aspects of Selection of Classification Space

Tatjana Lange

The paper describes a new methodology (known as Alpha-procedure [Vassilev/Lange]) of constructing algorithms for the detection of laws of similarities, equalities and order, which are hidden in given data, in forming of a decision rule. This methodology corresponds to modern ideas in mathematics, which use, on the one hand, projective transformations and their invariants, and, on the other hand, computer based discrete forms of data processing.

These ideas are connected with a new common definition of the n-dimensional space (by Riemann) as a special geometric space corresponding to a specific law, i.e. kind of transformation, e.g. differential transformation or kernel integral transformation. Along with it the concept of convergence and limit is replaced by the concept of transformation invariants, ensuring both the stability and the comparability of decisions.

The issue will be considered using the example of learning the recognition of two classes A and B by measured data. In this process, the random element may only (but not mandatorily) be the composition of the sample of classes A and B and the composition of the features, which were chosen for the measurements on the objects. The originally measured information can be represented in the form of vectors in a Euclidian vector space, where the vectors portray the objects of classes A and B. The coordinate axes represent the features of the objects (e.g. colour, weight, length, etc.). The coordinates of the points are the measured feature values.

It is required to find (during the learning phase) such a decision rule that will afterwards automatically and stably (faultlessly or at least with only few errors) assign new objects to the class A or B. Together with that, it is not mandatory at all, that the search of the decision rule has to be performed in the Euclidian vector space of the original data, or in their extended/rectifying space, e.g. with the minimax optimization method. The decision rule may also (but not mandatorily) be represented in the same space.

In other words, during the learning process working with a sample of measured object data it is necessary to learn to distinguish stably between two systems A and B, which generate the current data, i.e. it is necessary to distinguish between two kinds of law. Although the two laws are different, they have nevertheless one and the same mathematical nature. e.g. kernel nature, differential nature, statistical nature etc. Every kind of nature (i.e. kind of transformation) will also own its corresponding "geometric wrapping" (geometrics based on projective group of transformations).

The paper shows which kind of transformation (kernel or Lorentz) is more suitable for a certain task of Pattern Recognition and why it is so.

Further, it explains, why the family of Pattern Recognition tasks which are solved well by the kernel transformation can be enlarged by using the regularizations of Tikhonov type or of validation type.

Finally, it shows which tasks are suitable for the Alpha-procedure or for a certain combination of Alpha-procedure and kernel algorithms.

Computing Neural Reliability from EEG Recordings

Pieter Schoonees, Niël Le Roux

Evidence has emerged from the neuroscience literature that the level of inter-subject synchronization between the neural responses of different subjects to, for example, movies is related to population-level measures of movie success, such as box office performance. Measures of such inter-subject similarity are also known as neural reliability measures. The assumption is that the more engaging a naturalistic stimuli such as a movie is, the more similar the responses are even when comparing across subjects. Several studies have shown this empirically, using a variety of methods including correlation-based distance measures and component analysis techniques similar to canonical correlation analysis. In this paper, we discuss these approaches and how to validate them using simulated data.

Dimension Reduction Could be Used to Build Stable Models to Predict Sexual Activity Amongst Incoming First-Year Students

Humphrey Brydon, Retha Luus, Rénette Blignaut, Innocent Karangwa, Joachim Jacobs

Dimension reduction or variable selection is a necessary step in high dimensional data to avoid over-fitting and the construction of overly complex models. The selection of a subset of variables can often lead to reduced computational time as well as more accurate predictive estimates.

The tools used for variable selection varies in the literature (Guyon & Elisseeff, 2003). The selection of a best subset of variables can often be determined through domain knowledge, however, in the absence of this knowledge formal selection methods need to be undertaken.

This study reports on quantitative data collected at the University of the Western Cape over a period of three-years from 2013 to 2015. All students attending the first-year orientation sessions and those who were willing to complete the anonymous questionnaire were included in the study. The questionnaire measured the following aspects: sexual activity prior to entering university, age at first sexual encounter, number of sexual partners, condom usage, perception of HIV risk and HIV testing. Alcohol and drug use and smoking habits were also recorded. Variables measuring depressive symptoms and a scale of religiosity were included. The percentage of students reported having been sexually active prior to entering university was 51.82% compared to 48.18% reported not being sexually active.

The aim of this study is to compare the ability of Classification and Regression Trees (CART), Chi-Squared Automatic Interaction Detection (CHAID) trees and Stochastic Gradient Boosted Trees (SGBT) in identifying a subset of variables for predictive modelling. The variables identified through the CHAID, CART and SGBT procedures are then incorporated into a logistic regression modelling procedure to predict the sexual activity of incoming first-year university students. Initial results indicate that a naïve CHAID modelling procedure is more accurate at classifying sexually active students with a misclassification rate of 27.89% as compared to the CART and SGBT procedures.

The variables selected for the modelling procedures could assist in the development of prevention efforts to target incoming first-year students who are not yet sexually active. These results could assist in the development of HIV intervention programmes to curb the spread of HIV and other sexually transmitted infections amongst university students.

Classification Based on Dissimilarities Towards Prototypes

Beibei Yuan, Willem Heiser, Mark de Rooij

We introduce the delta-machine, a statistical learning tool for classification based on dissimilarities or distances. In the first step dissimilarities between profiles of the objects and a set of selected exemplars or prototypes in the predictor space are computed. In the second step, these dissimilarities take the role as predictors in a logistic regression to build classification rules. This procedure leads to nonlinear classification boundaries in the original predictor space.

In this presentation we discuss the delta-machine with mixed nominal, ordinal, and numerical predictor variables. Two dissimilarity measures are distinguished: the Euclidean distance and the Gower measure. The first is a general distance measure, while the second is a tailored dissimilarity measure for mixed type of variables. Using simulation studies we compared the performance of the two dissimilarity measures in the delta-machine using three types of artificial data. The simulation studies showed that overall the Euclidean distance and the Gower measure had similar performances in terms of the accuracy, but in some conditions the Euclidean distance outperformed the Gower measure.

Furthermore, we will show the classification performance of the delta-machine in comparison to three other classification methods on an empirical example and discussed the results in details. The empirical example is the Statlog heart data from the UCI database. The aim of the study is to accurately predict heart disease by using 13 predictor variables which describe 270 patients. From the 13 predictor variables; six are numerical, one is ordinal, three are binary, and three are nominal. For this empirical example, we will show a variable importance plot and partial dependence plots to reveal the relationship between the original predictors and the response variable.

Dimensionality Reduction

Comparing Two Brand Switching Matrices by Asymmetric Multidimensional Scaling

Akinori Okada, Hiroyuki Tsurumi

A brand switching matrix a square table which usually represents the number of consumers who changed brands they purchased in two consecutive periods in a group of consumers. The (j,k) element of the brand switching matrix show the number of consumers who purchased brand j in the first period and band k in the second period. The main diagonal element of the brand switching matrix shows the numbers of consumers who did not change the brand they purchased in the first and the second periods. The band switching matrix is inherently asymmetric, because the number of consumers who purchased brand j in the first period and brand k in the second period is not necessarily equal to the number of consumers who purchased brand k in the first period and brand j in the second period. When we regard the (j,k) element of the brand switching matrix as the similarity from brand j to brand k, the brand switching matrix is an asymmetric similarity matrix. To analyze (or not ignore) the asymmetry of the band switching matrix, the asymmetric multidimensional scaling and the asymmetric cluster analysis which can deal with one-mode two-way asymmetric similarities have been utilized. But when we have a set of brand switching matrices where each comes from different time points, or different conditions, different areas…, we have two-mode three-way asymmetric similarities. As a first step of analyzing a set of brand switching matrices (or two-mode three-way asymmetric similarities), a procedure of comparing two brand switching matrices by asymmetric multidimensional scaling based on the singular value decomposition is introduced An application of the procedure to real data is shown. The analysis shows which brand switching has increased or decreased graphically. The relationships between consumers and the brand switching are shown as well.

Investigating Quality measurements of projections for the Evaluation of Distance and Density-based Structures of High-Dimensional Data

Michael Christoph Thrun, Alfred Ultsch

Projections are conventional methods of dimensionality reduction for information visualization used to transform high-dimensional data into low dimensional space [1]. If the output space is restricted in the projection method to two dimensions, the result is a scatter plot. The goal of this scatter plot is a visualization of distance and density-based structures. As stated by the Johnson–Lindenstrauss lemma [2], the two-dimensional similarities in the scatter plot cannot coercively represent high-dimensional distances. Projections of several datasets with distance and density-based structures showed a misleading interpretation of the underlying structures [3]. Nonetheless, a scatter plot generated by a projection method remains a state of the art approach in cluster analysis to visualize data structures (e.g. [4-6]). Consequently, the evaluation of projections remains essential and there exist many quality criteria for an evaluation of the projection methods and their results [7]. Here, 19 quality measurements (QM) are grouped, with the aid of graph theory, into semantic classes. It is shown that QMs fail to evaluate the projections of simple data structures of benchmark datasets, called FCPS [8], using common methods such as Principal component analysis (PCA), Sammon mapping (MDS variant) or t-distributed stochastic neighbor embedding (t-SNE). Insights into graph theory indicate that QMs require prior assumptions about the underlying structures [3]. If not, an objective function could be defined using the best QM, in which case it would always be possible to obtain an optimal unsupervised two-dimensional visualization of structures by optimizing this objective function. In this work, new QM called Delaunay classification error (DCE) makes an unbiased evaluation of structures on FCPS: first, DCE utilizes the information provided by a prior classification to assess pro-jected structures. Second, DCE applies the insights drawn from graph theory. The DCE is available on CRAN in the R package “DatabionicSwarm”.

1. Venna, J., et al., Information retrieval perspective to nonlinear dimensionality reduction for data visualization. The Journal of Machine Learning Research. 11: p. 451-490, 2010.

2. Dasgupta, S. and A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures & Algorithms, 22(1): p. 60-65, 2003.

3. Thrun, M.C., Projection Based Clustering through Self-Organization and Swarm Intelligence, Doctoral dissertation, Heidelberg: Springer, 2018.

4. Handbook of cluster analysis. Handbook of Modern Statistical Methods, ed. C. Hennig, et al., New York, USA: Chapman&Hall/CRC Press, 2015.

5. Ritter, G., Robust cluster analysis and variable selection. Monographs on Statistics and Applied Probability, Passau, Germany: Chapman&Hall/CRC Press, 2014.

6. Mirkin, B.G., Clustering: a data recovery approach. Computer Science and Data Analysis Series, ed. J. Lafferty, et al., Boca Raton, FL, USA: Chapnman&Hall/CRC, 2005.

7. Gracia, A., et al., A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality. Information Sciences, 270: p. 1-27, 2014.

8. Ultsch, A. Clustering with SOM: U* C. in Proceedings of the 5th Workshop on Self-Organizing Maps, 2005.

Knowledge Discovery from Low-Frequency Stream Nitrate Concentrations: Hydrology and Biology Contributions

Michael Christoph Thrun, Lutz Breuer, Alfred Ultsch

Human activities modify the global nitrogen cycle, mainly through farming. These practices have unintended consequences; for example, nitrate lost from terrestrial runoff to streams and estuaries can impact aquatic life [Aubert et al., 2016]. A greater understanding of water quality variations can improve the evaluation of the state of water bodies and lead to better recommendations for appropriate and efficient management practices [Cirmo/McDonnell, 1997]. Here a high dimensional data set containing measurements on related chemical water quality indicators, i.e., nitrate and electrical conductivity, is analyzed. The dataset contained in total 32,196 data points for 14 different variables. In contrast to the high-frequency temporal analysis (with 15-minute intervals) of [Aubert et al., 2016], in this work the daily courses for each variable were calculated as the sums of all daily measurements, resulting in a low-frequency analysis. The databionic swarm (DBS) was used for projection and clustering of the data [Thrun, 2018]. This resulted in five clusters with small intracluster distances and large intercluster distances. The rules extracted from the CART decision tree were applied to the clustering [Breiman et al., 1984]. The classes suggest that the hydrological data can be explained by two variables related to biological processes and two variables related to hydrological processes. In particular temperature influences the activities of living organisms, such as soil microbial organisms [Zak et al., 1999]. Soil moisture determines microbial activities, such as long-term inactivity in dry soil followed by wetting [Borken/Matzner, 2009]. The groundwater level (or head, in m) is the primary factor driving discharge in a catchment [Orlowski et al., 2014]. Rainfall intensity triggers discharge and affects soil moisture as well as leaching of nutrients [Orlowski et al., 2014].

[Aubert et al., 2016] Aubert, A. H., Thrun, M. C., Breuer, L., & Ultsch, A.: Knowledge discovery from high-frequency stream nitrate concentrations: hydrology and biology contributions, Scientific reports, Nature, Vol. 6(31536) , 2016.

[Borken/Matzner, 2009] Borken, W., & Matzner, E.: Reappraisal of drying and wetting effects on C and N mineralization and fluxes in soils, Global Change Biology, Vol. 15(4), pp. 808-824. 2009.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A.: Classification and regression trees, CRC press, 1984.

[Cirmo/McDonnell, 1997] Cirmo, C. P., & McDonnell, J. J.: Linking the hydrologic and biogeochemical controls of nitrogen transport in near-stream zones of temperate-forested catchments: a review, Journal of Hydrology, Vol. 199(1), pp. 88-120. 1997.

[Orlowski et al., 2014] Orlowski, N., Lauer, F., Kraft, P., Frede, H.-G., & Breuer, L.: Linking spatial patterns of groundwater table dynamics and streamflow generation processes in a small developed catchment, Water, Vol. 6(10), pp. 3085-3117. 2014.

[Thrun, 2018] Thrun, M. C.: Projection Based Clustering through Self-Organization and Swarm Intelligence, Ultsch, A. & Hüllermeier, E. Doctoral dissertation advisors, Heidelberg, Springer, ISBN: 978-3658205393, 2018.

[Zak et al., 1999] Zak, D. R., Holmes, W. E., MacDonald, N. W., & Pregitzer, K. S.: Soil temperature, matric potential, and the kinetics of microbial respiration and nitrogen mineralization, Soil Science Society of America Journal, Vol. 63(3), pp. 575-584. 1999.

On Extracting Asymmetric Changes in Asymmetric Matrices

Tadashi Imaizumi

In our real world, we find many asymmetric relationships between two objects, for example, replacing car data, world trade data between nations, etc. And many models and methods have been proposed to analyze these data. Okada and Imaizumi (1997) proposed a non-metric asymmetric multidimensional scaling model and method for two-mode three-way similarity matrices (object× object× source) . The model consists of a common object configuration and two kinds of weights, and is applicable to many types of data matrices. However, we can extract some valuable characteristics of data matrix if we knew some constraints on similarity matrices and could assume a proper model for the data collected. For example, we want to analyze several trading data between nations of every 5 years from 1960 years, and this data are two-mode three-way similarity matrices (object× object× time points). It will be natural to assume that some time trends of these trading data between successive two years are. So, we proposed the constrained asymmetric multidimensional scaling to extract symmetric trends and asymmetric trends in data. The common configuration and associated weights represent symmetric parts of the model, and radius of each object and associated weights represent asymmetric parts of the model. We focus only a linear trend or a quadratic trend of time points in this paper. And two types of weights are also represented by those trends. Simulatio Study and Application to real data will be shown.

A. Okada and T. Imaizumi, Asymmetric multidimensional scaling of two-mode three-way proximities, Journal of Classification, 1997 Vol.14(2) pp 195-224.

Multivariate Gaussian Feature Selection

Helene Dörksen, Volker Lohweg

We concentrate our research activities on the multivariate feature selection, which is one important part of many machine learning tasks. In particular, Linear Discriminant Analysis (LDA) belongs to the state-of-the-art methods for the multivariate analysis. From the theoretical point of view, it is well-known fact that LDA is best suitable in the case the features are Gaussian distributed. In the theoretical part of the presented paper, we analyse the properties of the multivariate discriminant analysis with respect to the feature selection. In this context, we consider a binary supervised learning task and, in addition, assume that the features are independent Gaussian distributed. The discriminant analysis solves the mentioned supervised learning task by maximising of the discriminant value, which is calculated for the linear combination of the features. Maximising the discriminant corresponds to the searching for the such direction, which maximises the projected class means while minimising the classes variance in this direction. In our investigations, the initial LDA solution is considered for all given features. The discriminant value is calculated by the classical formula based on the projected class means and projected class variances. We proof several propositions with the aim to find subsets of the features having higher discriminant values as the original. For the suitability in the real world settings, e.g. in the industrial environments, here we are mostly interested in the fast searching for such subsets. The performance of the mentioned propositions is examined experimentally on datasets from UCI repository. In the paper, diverse application scenarios will be discussed and tested on the datasets. For example, scenarios for feature selection in the classical sense as well as combinations of feature subsets with or without dimensionality reduction will be illustrated. We present results of the cross-validation tests, which indicate that applying our feature selection procedure often leads to the increasing of the generalisation ability of classification. As the test results show, the performance can be often achieved also in the case the features are not Gaussian distributed. By this fact and by the low complexity of the computational time, we expect our feature selection procedure to be suitable in the real world environments.

Choosing Among Notions of Depth for Multivariate Data

Karl Mosler, Pavlo Mozharovskyi

Many statistical tasks involve measures of centrality, identifying the center of a data cloud and measuring how close a given data point is to the center. In a probabilistic setting, we are interested in the question how central a point is with regard to a probability distribution. Classical multivariate statistics measures outlyingness by the Mahalanobis distance in an affine invariant way. Since the early 1990's more general, nonparametric depth statistics have been developed for measuring centrality and outlyingness of data in d-space. A depth function is a function which, given a point and a distribution in d-space, yields a number between 0 and 1, while satisfying certain postulates. Accordingly, numerous notions of depth have been proposed in the literature, some of which are also robust against outlying data. The departure from classical Mahalanobis distance does not come without cost. There is a trade-off between invariance, robustness and computational feasibility. In the last few years, efficient exact algorithms as well as approximate ones have been made available in R-packages like ddalpha, DepthProc, and many others implementing specific depths and their applications. Consequently, the choice of a depth statistic is not any more restricted to one or two notions due to computational limits, but rather often to more notions among which we have to decide. We discuss aspects and general principles of this choice. The speed of exact algorithms is compared, the limitations of popular approximate approaches like the random Tukey depth are demonstrated, and guidelines are provided for the construction of depth-based statistical procedures as well as for practical applications when several notions of depth appear to be computationally feasible.

EuADS Symposium on Data Science Education 1

The Project “ExWoSt Digitale Lernlabore”: Smart Data Labs as a Method of Data Science Education

Katharina Schüller, Katrin Grimm

In the age of digitization data seem to be ubiquitous: We produce them with our mobile phones, with every movement on the Internet, via sensors or camera shots. What happens with these data and which risks and uncertainties are associated with data-based decisions is, however, hidden to most of us. Thus, statistical thinking and data literacy are more important competencies than ever and there is no doubt that it is important to start early when teaching data literacy. Data science education does not stop after graduation but needs to continue in adult life. But there is lack of good examples how data science and statistical literacy can be taught by means of real-life problems.

This contribution presents the project “ExWoSt Digitale Lernlabore” which investigates the role of so called smart data labs in promoting data literacy in German towns, i.e. in their municipalities but also in their civil society, for example in youth centers, libraries, or other public places.

Smart Data Labs are characterized by working on real-life problems with agile methods, professional support, and ongoing exchange and feedback. That is, a smart data lab provides a sheltered space free of expectations regarding the results. Data experts provide support where necessary. The key aspect in planning a smart data lab is to find a statistical problem which is suitable for building competences regarding data preparation, analysis and visualization and interpretation of statistical results but relates to the participants' daily life and at the same time.

The project “ExWoSt Digitale Lernlabore” explores the implementation of Smart Data Labs in an urban context. In doing so the data labs are subject to accompanying evaluation in order to identify factors of success and barriers with respect to three dimensions of success: solution for the problem at hand, competence improvement, and promoting citizens' participation and networking. Results could be helpful in identifying approaches to planning Smart Data Labs and choosing suitable problems for Smart Data Labs so that they can also be transferred as a model for data science education at schools.

Statistical Computing and Data Science in Introductory Statistics

Karsten Luebke, Matthias Gehrke, Norman Markgraf

In the last years there is a movement towards simulation-based inference (e.g. bootstrapping and randomisation tests) in order to improve students' understanding of statistical reasoning (see e.g. Chance et al., 2016) as well as a call to introduce statistical computing (e.g. Nolan & Temple Lang, 2010) and reproducible analysis (e.g. Baumer et al., 2014) within the curriculum. With help of R mosaic (Pruim et al., 2017) and the concept of Minimal R we were able to include all this in a Introductory Statistics course for people studying while working a business related major. Moreover, this also paves the road toward methods and concepts like Data Wrangling or Algorithmic Modeling, more related to Data Science than to classical statistics, a shift proposed by e.g. Cobb (2015) and Horton et al. (2015).

These courses are offered for different majors in different study centers across Germany. There, the acceptance of this shift towards more data literacy is evaluated and analysed by a survey among our heterogeneous faculty staff.

Baumer, Ben, et al. "R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics." Technology Innovations in Statistics Education 8.1 (2014).

Chance, Beth, Jimmy Wong, and Nathan Tintle. "Student performance in curricula centered on simulation-based inference: A preliminary report." Journal of Statistics Education 24.3 (2016): 114-126.

Cobb, George. "Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up." The American Statistician 69.4 (2015): 266-282.

Horton, Nicholas J., Benjamin S. Baumer, and Hadley Wickham. "Teaching precursors to data science in introductory and second courses in statistics." arXiv preprint arXiv:1401.3269 (2014).

Nolan, Deborah, and Duncan Temple Lang. "Computing in the statistics curricula." The American Statistician 64.2 (2010): 97-107.

Pruim, Randall, Daniel T. Kaplan, and Nicholas J. Horton. "The mosaic Package: Helping Students to 'Think with Data' Using R." R Journal 9.1 (2017).

From Computer Science and Statistics towards Data Science at LMU Munich

Thomas Seidl, Göran Kauermann

Big Data is certainly one of the buzzwords of the last five years. With the digital revolution nearly everything can today be measured, recorded or protocolled. The amount of data exceeds multiples of peta bytes with an annual increase, which has been unbelievable even a couple of years ago. But the pure data flood is without great use if no information is drawn from the data. This step is the real challenge for the next decades. While some early views in the gold rush times of Big Data were even postulating that Big Data calls for the end of theory, since with enough data every question can be answered, it becomes more and more apparent, that the step from Big Data to relevant information is full of obstacles and traps and demands for novel scientific routes.

This view is not new, but has been formulated by Cleveland (2001) more than 1 1/2 decades ago. He proposed to combine statistical approaches with computer science and labelled this as Data Science. In his article "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" he criticises the existing separation between statistics and computer science and writes: „The benefit to the data analyst has been limited, because the knowledge among computer scientists about how to (...) approach the analysis of data is limited, just as the knowledge of computing environments by statisticians is limited. A merger of the knowledge bases would produce a powerful force for innovation." This merger of the two disciplines is what is today understood as Data Science and the current Big Data challenges make Cleveland's proposal vivid and indispensable.

At Ludwig-Maximilians-University Munich a new master program in Data Science has been launched in 2016 (www.datascience-munich.de). The initiative is funded by the Elitenetwork Bavaria (www.elitenetzwerk.bayern.de). The program is jointly run by the Department of Statistics and the Department of Computer Science at the LMU and curriculum is, as proposed by Cleveland, a real merge of statistics and computer science. The talk will underpin the benefits of the collaboration exhibit a curriculum for Data Science which combines statistics and computer science, following Cleveland's proposal (see Kauermann & Seidl, 2018).

W.S. Cleveland (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Review, 69:21-26.

G. Kauermann, T. Seidl (2018). Data Science – A proposal for a Curriculum. International Journal of Data Science and Analytics. https://doi.org/10.1007/s41060-018-0113-2

Data Science and Big Data in Upper Secondary Schools: What Should Be Discussed from a Statistics Perspective?

Rolf Biehler, Daniel Frischemeier, Susanne Podworny, Thomas Wassong

In the present days and due to the increasing availability of Big Data, Data Science is becoming more and more relevant in our daily lives. That is why it is important to address Data Science already in school. The domain of Data Science is a large field, combining statistics, computer science and sociocultural issues. It has to be discussed which topics and which contents can and should be implemented in school, e.g. from the perspective of statistics.

Within the frame of a Design-based Research project, computer science educators and statistics educators at Paderborn University will design a pilot course on the subject of Data Science and Big Data. It will address upper secondary students and will be realized by weekly sessions (three hours) over seven months in the upcoming school year 2018/2019. The whole course which is meant to introduce upper secondary school students in the field of Data Science consists of four modules. In module 1, the learners are introduced into the basics of statistics and big data and it aims at developing their data competence and data awareness. The second module will introduce learners into machine learning and programming based on examples from module 1. In the third and fourth module learners can apply their knowledge gained in modules 1 and 2 and will work in small groups on real and meaningful Data Science projects. In this presentation, we want to concentrate on the statistics components, especially of module 1, and we will present how we develop the data competence and data awareness of upper secondary school students to prepare them to work on Data science projects in modules 3 and 4.

EuADS Symposium on Data Science Education 2

Digitally Fit?: Applied Machine Learning Academy for Industry

Ralph Ewerth, Marc Dittrich, Wolfgang Nejdl, Claudia Niederee, Hendrik Noske, Jan-Hendrik Zab, Sergej Zerr

The digitization and the increasing inter-connectivity of the industry are fundamentally changing the efficiency and complexity of technical systems and operations in intelligent manufacturing. Industry 4.0 allows companies to improve business processes, optimize facility utilization, increase flexibility, detect anomalies and predict failures.

The core topic in the field of intelligent systems is machine learning. However, recent studies show that majority of German companies, and many worldwide, do not yet use state of the art solutions that research is offering. In particular, small and medium-sized enterprises (SMEs) are often rejecting digitalization for fear of investments, complex integration of new tools and due to the lack of suitably educated IT staff. The lack of skilled employees is seen as one of the biggest barriers for integrating appropriate solutions. Therefore, their further qualification is necessary, which builds on their professional competencies and experience, making them fit for digitization.

In this work, we will present initial results achieved within the project “Applied Machine Learning Academy (AMA)” - a qualification program with research activities carried out by Leibniz University of Hannover that focus on Machine Learning competences within the Industry 4.0 environment in Germany. The goals of the program are on one hand, further qualification of employees in SMEs and industry in the area of Machine Learning and Artificial Intelligence and, on the other hand, the development of Machine Learning solutions for SMEs along with industrial partners in the form of small research projects.

In our work we (a) report the current situation in the industrial area based on a questionnaire results from 500 industrial companies in Germany (mainly in the area of manufacturing), (b) we will evaluate approaches within our courses to transmit theoretical ML principles in the way that build on employees present professional knowledge and skills and (c) present the status of running research projects on the topics of “Anomaly Detection” and “Predictive Maintenance” in industrial manufacturing processes that are currently carried out between the University and Industry.

Data Science and Big Data in Upper Secondary Schools: What Should Be Discussed from a Perspective of Computing Education?

Birte Heinemann, Lea Budde, Carsten Schulte

In the present days and due to the increasing availability of Big Data, Data Science is becoming more and more relevant in our daily lives. That is why it is important to address Data Science already in school. The domain of Data Science is a large field, combining statistics, computer science and sociocultural issues. It has to be discussed which topics and which contents can and should be implemented in school, e.g. from the perspective of computing education.

Within the frame of a Design-based Research project a pilot course will be designed by computing and statistics educators at the University of Paderborn, addressing upper secondary students. The whole course consists of four modules. In module 1, the learners are introduced into the basics of statistics and big data and it aims at developing their data competence and data awareness. The second module will introduce learners into machine learning and programming. In the third and fourth module learners can apply and deepen their knowledge by active and self-directed work on concrete Data Science projects.

In this session, we concentrate on module 2, which shows how Data Science is transforming computing, by highlighting the change from classical algorithmic problem solving to data-driven processes, using the example of machine learning in a tangible way. This includes to familiarize students with the associated role of humans in interaction with data driven artifacts. The students should learn basic knowledge in the field of data-driven problem-solving technology, get an understanding of neural networks, deep learning and machine learning. We believe this requires developing an adequate picture of hybrid systems – a kind of data-driven, emergent ecosystem which needs to be made explicit to understand the transformative role as well as the technological basics of these tools from artificial intelligence and how they are related to Data Science. Some individual phases of the module are presented in this session. Broad Experiences, challenges, stumbling blocks and opportunities of these elements are discussed and reflected with regard to the topic and objectives of Data Science.

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

Nadja Bauer, Malte Jastrow, Daniel Horn, Lukas Stankiewicz, Kristian Kersting, Jochen Deuse, Claus Weihs

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. Data administration skills are demanded to handle big data in diverse storage structures in data bases. 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 newly acquired project ”Industrial Data Science” (InDaS) we aim to develop and implement a qualification concept for machine learning in industrial production based on the explicit demands coming up in industrial environment. The resulting qualification program is targeted at two different groups. On the one hand advanced students from any of the three fields mentioned above and on the other hand experienced professionals working in industrial production. For the first group a one term lecture is going to be designed while coping with different levels of knowledge in the inhomogeneous audience. It will be followed by a seminar with focus on use cases delivered by partners from industrial production. Separately a workshop concept for the second target group will be developed taking into account the strong domain knowledge of the participants.

The contents of the qualification concept should be selected according to the needs of industrial companies. Therefore a survey was created to inquire the use and potentials of machine learning and the requirements for future employees in industrial production. The evaluation of the survey and the resulting conclusions affecting the qualification concept are going to be presented in this talk.

Image and Music Data Analysis

Enhancing Flood Risk Analysis using Interactive Retrieval of Social Media Images

Björn Barz, Bin Yang, Kai Schröter, Moritz Münch, Andrea Unger, Doris Dransch, Joachim Denzler

The analysis of natural disasters such as floods often suffers from limited data due to sensor failures or a coarse distribution of sensors. At the same time, valuable complementary information could be extracted from images of the event posted by volunteers on social media platforms such as Twitter and Instagram, so-called "Volunteered Geographic Information (VGI)". However, inspecting the entire stream of images posted online during the time period in question and in the relevant area is infeasible due to the huge amount of images posted in a very short time, most of them being irrelevant for the analysis to be performed.

In this work, we propose to use a computer vision approach for retrieving relevant images of the event to be analyzed based on the image content only. The analyst starts by providing a suitable query image and the system then returns a ranked list of similar social media images. The user can provide relevance feedback by flagging certain images as relevant or irrelevant to obtain a refined list of results. This content-based image retrieval (CBIR) approach is very flexible and can adapt to different search objectives easily: Sometimes, the analysts want to determine inundation depth, where images of partially flooded traffic signs can be helpful, and sometimes they want to assess the grade of water pollution, for which different images are relevant.

We introduce a new dataset of about 3,500 images of the European Flood 2013 collected from Wikimedia Commons and annotated by domain experts regarding their relevance with respect to three tasks. We compare several image features and metric learning methods for relevance feedback on that dataset, mixed with 100,000 distractor images, and are able to improve the precision among the top 100 retrieval results from 50% with the baseline retrieval to 82% after 5 rounds of feedback.

Handwritten Formula Recognition with Pixel-Wise Generative Adversarial Networks

Matthias Springstein, Clemens Pollak, Ralph Ewerth

Recently, researchers have combined different deep learning techniques to achieve new advances in the field of machine learning and computer vision. However, the training of deep neural networks typically requires a large amount of data to adapt to a certain task. If training data are not freely available or cannot be generated, it has to be annotated manually, which is a very time-consuming and expensive task. This also applies to the problem of handwritten formula recognition, where it is easy to create printed formulas, but it is very difficult to create handwritten training data.

In this work, we suggest to use a generative adversarial network for data augmentation to change the appearance of synthetically generated equations. This step allows us to create new examples for a deep learning recognition system and to train models that adapt better to unseen equations.

Our model relies on a pixel-level domain transfer model that can work with images of arbitrary size. Then, we use a model to classify the domain of the source image and two recognition systems to solve the problem of handwritten and synthetic equation classification simultaneously. This network structure is able to apply the properties of handwritten formula images to rendered formula images. The recognition model uses an attention mechanism to change the reception field during the classification process depending on the symbols already perceived.

Experimental results show that these models allow us to create a system that can recognize synthetic as well as handwritten ones.

Visual Stylometry of Comics Using CNN Features

Jochen Laubrock, David Dubray

Stylometry including authorship attribution has been successfully applied to text, with application ranging from literary studies to forensics (e.g., Juola, 2006). Stylometry of visual material is not yet as well-established, despite recent advances (Saleh & Elgammal, 2015; Gatys et al., 2015). Developments in machine learning and availability of large corpora of annotated images may change the scene. Rather than using hand-crafted engineered features, discovery of features relevant for classification can be automated by training deep convolutional neural networks (CNNs; LeCun et al., 2015; LeCun, 1989), inspired by processing in the human visual system.

Here we propose a method for a visual stylometry based on CNN features. Models and parameters of CNN-based winning entries in large-scale photograph classification challenges like ImageNet are freely available and can be adapted to new material by re-training just a few layers, assuming that basic features at the lower levels are more or less generic. Since CNNs learned filters may represent invariants of our environment, we hypothesized that they might also be useful for the analysis of drawings, which, as abstractions, are related to our environmental reality. We tested transfer learning to classify about 50,000 digitized pages from about 200 comics of the Graphic narrative corpus (GNC, Dunst et al., 2017) by their illustrators. We trained SVM and fully-connected neural network classifiers to classify drawing style, given the visual features coded in each of the main (mixed) layers of ImageNet pre-trained Inception V3 (Szegedy et al., 2015).

Overall, the top-1 test-set classification accuracy in the artist attribution analysis increased from 92% for mixed-layer 0 to over 95% when adding mixed-layers higher in the hierarchy. Above mixed-layer 5, there were signs of overfitting. Performance of NN and SVM classifiers was on a par. An in-depth analysis and the corresponding visualizations of which features are the most discriminative and most strongly associated with a given artist are underway. The few misclassifications were instructive. For example, a page of “Batman: The Long Halloween” drawn by Tim Sale was mistakenly attributed to David Mazzuchelli, the artist responsible for “Batman: Year One”; Tim Sale got only the second highest vote.

Our analysis illustrates that stylistic classification of comics artists is generally possible using CNN features pre-trained on photographs and given only a limited amount of additional training material. Texture-like mid-level vision features are sufficient for fairly good performance. Given this successful transfer, we suspect that CNN features are general enough to be applied in a wide variety of other domains in which a visual stylometry may be useful. For example, art historians may be interested in combining the method with nearest neighbor search to describe how close different artist are in feature space.

Measurement of Robustness of Features and Classification Models on Degraded Data Sets in Music Classification

Igor Vatolkin

There exists a large number of supervised music classification tasks: recognition of music genres and emotions, playing instruments, harmonic and melodic properties, temporal and rhythmic characteristics, etc. In recent years, many studies were published in that field, which are either focused on complex feature engineering or application and adjustment of various classification methods. However, less work is done on the evaluation of model robustness, and music data sets are often limited to music with some common characteristics, so that the question about the generalisation ability of proposed models usually remains unanswered. In this study, we examine and compare the classification performance of audio features and classification models when applied for recognition of different music categories on music data sets which were degraded by means of techniques available in the Audio Degradation Toolbox (Mauch and Ewert, 2013) like attenuation, compression, filtering, or adding of noise.

M. Mauch, S. Ewert (2013). The Audio Degradation Toolbox and Its Application to Robustness Evaluation. Proc. ISMIR, 14th International Society for Music Information Retrieval Conference, pp. 83-88.

Classifying Music Genres Using Image Classification Neural Networks

Alan Kai Hassen, Hilko Hermann Janßen, Dennis Assenmacher, Mike Preuß

Since the rise of online music streaming platforms, music genre classifcation, as part of music information retrieval, has attracted researcher’s attention in the field of computer science.

Lately, domain tailored Convolutional Neural Networks (CNN) have been applied using spectrograms as visual audio representation, outperforming traditional classification methods that utilized hand-crafted audio features (Costa et al., 2017). Up to this point, it is open to debate, whether domain tailored CNN architectures (Costa et al., 2017; Zhang et al., 2016) are superior to network architectures used in the field of image classification. The question arises, since image classification architectures have highly influenced the design of domain tailored network architectures in, for example, the adaptation of parts from ResNet (He et al., 2016) by Zhang et al. (2016). This development is in line with the reasoning for using CNNs for audio analysis tasks which assumes that the network detects auditory events by identifying their time-frequency representations (Choi et al., 2016). We expect that the genre classification task using spectrograms as audio representation, resembles an image classification task. Therefore, we examine, whether CNN architectures, applied in image classification are able to achieve similar performance compared to domain tailored CNN architectures used in genre classification. Since the design of domain tailored networks is a time extensive process, skipping the neural networks’ design phase and using state of the art image classification networks instead, is of great value.

Within our work, we are comparing domain tailored and image classification networks by testing their performance on two different datasets after hypertuning the networks' respective parameters. The datasets are the most frequently used scientific dataset GTZAN (Tzanetakis and Cook, 2002) and a newly created dataset with ten times more songs mirroring the genre structure of GTZAN - the SoundCloud dataset. The SoundCloud dataset is created since the GTZAN dataset comprises only 1000 songs and it has been shown that a low number of available data instances is a limiting factor for the generalization capabilities of neural network architectures (Srivastava et al., 2014).

Choi, K., Fazekas, G., Sandler, M. (2016). Automatic Tagging Using Deep Convolutional Neural Networks. Proc. 17th International Society for Music Information Retrieval Conference, ISMIR 2016, pp. 805–811.

Costa, Y. M. G., Oliveira, L. S., Silla, C. N. (2017). An Evaluation of Convolutional Neural Networks for Music Classification Using Spectrograms. Applied Soft Computing, 52(C): 28–38.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), pp. 770–778.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(1):1929–1958.

Tzanetakis, G., Cook, P. (2002). Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing, 10(5):293–302.

Zhang, W., Lei, W., Xu, X., Xing, X. (2016). Improved Music Genre Classification with Convolutional Neural Networks. Proc. 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), pp. 3304–3308.

Multi-Objective Optimization of Tone Onset Detection and Pitch Estimation Algorithms

Kerstin Wintersohl, Nadja Bauer, Daniel Horn, Claus Weihs

Automatic music transcription is one of the most challenging signal processing applications. There exist a lot of works where single music transcription components like algorithms for tone onset detection or pitch estimation are improved or optimized. However, for such a complex problem the simultaneous optimization of all relevant components is essential. This talk - as a first step in this direction - presents a model-based multi-objective optimization approach for simultaneous tuning of an onset detection and a pitch estimation algorithm. For this purpose, 16 parameters are optimized belonging either to onset detection or pitch estimation, while two of them (window size and hop size) are the same for the both algorithms. The optimization is performed on five different sound sequences that exist in two velocity versions. Each of these ten sound sequences is played in the following six musical instruments: clarinet, flute, guitar, piano, trumpet and violin yielding a total of 60 sound sequences. The goal is to find the pareto set of the best parameter settings. In order to be able to compare the results of the different parameter settings, the goodness of both detection methods has to be measured. For onset detection the F-measure is calculated that is the harmonic average of precision and recall. For pitch detection an accuracy value is computed. The result of the optimization is the pareto front consisting of three parameter settings leading to a good performance of both onset detection and pitch estimation algorithms.

Industrial Applications

Data and Model Management Architecture for the Steel Industry

David Arnu, Fabian Temme, Edwin Yaqub, Gabriel Fricout, Marcus Neuer

Steel manufacturing is a highly complex process. Slight deviations of the process conditions can lead to the occurrence of metallurgical defects in the final product. Thus ensuring and improving the quality remains a major challenge faced by the steel industry. The EU project PRESED (Predictive Sensor Data mining for Product Quality Improvement) focuses on the usage of machine learning algorithms to predict the quality of the product during early processing steps. We aim at training on the complete production data instead of using aggregated values, which has been the approach in the majority of prior art. The variety and veracity of the production data requires the development of a sophisticated data model. It has to comprise time series of sensor data and process condition values. The change of properties of the product over the different production steps adds up to the complexity of the data model. Both the data and the predictive models have to be managed in a manner, which supports non-rigid schematic representation of data. We present a data and model management architecture which is developed within the PRESED project. Using a NoSQL database, we are able to mirror the diversity of the described data model. The data structure can map the different production steps as independent nodes and is not limited to a rigid schema. It is also possible to store raw and transformed production data in a common hierarchy. As such it is possible to retrieve the best data representation fitting the required usage. In addition, we have designed a predictive model representation layer for organizing trained models. This allows for an easy to use evaluation of different methods. This is important to detect drift of concepts during the production process and retrain the models if needed. To operationalize the predictive models, we are using browser-based dashboards as an intuitive interface. The hierarchical data structure transfers seamlessly into such a graphical representation. The objective of the architecture is to support the complete life-cycle of a predictive analytics project by including a smart data layer and a prediction model layer for model management. With the help of this architecture we are able to manage a complete processing pipeline. Starting from the raw production data we can select the desired transformation and cleaning steps, train feature extraction and prediction models and apply them on new batches of incoming data. The visual guidance of these steps encourages an on side deployment for non data-scientist personal. On top we could significantly increase the prediction quality compared to a reference model.

Unsupervised Learning Approach to Assign Error Types to Automobile Engine Failures

Shailesh Tripathi, Sonja Strasser, Lukas Schimpelsberger, Matthias Dehmer

Vibration data provides useful information about the functioning of different components (e.g. crankshaft, camshaft) of an automobile engine. This data allows us to develop predictive/classification models for the prediction of failure-types in the engine based on the pattern of the noise. However, engine failure can happen due to the faults in multiple components in various combinations. In such cases of failures, it is difficult to label the failure type correctly. Other than that, not labeling data correctly or pseudo-failures are other problems where error types are not identified correctly. The predictive modeling for the identification of error-types is not a straightforward approach when the failures are not identified correctly, or the data is misclassified. In our analysis, we provide an unsupervised learning approach to identify different groups (clusters) of failures which are unknown, not labeled correctly and misclassified by utilizing features from vibration data. Further, we perform enrichment analysis to label these groups with possible error types by using the known-error type information from the vibration data. The unsupervised learning approach of labeling unknown and misclassified data to describe error-types is a useful approach to develop predictive/classification models accurately and to improve their prediction power.

Applications of Machine Learning and Predictive Analytics in the Automotive Industry

Ralf Klinkenberg

This paper provides an overview of application use cases of machine learning and predictive analytics in the automotive industry. The author has more than 25 years of experience in machine learning and its applications in various industries. This paper covers application use cases encountered in the automotive industry in the last 20 years in order to show the breadth and diversity of these applications with the goal to inspire companies in the automotive sector and in other industries, where and how they can leverage machine learning to create value from their data. Customer- and business-oriented as well as product- and manufacturing-oriented use cases, their challenges, and solution approaches are discussed. The spectrum ranges from standard use cases like predictive maintenance, i.e. machine failure prediction and prevention, failure root cause analysis, and demand forecasting to more complex tasks like automatically predicting assembly plans for new product designs for e.g. engines and their components as well as to web mining and social media monitoring to generate market and customer insight as well as competitive intelligence. Applications leverage machine learning, classification, regression, clustering, time series analysis and forecasting, text analytics, audio mining, image mining, web mining, and social media monitoring. For each application use case, the paper describes the task, its challenges, and a high level description of the algorithmic solution.

Development of a Data-Driven Software Tool to Detect Optimal Electrode Sheets in Lithium-Ion Battery Production

Oliver Meyer, Claus Weihs, Sarah Schnackenberg, Michael Kirchhof

Rechargeable Lithium-Ion Battery Cell Production is one of the most important processes in the field of electro mobility. The most important parts of these battery cells are the positive and negative electrodes that allow for storage and release of electronic charge. Electrodes are produced in the form of long coated foils which are then cut into pieces of a predefined length called electrode sheets and used to build single batteries. As a member of the battery research cluster ProZell, we are focusing on the optimization of the production process of such coated foils at the Center for Solar Energy and Hydrogen Research Baden-Württemberg (ZSW). The production process of the coated foils consists of several sequential process steps (e.g. coating and drying) and quality parameters are measured frequently along the foil after each sub-process. This means that for each of the smaller sheets several values of the same quality parameter are measured. For example, if a 5 m sheet of material is needed for one battery cell, a quality parameter might be measured every 10 cm, resulting in about 50 values for each sheet. We have implemented a software tool to monitor the production over all relevant sub-processes based on the values of those quality parameters. We aim at determining the maximum number of battery cells that can be built from a produced foil of a certain length, with respect to given quality requirements. In a second step, by treating the task as a knapsack problem and using dynamic programming techniques, our software tool is able to determine the optimal positions of the sheets to be cut out based on any of the observed quality parameters.

Automated Prediction of Assembly Plans for New 3D Product Designs

Ralf Klinkenberg

In many industries like for example the automotive industry, the competition is fierce and customers expect ever more variations of products and new products in ever faster cycles. In the typical process, product designers deliver a 3-dimensional design for each new product or product varation, for which an assembly planer then creates the according assembly plan for the manual parts of the product assembly. Variations in design can lead to less or more complicated assembly steps and hence shorter or longer assembly times, i.e. lower or higher assembly costs and hence overall product costs. Hence, nowadays, product designers and assembly planers often have to iterate the process to refine the design to reduce the assembly time and cost. This iterative process is time-intensive and costly.

In this paper, we propose an approach to automatically predict the assembly time for new product designs and variations and to also automatically predict assembly plans for new product designs. This allows the designer to quickly check already during design time, how costly each design option is in terms of assembly time and cost, and hence to decide for favourable design options, before even consulting the assembly planer. This also enables the assembly planer to not start with an empty assembly plan, but with candidate assembly plans automatically generated by the system. This significantly accelerates the product design and assembly planning process and hence makes the company more agile by reducing the time from product idea to go-to-market. This also reduces the design and assembly planing costs for new products and product variations and hence makes this process not only faster but also more cost-effective and hence profitable even for smaller product serieses. This paper describes the machine learning approaches chosen for predicting assembly times and assembly plans as well as results from an application in the automotive industry for predicting assembly plans for new truck engine component designs. The proposed approach leverages experience from text mining and other data mining applications in its choice of the data representation and its choice of algorithms. Any approach leveraging for example sub-graph matching to determine the similarity between hierarchical product component designs would be NP-complete and hence not scalable to large data sets and hence infeasable in practice. Hence we chose a vector-based data representation inspired by text mining and we use standard machine learning algorithms to make the problem computationally feasible and scalable to large data sets. We decompose the overall task of assembly plan prediction into sub-tasks, for which we can leverage and combine standard machine learning algorithms for clustering and classification. This paper describes the chosen data representation, the chosen algorithmic approach, and the results obtained in the truck engine component assembly prediction use case.

Interpretable Machine Learning 1

Cognitive Bias vs Inductive Bias: Use of Cognitive Heuristics in the Design of Machine Learning Algorithms

Tomas Kliegr

Cognitive bias is a term used by cognitive scientists and psychologists to characterize rules that humans apply during decision making [Mata, 2012], while inductive bias [Mitchell, 1997] is a key concept used in machine learning research to characterize the set of assumptions a learning algorithm makes. Both cognitive biases and inductive biases have certain scope, set of problems, for which they are suitable - ecologically valid - and for other problems they result in errors. There are thus some correspondences between the two concepts, but there are also marked differences. For general artificial intelligence or machine learning problems, creating versatile yet strong enough inductive bias runs into the limits imposed by the No free lunch theorems - all optimization algorithms are equally good if all problems are considered [Wolpert, 1996]. How could machine learning benefit from the empirical results obtained in cognitive science? As follows, effective design of inductive bias depends on the analysis of the distribution of types of tasks that will be handled by the learning algorithm. In case of highly structured and narrow domains, existing methodology may be sufficient. However, for general artificial intelligence or machine learning problems, creating versatile yet strong enough inductive bias is a challenge. Analysis of cognitive biases and heuristics that the evolution has coded into the human mind provides a conceivable starting point for such an endeavour.

In the context of model interpretability, inductive biases have also been already discussed in relation to human cognition by Griffiths et al. [2010] in the sense of using representations used by machine learning to explain human reasoning.

As a particular example of cognitive-bias inspired machine learning algorithm, we will review the Take-The-Best [Gigerenzer and Goldstein, 1996], which is based on the take-the-best heuristic assumed to be used by humans. This heuristic works as follows: when multiple dimensions are relevant to a judgment, the dimensions are tried one at a time according to their cue validity, and a decision is made that favours the first dimension that discriminates the alternatives. The TTB algorithm adopts the bounded rationality model [Simon, 1956, 1982] stating that information processing systems typically need to satisfice rather than optimize.

The discussion will be concluded with a review of possible implications of better understanding of the relation between cognitive and inductive biases for interpretability of machine learning models. Already, Griffiths et al. [2010] discusses the relation between inductive biases and cognitive science suggesting that the knowledge representations used in machine learning, such as rules or trees, can be useful for explaining human inferences.

Langley [1996, p. 384], Brighton [2006] make the general observation that machine learning suffers from lack of cognitive-inspired influences. To the best of our knowledge, our work is the first to specifically point at cognitive biases as the source of inspiration for machine learning with the justification that these reflect the probability of occurrence of individual types of problems in the physical world, alleviating the limitations imposed by the No free lunch theorem.

Rule Editor for Cognitive Experiments: Towards Better Understanding of Rule Interpretability and Comprehension

Stanislav Vojir, Patrik Kopecky, Tomas Kliegr

The presented rule editor is a web-based software under development that will allow to perform cognitive experiments with rules either directly discovered from data, or manually input by the experimenter. The interface presents a list of rules and asks the user to perform some editing action to achieve stated objective. The typical task will consist of the following editing actions: delete rule, remove attribute, add attribute, reorder rules. The system can be configured to provide different levels of editing operations and it also supports the possibility to evaluate the accuracy of the rule list on a pre-uploaded test datasets. The system will record changes to rules performed by the user. The software will implement the principles of randomized controlled trials. A prototypical experiment run with the software can look as follows. One group of users will be asked to improve understandability of the rule list. The second group of users will be asked to balance understandability of the rule list with classification accuracy. The rule editor can be used to empirically test hypotheses related to interpretability such as semantic coherence (Gabriel, 2014). According to this hypotheses, when faced with a rule such as “area > 6720, population > 607430, latitude < 44.1281 -> Unemployment = low”, the users will tend to remove the literal “latitude < 44.1281” because of its low relation to the other attributes. The rule editor is developed on top of the EasyMiner module for business rules (Vojir et al, 2014), for which a screencast is available at goo.gl/fMLLWo.

Interpretable Classification of Facial Expressions of Pain

Michael Siebers, Ute Schmid, Dominik Seuss, Jens Garbas, Teena Hassan, Miriam Kunz, Stefan Lautenbacher

We present the PainFaceComprehender – a prototype system for classification of facial expressions of pain. In contrast to end-to-end machine learning approaches, our approach is white-box, relying on action units (AUs) as intermediate and human understandable representation. Data are provided in form of videos gained from experimentally induced episodes of pain and other mental states. AUs and their intensities are extracted from the video recordings by a novel approach fusing probability estimates from geometric and appearance classifiers. The classifiers are learned on an extensive training data set. The AUs and their time stamps are input for the

PainFaceComprehender. We use the inductive logic programming (ILP) approaches Metagol and Aleph to learn classification rules to discriminate pain from other mental states. Learning in ILP can exploit various types of background knowledge. We provide a background theory of events where we specify temporal relations and a theory of AUs which we allocate with different regions of the face. Furthermore, background knowledge about the individuals, such as gender, age, or illness can be taken into account. The learned rules provide comprehensible information about which aspects of facial expressions are relevant to identify pain. Furthermore, we investigate explanation generation. The rule which is applied to classify a current image can be explained verbally as well as by back reference to the image. Verbal explanations can be shallow – referring only to the predicates in the body – or deep by expanding them to additional rules or facts. Currently we are exploring how invented predicates can be assigned meaningful names by user interaction. As a further step, we plan to build a training system for nurses on top of the explanations.

Interpretable Instance Based Text Classification for Social Science Research Projects - An Evaluation

Helena Löfström, Tuwe Löfström, Ulf Johansson

It is common within Social Science research projects to manually classify documents using human coders. The cost in time and money is considerable, taking resources from the time spent on researching, but it is seen as the best and most accurate way. Automatic classification could decrease the cost of classification, making it possible to use a larger part of the funds to research. Many text classifiers are behaving like black boxes. This could make users unwilling to use text classification, in spite of good results. Martens and Foster show that it is very important to explain the predictions of individual documents for the users and thereby making them more inclined to use text classification. Identifying and referring difficult documents to humans supervising the process together with the predictions and an explanation is often necessary. This study addresses the need for manual text classification within research projects. The data used is a small subset of a large corpora of manually classified news articles concerning politics or social issues. The documents vary a lot in size, ranging from 27 words to 1446 words. The paper applies an algorithm inspired by the algorithm presented by researchers from Astra Zeneca on top of an underlying predictive model (random forest) to get both predictions and explanations for the documents classified. The explanations are presented as a list of the words most important for the prediction. Leave-one-out was used during experimentation. Six articles were selected and presented to 3 researchers familiar with the used corpora. The six articles were selected to represent clearly correct, clearly incorrect and ambiguously predicted documents from each class, to cover different kinds of situations. The respondents were presented with the article text (with important words highlighted), the prediction (and the probability estimates) and the list of top explanatory words and their weights. They were asked to reflect on the prediction in relation to the text, if the most important words had been selected and if they agreed with the prediction. The manual classification was not revealed. The respondents used the explanation words in their analysis of the documents, either to question or agreeing with the predictions. The clearly correctly predicted documents (representing the vast majority of the predictions) were also correctly classified by all three respondents, even if one respondent had some questions about less important words being highlighted. The incorrectly predicted documents got mixed feedback and were considered as hard to manually classify due to length and wandering topics. The uncertain documents had several important words highlighted, but they were ranked low and also contradicting each other. Additionally a web survey (with 24 respondents) of the explanations aggregated over all documents were also conducted to evaluate the overall performance. To conclude, the evaluated solution provided the respondents with the most important reasons why the documents were predicted as they were. The web survey indicated that on an aggregated level, the top ranked words from each class was generally identified as representative of that class.

How Interpretable Are You?: A Framework for Quantifying Interpretability

Amit Dhurandhar

We provide a novel notion of what it means to be interpretable, looking past the usual association with human understanding. Our key insight is that interpretability is not an absolute concept and so we define it relative to a target model, which may or may not be a human. We define a framework that allows for comparing interpretable procedures by linking it to important practical aspects such as accuracy and robustness. We characterize many of the current state-of-the-art interpretable methods in our framework portraying its general applicability. Finally, principled interpretable strategies are proposed and empirically evaluated on synthetic data, as well as on the largest public olfaction dataset that was made recently available. We also experiment on MNIST with a simple target model and different oracle models of varying complexity. This leads to the insight that the improvement in the target model is not only a function of the oracle models performance, but also its relative complexity with respect to the target model.

Interpretable Machine Learning 2

Explanation Methods in Deep Neural Networks: An Overview

Gabrielle Ras, Marcel van Gerven, Pim Haselager

Issues regarding explainable AI involve four components: users, laws & regulations, explanations and algorithms. Together these components provide a context in which explanation methods can be evaluated regarding their adequacy. In this talk we set out to analyze the question ”What can be explained?” given these four components. The goal of this session is to give an overall view on the current state of explainability in DNNs. Different kinds of users are identified and their concerns revealed, relevant statements from the General Data Protection Regulation are analyzed in the context of Deep Neural Networks (DNNs), a taxonomy for the classification of existing explanation methods is introduced, and finally, the various classes of explanation methods are analyzed to verify if user concerns are justified. Overall, it is clear that (visual) explanations can be given about various aspects of the influence of the input on the output. However, it is noted that explanation methods or interfaces for lay users are missing and we speculate which criteria these methods / interfaces should satisfy. Finally it is noted that two important concerns are difficult to address with explanation methods: the concern about bias in datasets that leads to biased DNNs, as well as the suspicion about unfair outcomes.

A Manifold Perspective for Debugging and Interpreting Deep Learning Models

Ning Xie, Derek Doran

Owing to the overwhelming success of deep learning, the diagnosis and explanation of opaque machine learning models has arisen great interest in both industry and academia. Many researchers focus on exploring and visualizing the features learnt by an opaque model, in order to build “explanations” of its actions. Rather than focusing on learned features, we propose a novel method for model debugging and interpretation that leverages the training dataset and the manifold of the network activation space. Based on the network activations for a test example and all training dataset, a robust distance metric is designed to find the closest training examples (referred as “training supports”) for the test example in the approximated intrinsic network activation space. Such “training supports” for a test example can be used as a new resource to help explain model decisions. An explanation of how the model makes a specific decision can be developed by examining similarities between the features of the “training supports” and the test example.

The “training supports” of a test example can also shed light on new strategies for training data debugging and model improvement. For example, for each wrongly predicted test example, examination of the labels and the model predictions for its “training supports” can be used to decide if a “training support” is improperly labeled. When the labels of most “training supports” are different from the label of the test example, one may conclude that the reason of incorrect predictions may be due to: (i) poor performance on “training supports”, which may be symptomatic of model underfitting in this region of the data space. This can be verified by comparing the labels and predictions of “training supports”; (ii) an insufficient number of examples in the training dataset that are “similar” to a test example in the network activation space. This indicates an opportunity to improve the performance of the model by collecting or generating (via a GAN, for example) additional training examples with the same label as the test example. Such additional examples should teach the network the proper activation patterns so that the test example can be classified correctly by the model.

Rule Extraction from a Convolutional Neural Network in Sentiment Analysis

Guido Bologna

A natural way to explain classification responses provided by a Multi Layer Perceptron (MLP) is by producing propositional rules that represent the knowledge embedded within it. Since the eighties many rule extraction techniques have been proposed in this context, especially for shallow architectures. Convolutional Neural Networks (CNNs) can be regarded as a sophistication of ”classical” MLPs. At the functional level, the main difference is the greater number of hidden layers typically used with several particular operators, such as convolution and ”Max-Pooling”. Currently, rule extraction from CNNs is an unexplored research domain. To fill this gap, we propose to use a transparent MLP placed after the Max-Pool layer. For a sentiment analysis problem, we use before the MLP network a CNN architecture having a unique convolutional layer followed by a Max-Pool layer. Rule extraction is performed after the Max-Pool layer with the use of the Discretized Interpretable Multi Layer Perceptron (DIMLP). This special MLP architecture allows us to generate symbolic rules by precisely locating axis-parallel hyperplanes. The antecedents of the extracted rules represent responses of convolutional filters, which are difficult to understand, in general. However, we show that from these ”meaningless” values it is possible to obtain rules that make sense. Specifically, with a dataset on tweets of movie reviews, the inputs of the CNN are represented by word embedding vectors. As a result, from the values of the Max-Pool layer represented in the extracted rules' antecedents, it is possible to determine relevant words taken into account in the decision process. In the experiments we illustrate several examples of rules with their characteristics in terms of accuracy, fidelity and complexity. Overall, the presented approach represents a first step toward rule extraction from deeper CNN architectures used for instance in image recognition problems.

Rule Extraction with Guarantees

Ulf Johansson, Henrik Boström, Cecilia Sönströd

The main advantage of black-box rule extraction algorithms is their generality, specifically they are agnostic to the underlying opaque model. There is, however, one important problem associated with black-box rule extraction; even if the algorithm aims for maximizing fidelity in the learning phase, there is no guarantee that the extracted model will be faithful to the opaque model on test data. Instead, extracted models may overfit or underfit the training data, which stands in sharp contrast to open-box methods, where the rules should have perfect fidelity, even on the test set.

Ideally, we would like to have the efficiency and generality of black-box rule extractors, while guaranteeing test set fidelity. The purpose of this paper is to show how conformal prediction can be utilized for achieving this.

Conformal predictors output prediction regions, i.e., a label set in classification and an interval in regression. All conformal predictors are valid under the standard i.i.d assumption, i.e., the probability of making an error is bounded by a predetermined confidence level. A prediction error, in this framework, occurs when the correct value is not included in the prediction region.

In the empirical investigation, we show how conformal prediction can be used to guarantee either the fidelity or the accuracy of extracted interpretable models.

Finding an appropriate representation language for extracting regression models is not trivial. Regression trees are typically too weak, while more complex alternatives, like model trees, are not truly comprehensible. In this study, we suggest using standard regression trees, but modified as conformal predictors, i.e., each leaf will contain a prediction interval. We argue that the result is a very informative model, especially when combined with the provided guarantees for accuracy or fidelity.

Finally, we demonstrate how so-called Mondrian conformal prediction can be used to provide guarantees even for individual rules.

Understanding Learned Models by Identifying Important Feature Frontiers

Mark Craven, Kyubin Lee, Sid Kiblawi, Akshay Sood

In many application domains, the learned models that provide the best predictive accuracy are complex and difficult to interpret. Such models include neural networks, random forests and other ensembles. In some domains, it is critical to be able to provide descriptions of how such models make their decisions across the distribution of instances, as opposed to providing explanations for individual instances. One common and general approach to this task is to calculate measures of feature importance, and to summarize the decisions of a model in terms of which features are most important.

We have been developing methods that addresses several key challenges that arise in this type of approach: (1) the feature space may be very large, and there may be many features that have some degree of importance in the learned model, (2) there may be strong dependencies and known relationships among the features, (3) interactions among features may be important to the model's decisions, (4) it would be valuable to assess the statistical significance of feature importance values, while controlling the false discovery rate. Our approach attempts to characterize a learned model at the “right” resolution by finding a frontier in a hierarchy of features such that each node in the frontier has a statistically significant effect on the model's accuracy, this condition does not hold for children of the given node, and the false discovery rate across the frontier is controlled.

We have evaluated our approach in two challenging application domains in which it is important to characterize learned models – predicting patients' risk for adverse events given their past clinical histories, and modeling associations between viral genotypes and the disease phenotypes they induce. In these domains, the most accurate models we have learned are random forests and LSTM neural networks. In both domains, we have gained novel insights by analyzing the learned models using our methods.

Explaining Random Forest Predictions using Frequent Itemset Mining

Henrik Boström, Ram Gurung, Tony Lindgren, Ulf Johansson

Random forests have frequently been shown to achieve state-of-the-art predictive performance. This however comes with a cost that is shared with many other high-performing techniques; the logic behind their predictions cannot be easily understood. In order to provide some insights into random forests, techniques for estimating variable importance have been proposed. These highlight what features have the highest impact, but do not explain how they affect the predictions. Rule extraction aims to explain opaque models through interpretable approximations. Such approaches are commonly divided into black-box approaches, which only consider the input-output of the underlying model, and open-box approaches, which also exploit the structure of the underlying model. When applied to random forests, the former have mainly focused on approximating forests using single trees, while the latter approaches derive sets, or linear combinations, of rules extracted from the forests. As an alternative to approximating the underlying models, we here instead investigate means of understanding the predictions of random forests through approaches that aggregate information from predictions of the individual trees. In particular, we consider approaches that represent paths from roots to leaves in the trees as itemsets. This enables the use of frequent itemset mining and association rule discovery techniques to analyze both individual predictions as well as sets of predictions, e.g., finding association rules of sufficient support and confidence to explain predictions relative to a certain class. This idea has to the best of our knowledge not been explored in the literature, although the very similar idea of using association rules to summarize forests was earlier proposed, but not investigated, in an unpublished manuscript. It should however be noted that summarizing rules in a forest and explaining predictions lead to different types of insight. Results from applying the proposed techniques are presented and discussed. The potential utility of the techniques is discussed, as well as their limitations and possible extensions.

Machine Learning 1

A Comparison of Automatic Algorithms for Occupation Coding

Malte Schierholz

Occupation coding refers to the assignment of respondents' textual answers from surveys into an official occupational classification. It is time-consuming and expensive if done manually and, as a remedy, several algorithms have been suggested to automate this process. To overcome deficiencies of existent techniques and to provide probabilistic predictions, we introduce yet another method that combines training data from previous studies and job titles from a coding index. Using data from various German surveys, we compare our new method with some of the main algorithms described in the literature, including regularized logistic regression, gradient boosting, nearest neighbors, memory-based reasoning, and string similarity. Strengths and weaknesses of each algorithm are discussed.

Stopping Criteria for Active Learning with a Robot

Marek Herde, Adrian Calma, Daniel Kottke, Bernhard Sick, Maarten Bieshaar

The employment of robots plays a central role when improving the productivity in industry. The advancement of robot applications results in the automatization of previously manually executed processes.

Recently, we developed a method that enables a robot to learn a sorting rule taught by human. The robot actively points out objects and asks a human for the corresponding label (i.e., class affiliation). By selecting the objects, the robot learns actively, i.e., it is allowed to be curious and to select the data from which it learns. This learning paradigm is called active learning and aims at deriving high-performance models with minimal costs. In our case, the cost is a result of interaction with the human. Consequently, we aim at minimizing the number of interaction and, thus, the number of acquired labels. Still, the current method relies on the human to decide when the learning process should be stopped.

Our goal is to let the robot decide by its own when to terminate the learning process. We investigate techniques for assessing (estimating) the learning progress of the robot. Our survey of current research regarding stopping criteria for active learning shows that state of the art methods require user-defined thresholds. Moreover, these methods evaluate the current performance of the learning model, not the change in performance. The two aforementioned considerations offer valuable opportunities for improvement.

Hence, we emphasize the need for research on a new, self-contained stopping criterion. It has to assess the change in model performance and, based on this assessment, to stop further label acquisitions. Still, two goals have to be reached: minimal number of queries and high performance. That is, the learning should stop as early as possible without any loss in classification performance.

As the robots learns to sort a set of objects, the performance of the robot is evaluated on this set (transductive active learning). We propose a technique based on classification uncertainty. It estimates the label uncertainty of the remaining unlabeled objects and investigates the changes in uncertainty as proxy of the learning improvement.

We evaluate the performance of the stopping criteria on a Baxter robot. The integration of these stopping criteria into the training process of a robot is a further step towards full-autonomous learning.

Evaluating Ordinal Classifiers on Repetitive Class Structures

Lisa Schäfer, Hans A. Kestler, Ludwig Lausser

Ordinal classifiers are based on the assumption that a predefined (total) order of class labels is reflected by the embedding of the samples in feature space. This order constrains the layout of the decision regions and guides the classifier's training procedure making the classifier highly susceptible to wrong assumptions. Incorrect class orders will lead to decreased class-wise sensitivities. Ordinal classifiers can therefore be used to reject non-reflected class orders and propose a small set of candidates [1].

However, these candidates are not guaranteed to reflect real ordinal embeddings. In this work, we give examples for varying repetitive (non-ordinal) structures that mimic the characteristics of ordinal ones. Ordinal classifiers cascades applied to both types of embeddings can end up with comparable class-wise sensitivities and candidate sets.

We provide cross-validation results for various controlled artificial datasets and ordinal classifier cascades build of different base classifiers. The repetitive structures are not rejected by cascades of classifiers with disconnected decision regions. These classifiers propose unique class orders in all experiments.

[1] Lattke R, Lausser L, Müssel C, Kestler HA. Detecting ordinal class structures. In Schwenker F, Roli F, Kittler J, (eds), Multiple Classifier Systems (MCS 2015), volume LNCS 9132, pp. 100-111. Springer, 2015.

Leela Chess Zero: A Crowd-Sourced Effort to Replicate and Improve AlphaZero

Folkert Huizinga, Karlson Pfannschmidt

Progress in artificial intelligence in recent years has been driven by supervised learning of data produced by humans. This severely limits the applicability to domains where human supervision is lacking or unfeasible. Furthermore, this human supervision could implicitly limit the performance of the final system. Computer chess is dominated by alpha-beta search engines which employ handcrafted evaluation functions and aggressive search heuristics to achieve state-of-the-art performance. Silver et al. (2017) demonstrate with AlphaZero that it is possible to improve by replacing the evaluation function with a deep neural network trained on millions of self-play games through reinforcement learning. To explore the search tree efficiently, Monte Carlo tree search is employed.

The development of AlphaZero marks an important step in the field of artificial intelligence. For the scientific community to benefit, it is necessary to verify the design decisions which went into the approach and to replicate the results. Replication here is difficult because Google used custom tensor processing units with significantly more floating point operations than publicly available.

We present our crowd-sourced and open source effort (http://lczero.org) for replicating the approach. Volunteers from all over the world can choose to contribute computational power in the form of self-play games. At the time of writing around 500 contributors have produced 5 million self-play games. Architectural and methodological questions are solved collaboratively. Since the game generation needs to be carefully balanced with the training of new networks, we employ techniques like loss landscape visualization and Bayesian optimization to adapt neural network architecture and hyperparameters. We present further insights on how to improve the exploration/exploitation trade-off in the Monte Carlo tree search by employing first-play urgency. The collaborative, distributed nature of this project allows ideas to quickly be discussed, rigorously evaluated and implemented, which results in a rapid increase of knowledge gained during the process.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T. P., Simonyan, K., Hassabis, D.: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. CoRR abs/1712.01815 (2017)

Jouppi, N. P., Young, C., Patil, N., Patterson, D. A., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., Ghaemmaghami, T. V., Gottipati, R., Gulland, W., Hagmann, R., Ho, C. R., Hogberg, D., Hu, J., Hundt, R., Hurt, D., Ibarz, J., Jaffey, A., Jaworski, A., Kaplan, A., Khaitan, H., Killebrew, D., Koch, A., Kumar, N., Lacy, S., Laudon, J., Law, J., Le, D., Leary, C., Liu, Z., Lucke, K., Lundin, A., MacKean, G., Maggiore, A., Mahony, M., Miller, K., Nagarajan, R., Narayanaswami, R., Ni, R., Nix, K., Norrie, T., Omernick, M., Penukonda, N., Phelps, A., Ross, J., Ross, M., Salek, A., Samadiani, E., Severn, C., Sizikov, G., Snelham, M., Souter, J., Steinberg, D., Swing, A., Tan, M., Thorson, G., Tian, B., Toma, H., Tuttle, E., Vasudevan, V., Walter, R., Wang, W., Wilcox, E., Yoon, D. H.: In-Datacenter Performance Analysis of a Tensor Processing Unit. ISCA 2017: 1-12

Extraction of Classification Rules Using a Bee Swarm Approach

Sadjia Benkhider

Our paper provides a metaheuristic-based approach to extract a set of classification rules. Indeed the classification problem is a data mining task which belongs to the NP-complete Problem class. This problem is combinatorial and then may find good solutions using metaheuristic approaches. We are interested in adapting the Bee Swarm Optimisation (BSO) algorithm which is relatively a new metaheuristic based on the life of bees. Our aim is to discover a classification model. This last is constituted by a set of rules as follows:

R_1: If $\text{att}_i = \text{val}_{ij}$ and $\text{att}_k < \text{val}_{kl}$ and ... then class = C_1
R_2: If $\text{att}_m > \text{val}_{mi}$ and $\text{att}_l = \text{val}_{li}$ ... then class = C_2
$\vdots$
R_k: If $\text{att}_i \geq \text{val}_{ik}$ and $\text{att}_m \leq \text{val}_{mk}$ then class = C_k

A rule presents the following form $A \quad C$ Where A is called the antecedent of the rule constituted by a set of attributes as following : att_i operator val_{ij}

With att_i is the ith attribute and operator is = or $\neq$ if att_i is a nominal attribute, and = or $\leq$ or $\geq$ if att_i is a numeric or continuous attribute. val_{ij} is the jth value of the ith attribute in the domain of its values. The methods of rule-based classification derive from explicit knowledge data directly. This is why users of these classifiers can often assess the accuracy and utility of the discovered knowledge. Otherwise, the BSO approach already gave some interesting solutions for the well known SAT problem. This motivated our choice to adapt BSO to this problem. In this paper we show the efficiency of the method for the Problem of the Extraction of Classification Rules, using some benchmarks of the UCI repository. Indeed the obtained results are very promising when compared with other methods of classification on the same benchmarks and the accuracies are around 90%.

Machine Learning 2

Learning Choice Functions

Karlson Pfannschmidt, Pritha Gupta, Eyke Hüllermeier

We study the problem of learning choice functions, which play an important role in various domains of application, most notably in the field of economics. Formally, a choice function is a mapping from sets to sets: Given a set of choice alternatives as input, a choice function identifies a subset of most preferred elements. Learning choice functions from suitable training data comes with a number of challenges. For example, the sets provided as input and the subsets produced as output can be of any size. Moreover, since the order in which alternatives are presented is irrelevant, a choice function should be symmetric. Perhaps most importantly, choice functions are naturally context-dependent, in the sense that the preference in favor of an alternative may depend on what other options are available. We formalize the problem of learning choice functions and present two general approaches based on two representations of context-dependent utility functions. Both approaches are instantiated by means of appropriate neural network architectures, and their performance is demonstrated on suitable benchmark tasks.

Data Augmentation for Discrimination Prevention

Vasileios Iosifidis, Eirini Ntoutsi

Data-driven decision support systems are part of our daily life. Machine Learning models are built upon historical data in order to provide predictions for future unknown problem instances. However, the training data might be already biased and sensitive groups like gender or race minorities might be underrepresented. Such biases in the training data, might be inherently learned by a (machine learning) model and propagated to decisions on future unseen instances. This topic has raised a lot of recent attention in the society and the machine learning community. As a result, a variety of methods have been proposed over the recent years in order to eliminate (machine learning) discrimination [1-3]. In this work, we focus on supervised learning where the class label might be correlated to sensitive attributes like race or gender. Just removing those attributes from the training set is not adequate as there might be proxy attributes or combinations of attributes. Therefore, more corrective actions are required in order to eliminate bias.

In this work, we propose data augmentation techniques to correct for bias at the input/data layer [4]. By augmenting under-represented groups, we aim to facilitate the classification model to learn beyond the obvious. We employ two well-known augmentation techniques: - oversampling, where randomly selected original data instances are duplicated, and - SMOTE [5], where new synthetic instances are generated in the neighborhood of existing instances.. . Both methods have their limitations. On the one hand, oversampling does not add new information, but just multiplies existing information. On the other hand, SMOTE adds new information, however the synthetic instances might distort the real data distribution as the generation process is prone to errors. Despite their limitations, both methods are helpful for balancing the dataset. Our experiments on two real world datasets with bias, namely, : census-income [6] and german-census [6], show that data augmentation is a promising technique for correcting for discrimination.

[1] Kamiran, Faisal, and Toon Calders. ”Data preprocessing techniques for classification without discrimination.” Knowledge and Information Systems 33.1 (2012): 1-33.

[2] Hajian, Sara, Francesco Bonchi, and Carlos Castillo. ”Algorithmic bias: From discrimination discovery to fairness-aware data mining.” Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

[3] Luong, Binh Thanh, Salvatore Ruggieri, and Franco Turini. ”k-NN as an implementation of situation testing for discrimination discovery and prevention.” Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011.

[4] V. Iosifidis, E. Ntoutsi, ”Dealing with Bias via Data Augmentation in Supervised Learning Scenarios”, BIAS workshop in conjunction with iConference 2018.

[5] Chawla, Nitesh V., et al. ”SMOTE: synthetic minority over-sampling technique.” Journal of artificial intelligence research16 (2002): 321-357.

[6] Merz, Christopher J., and Patrick M. Murphy. ”UCI Repository of machine learning databases.” (1998).

Label Noise Filtering based on Cluster Validation Measures

Veselka Boeva, Lars Lundberg, Jan Kohstall, Milena Angelova

When data is modeled using machine learning algorithms, existing noise and outliers can affect the generated model. Improving how learning algorithms handle noise and outliers can produce better models. In this study, we propose an outlier mining technique, entitled Cluster Validation Index (CVI)-based Outlier Mining which finds suspicious instances considering the class label. The proposed approach identifies and eliminates outliers from the training set, and a classification hypothesis is then built from the set of remaining instances.

Cluster validation measures are usually used for evaluating clustering solutions in unsupervised learning. We apply these well-known and scientifically proven measures for filtering outliers in training sets in supervised learning scenarios with known classes/clusters. The intuition is that instances in the training set that are not strongly connected to their clusters are outliers and should be removed prior to training. Our approach assigns each instance in the training set several cluster validation scores representing its potential of being an outlier with respect to the clustering properties the used validation measures assess. To improve the results further we combine those scores either by ranking or on a score base. In this way we managed to reflect different aspects of the clustering model determined by the CVIs on the training set. Based on this joint score we perform the filtering.

We examine the effects of mining outliers for five commonly used learning algorithms on ten data sets from the

UCI data repository using three different cluster validation indices. We study how filtering based on different combination of the three cluster validation indices affects the classification performance. The obtained results demonstrate that the proposed approach is a robust outlier filtering technique that is able to improve classification accuracy of the learning algorithm. Our approach can easily be adapted to different data sets and learning algorithms by using a proper combination of cluster validation measures reflecting different clustering properties. It turns out that the CVI-based algorithm outperforms LOF in most cases, particularly when we combine several cluster validation measures.

A Forest of Stumps

Amirah Alharthi, Charles Taylor, Jochen Voss

Many numerical studies indicate that bagged decision stumps preform more accurately than a single stump. In this work, we will introduce a new stump-based ensemble method for classification which is: A Forest of Stumps ”Gini-Sampled Splits”. A stump within this forest uses a split that is generated from transformed Gini indices for each possible cut points. The choice of variable which is chosen on which to generate a split has a probability proportional to that variable Gini index values. The final decision of these stumps is aggregated using weighted vote rather than majority vote. We compared between this method and other tree-based ensemble classification methods in terms of the accuracy and the results are promising.

On the Projection of Machine Learning Scores to Well-Calibrated Probability Estimates

Johanna Schwarz, Dominik Heider

Machine learning models, in particular computer-assisted clinical decision support systems (DSS), have been shown to be powerful tools to reduce medical costs and errors (Tsai et al., 2003) and have been applied in numerous fields, ranging from endoscopy (Dechêne et al., 2014) towards prediction of drug resistance in pathogens (Heider et al., 2014). Nevertheless, these models typically have a caveat: many of them are perceived as black boxes by clinicians and, unfortunately, the resulting classifier scores cannot usually be directly interpreted as class probability estimates. Consequently, various calibration methods have been developed in the last two decades from the most basic approaches such as Platt scaling to more advanced Bayesian histogram binning strategies. However, finding the best-suited calibration method for a specific classification problem can be a tedious task as there is no easy-to-use tool available that allows a quick and comparative analysis of different calibration methods.

In this work, we present the R-package CalibratR, which can be used to automatically transform machine learning scores to well-calibrated probability estimates. Furthermore, we compare the calibration performance of four different state-of-the-art calibration methods, namely scaling, transforming, histogram binning and BBQ (Naeini et al., 2015) with our novel probability estimation method GUESS.

CalibratR was evaluated on simulated and real data sets where it successfully projected uncalibrated machine learning scores to reliable probability estimates and thus minimized the calibration error in the training and test sets. With the help of CalibratR, we were able to identify the optimal calibration method for each data set and application in a timely and efficient manner.

Using calibrated probability estimates instead of original classifier scores will contribute to the acceptance and dissemination of machine learning based classification models in cost-sensitive applications such as clinical research where easy-to-use yet reliable computer-assisted DSS are urgently needed to reduce preventable human errors.

A. Dechêne, C. Jochum, C. Fingas et al. (2014). Endoscopic management is the treatment of choice for bile leaks after liver resection. Gastrointest Endosc., 80(4):626-633.

D. Heider, J.N. Dybowski, C. Wilms, D. Hoffmann (2014): A simple structure-based model for the prediction of HIV-1 co-receptor tropism. BioData Min., 7:14.

M.P. Naeini, G.F. Cooper, M. Hauskrecht (2015): Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of the AAAI Conference on Artificial Intelligence, 2901–2907.

T.L. Tsai, D.B. Fridsma, G. Gatti (2003): Computer decision support as a source of interpretation error. The case of electrocardiograms. Journal of the American Medical Informatics Association, 10(5):478–483.

Machine Learning and Optimization

A Branch and Bound Algorithm for Decision Trees with Optimal Cross-Splits

Ferdinand Bollwein, Martin Dahmen, Stephan Westphal

State-of-the-art decision tree algorithms are top-down induction heuristics which greedily partition the attribute space by iteratively choosing the best split on an isolated attribute. Despite their attractive performance in terms of runtime, simple examples, such as the XOR-Problem, point out that these heuristics often fail to find the best classification rules if there are strong interactions between two or more attributes from the given datasets.

In this context, we present a branch and bound based decision tree algorithm (BBTree) to identify optimal bivariate axis-aligned splits according to a given impurity measure. In contrast to a single split that can be found in linear time, such an optimal cross-split has to consider every combination of values for every possible selection of pairs of attributes, which leads to a combinatorial optimization problem that is quadratic in the number of values and attributes. To overcome this complexity, BBTree uses a branch and bound procedure, a well-known technique from combinatorial optimization, to divide the solution space into several sets and to detect the optimal cross-splits in a short amount of time.

These cross splits can be used either directly to construct quarternary decision trees or they can be used to select only the better one of the individual splits. In the latter case, the outcome is a binary decision tree with a certain sense of foresight for correlated attributes. We test both of these variants on various datasets of the UCI Machine Learning Repository and show this way that cross splits can consistently produce smaller decision trees than state-of-the-art methods with comparable accuracy. In some cases, our algorithm produces considerably more accurate trees due to the ability to draw more elaborate decisions than single-variate induction algorithms.

Deep Learning Assisted Heuristic Tree Search for the Container Pre-marshalling Problem

André Hottung, Shunji Tanaka, Kevin Tierney

One of the key challenges for operations researchers solving real-world problems is designing and implementing high-quality heuristics to guide their search procedures. In the past, machine learning techniques have failed to play a major role in operations research approaches, especially in terms of guiding branching and pruning decisions. We integrate deep neural networks into a heuristic tree search procedure to decide which branch to choose next and to estimate a bound for pruning the search tree of an optimization problem. We call our approach Deep Learning assisted heuristic Tree Search (DLTS) and apply it to a well?known problem from the container terminals literature, the container pre-marshalling problem (CPMP). Our approach is able to learn heuristics customized to the CPMP solely through analyzing the solutions to CPMP instances, and applies this knowledge within a heuristic tree search to produce the highest quality heuristic solutions to the CPMP to date.

Gaussian Process Emulation of Computer Experiments with Both Continuous and Categorical Inputs

Dominik Kirchhoff

This talk deals with several approaches to incorporating categorical input variables into Gaussian Process Emulators (also known as Kriging models).

Kriging is an important tool for the global optimization of black-box functions. The original model can only cope with purely continuous input variables. Since many applications include also categorical variables, we consider extensions that are able to deal with this case.

A brief overview of different existing approaches called Exchangeable Correlation (EC), Multiplicative Correlation (MC), and Unrestrictive Hypersphere-based Correlation (UC) is provided. We then introduce two new methods. The first one lets the practitioner choose the number of parameters to be estimated freely. The other method is as parsimonious in terms of computational effort as MC but is able to capture negative cross-correlations.

The various methods are applied to a synthetic test function in order to illustrate their strengths and weaknesses.

A First Analysis of Kernels for Kriging-based Optimization in Hierarchical Search Spaces

Martin Zaefferer, Daniel Horn

Many real-world optimization problems, including hyperparameter tuning and other machine learning related topics, require significant resources for objective function evaluations. This is a challenge to evolutionary algorithms, as it limits the number of available evaluations. One solution are surrogate models, which replace the expensive objective.

A particular issue in this context are hierarchical variables. Hierarchical variables only influence the objective function if other variables satisfy some condition. For instance, parameters of an SVM kernel only influence the performance if that kernel is utilized. We study how this kind of hierarchical structure can be integrated into the model based optimization framework. We discuss an existing kernel and propose alternatives. An artificial test function is used to investigate how different kernels and assumptions affect model quality and search performance.

Machine Learning for Dynamic Systems

Ensemble Methods for Tracking Various Concept Drift Structures

Dhouha Mejri, Mohamed Limam, Claus Weihs

Concept drift is one of the most relevant challenges in data stream mining. Ensemble methods have been effectively used to handle it due to their ability to learn from continuously arriving data and to incorporate dynamic updates. This article presents a new application of Dynamic Weighted Majority-Winnow (DWM-WIN) ensemble method on datasets with gradual, sudden and nonlinear drifts. A comparison of DWM-WIN based on ensembles of naiveBayes, decision trees and kknn learners is conducted. A clear vision of the ensemble's behavior based on misclassification error rates is provided. Experiments show that DWM-WIN is not only an effective concept drift identification method but has also a high adaptation capacity to noise and learner's type.

Structure Identification of Dynamical Takagi-Sugeno Fuzzy Models by Using LPV Techniques

Matthias Kahl, Andreas Kroll

Takagi-Sugeno (TS) fuzzy models (Takagi and Sugeno, 1993), which approximate nonlinear systems by a weighted superposition of local linear models, have successfully been applied in many industrial problems. In order to obtain a model which performs well for the considered task an appropriate model structure has to be chosen. The problem of selecting significant subsets of regressors or features is known from multivariate data analysis such as linear regression or pattern recognition and investigated in the areas of data mining and machine learning (see, e.g., Hastie et al., 2009; May et al., 2011). In case of identification of dynamical systems, this involves the choice of relevant physical variables and individual time lags of each variable resulting in a large set of potential model candidates. The structure selection problem of TS models consists of three tasks: 1) the choice of appropriate scheduling variables, 2) the partitioning of the scheduling space by an appropriate parameterization of the membership functions of a predefined type as well as the choice of the number of local models, and 3) the selection of a suitable local model structure. In this contribution the problem of order selection for TS fuzzy models is investigated and solved by formulating the TS model in its Linear Parameter Varying (LPV) form and applying a recently proposed regularized Least Squares Support Vector Machine technique for LPV models (Piga and Toth, 2012). In contrast to parametric identification approaches this method does not need to solve the TS partition design task in terms of a priori specified basis functions by application of the kernel trick and enables the selection of the model order by solving a regularized convex optimization problem. Once the correct model order is found a parametric TS model can be estimated e.g. by clustering and estimation of the local model parameters. The proposed approaches are illustrated in a case study.

T. Hastie, R. Tibshirani, J. Friedmann (2009). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.

R. May, G. Dandy, H. Maier (2011). Review of input variable selection methods for artificial neural networks. The Artificial Neural Networks – Methodological Advances and Biomedical Applications, InTech.

D. Piga, R. Toth (2013). LPV model order selection in an LS-SVM setting. 2013 IEEE 52nd Annual Conference on

Decision and Control, pp. 4128-4133.

T. Takagi, M. Sugeno (1993). Fuzzy identification of systems and its application to modelling and control. Readings in Fuzzy Sets for Intelligent Systems, pp. 387-403.

OpenCL Deep Learning Framework for Automated Text Recognition

Patrick Kappen, Lester Kalms, Diana Goehringer

Automated handwritten character recognition is a challenging task that cannot easily be solved by using classical algorithms. An alternative approach is deep learning, which has proven to be successful, but requires many computational resources. In this work, automated handwritten character recognition was solved by applying various neural networks. Normal encoder-based neural networks suffer from the high dimensionality of images and the requirement of fixed image sizes. Thus, a contribution of this work is the use of sliding windows to increase the flexibility of automated handwritten character recognition by allowing flexible image sizes. To be usable on different hardware devices like CPUs, GPUs and FPGAs, an OpenCL deep learning framework was developed. The framework supports fast creation of different neural networks including CNNs, RNNs and combinations of both. The implemented framework and neural networks where successfully applied to the handwritten character recognition problem, with an accuracy of 89,6%.

A Reinforcement Learning Strategy for the Swing-Up of the Double Pendulum on a Cart

Michael Hesse, Julia Timmermann, Ansgar Trächtler, Eyke Hüllermeier

The effective control design of a dynamical system traditionally relies on a high level of system understanding, usually expressed in terms of an exact physical model. In contrast to this, reinforcement learning [SB12] adopts a data-driven approach and constructs a control strategy by interacting with the underlying system. Therefore complex control tasks can be solved without the need of much expert knowledge. In our research, we used the state-of-the-art model-based reinforcement learning method PILCO (Probabilistic Inference for Learning Control [DFR16]), which is able to learn a probabilistic model of the dynamical system and to design a feedback control strategy from scratch. In comparison to other reinforcement learning methods (e.g. model-free Deep Deterministic Policy Gradients [SLH+14]), PILCO is very data efficient and requires only a few interactions with the system. In view of technical systems, this is of great importance, as the wear and tear of the system should be as low as possible. We realized the swing-up from the lower stable to the upper unstable equilibrium of a double pendulum on a cart, both on a validated simulation model and on the real system by using PILCO. In order to achieve this task, we introduced state restrictions via penalty terms into the PILCO framework, so that the limited cart distance of the test bench can be taken into account during the learning process. By using the cart acceleration instead of the cart force as the input to the system, an improvement in terms of learning speed could also be achieved. With these adjustments, the swing-up task was successfully solved in 27 learning iterations, which corresponds to a total interaction time of 88 seconds.

Development of a Concept for Process Improvement Based on Large Amount of Data

Steffen Hovestadt

The current market situation and the industrial environment are shaped by the globalization of the economy. This results in new challenges for industrial companies, which are characterized by high competitive pressure. Due to these circumstances, industrial companies are confronted with numerous requirements, which consist in a constant high quality of the products, an efficient and flexible production and the fast development of their products.

Many companies meet these requirements by continuously improving their processes in terms of increasing the quality of their products and the efficiency of their production processes. Another aspect is the growing use of digital tools for planning, monitoring and controlling of the production as well as supporting business processes. These digital tools enable the handling of the rising product and process complexity.

To improve their processes, industrial companies often use standardized procedures such as the DMAIC cycle. With its help the actual state of a process is determined based on measurable data so that improvements for the process can be developed and implemented. On the other hand data mining process models such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) are used, in order to exploit the large volumes of data that are generated by the consistent use of digital tools. Their goal is to generate knowledge from the data as efficiently as possible, which is nowadays of highest importance for the company’s success.

Against the background of the current industrial environment, the DMAIC cycle has weaknesses in process improvement. The reason is that it is not designed for the use of larger and more complex data structures and thus cannot be used to evaluate this data, without further preparation. By contrast, the CRISP-DM is designed solely for the generation of knowledge from data and cannot be applied as a process improvement method because therefore it lacks several relevant components.

As a result, both of the standard approaches have deficits in terms of process improvement based on large amounts of data. Additionally despite the enormous potential of improving such processes, there is currently no adequate standardized method for this objective. Consequently, the development of a general concept for process improvement, based on large amounts of data, is sought. The concept is to build on the DMAIC cycle and CRISP-DM and combine their advantages. For this purpose, elements of the CRISP-DM are integrated into the DMAIC cycle in order to make it usable for the evaluation of larger amounts of data.

Behavioural Profiling of Industrial Assets Through Numerical Encoding of Event Logs

Pierre Dagnely, Tom Tourwé, Elena Tsiporkova

More and more industrial assets continuously report events about their status to raise alarms and warnings. These events are stored in so-called event-logs that contain valuable information about the asset behavior and could be explored to find e.g. assets sharing similar behavior. One challenge specific to the exploitation of event logs is their textual nature. Most machine learning (ML) methods have been developed for numerical (i.e. sensor) data and can not be applied as such on event logs without some pre-processing e.g. extracting numerical features as event frequency and durations. Another approach, not yet widely explored to the best of our knowledge, is to employ more sophisticated techniques in order to convert event logs into numerical vectors. For instance, Fronza et al. [Journal of Systems and Software 86, 2–11 (2013)] utilized random indexing to transform event logs from software runs into high dimensional bit vectors which embed the event context, i.e. surrounding events. Subsequently, SVM was trained on the high dimensional context vectors to predict failures. The advantage of such encoding approach is twofold: 1) allowing to use powerful and well validated ML methods 2) embedding additional knowledge (e.g. context) in the numerical encoding. In this study, we develop and explore further several different encoding techniques specially dedicated to event log data originating from "real-life" industrial assets e.g. portfolio of photovoltaic installations and fleet of wind turbines. These techniques include the estimation of relevance scores of the events across the whole portfolio/fleet allowing to arrive at typical behavioural fingerprints in the form of numerical vectors. Combining this approach with pattern and process mining enables the derivation of some characteristic behaviour profiles allowing for labeling, quantifying and interpreting performance.

Mining Streaming and Time-Evolving Data

A General Extension for Online Discriminant Analysis Methods for Data Streams with Concept Drift

Sarah Schnackenberg, Uwe Ligges, Claus Weihs

Various methods based on classical classification methods such as discriminant analysis (DA) have been developed for working on data streams in situations with concept drift (Hoens et al., 2012). Some of these methods have been summarized by Schnackenberg et al. (2018). Although such methods may lead to a good goodness of fit, their classifiers may result in a bad prediction error rate in case the underlying distribution gradually changes further on. Therefore we invented a rather general extension to such methods to improve the forecasting quality. An example how to extend the method proposed by Kuncheva and Plumpton (2008) has been given by Schnackenberg et al. (2018). In the talk we describe the general extension that leads to considerably improved error rates in concept drift situations for all methods we tried. Under some assumptions we estimate a model for the time depending concept drift that is used to predict the forthcoming distribution of the features. These predictions of distributions are used in the DA to learn the classification rule and hence for predicting new observations. In a simulation study we consider different kinds of concept drift and compare the new extended methods with the methods these are based on.

T.R. Hoens, R. Polikar, N.V. Chawla (2012). Learning from streaming data with concept drift and imbalance: an overview. Progress in Artificial Intelligence, 1(1):89-101.

L.I. Kuncheva, C.O. Plumpton (2008). Adaptive Learning Rate for Online Linear Discriminant Classifiers. Structural, Syntactic, and Statistical Pattern Recognition. Lecture Notes in Computer Science, 5342:510-519.

S. Schnackenberg, U. Ligges, C. Weihs (2018). Online Linear Discriminant Analysis for Data Streams with Concept Drift, submitted to: Archives of Data Science, Series A.

Temporal Density Extrapolation in Data Streams with Basis Expansion and Compositionally Modelled Coefficients

Dominik Lang, Vera Hofer

In the dynamic environment of data streams the underlying distribution of a population is likely to change over time. This results in the need to adjust trained models, since the original data they were trained on has become outdated and less representative of the underlying distribution. Both descriptive and predictive models require an update, however acquiring new and more representative data right away can be costly or simply not an option, for example when new data only become available with a certain delay. Therefore it makes sense to model the density distribution's development itself over time and to make predictions regarding the distribution at a future time point. This way less data overall has to be acquired or processed for adaptation, a better understanding of the dynamics in the stream can be acquired and the issue of delayed access to new data is avoided. To this end we present a non-parametric approach in order to model the streams dynamics. The model is based on a univariate basis expansion fit to historic data. The evolution of the basis coefficients is approached as a compositional data problem, which is solved for the entire set of historic data instead of batches. We demonstrate the performance of the model on artificial data, which is used to simulate relevant phenomena in data streams.

Scalable Implementation of Dynamic Factor Machine Learning for Very High Dimensional Forecasting,

Gianluca Bontempi

Big data forecasting (e.g. in IoT) requires computational solutions able to process a massive number of time series in accurate and scalable manner. Although both computational and statistical challenges need to be jointly addressed, most existing works consider only a portion of the problem. For instance, statistical forecasting literature proposes well-founded strategies to improve prediction accuracy but are often limited to consider low dimensionality (at most hundreds of series) and single processor implementation on conventional architectures. At the same time the analytics methodology of most recent works on big data streaming is often trivial and boils down to some summary statistics.

This paper aims to go one step further by introducing and assessing a scalable implementation of an accurate forecasting solution for very large dimensional forecasting problems. The goal is to scale up the size of multivariate task to a huge number (e.g. millions) of time series and to assess jointly two aspects: the accuracy of the prediction and the scalability of the solution.

In a recent publication we proposed DFML (Dynamic Factor Machine Learner) (Bontempi et al., 2017), an extension of the conventional Dynamic Factor Model (DFM) which originated in the economic forecasting community. The basic idea of DFML is that a small number of unobserved series (or factors) can account for the time behavior of a much larger number of variables. If we are able to obtain accurate estimates of these factors by machine learning, the forecasting endeavor can be made simpler and scalable by using the estimated dynamic factors for forecasting instead of all the series.

In this paper we propose a scalable and streaming implementation of DFML to deal in an accurate and efficient way with very large dimensional forecasting. The scalable implementation relies on an online implementation of PCA, a distributed prediction of the factors and a distributed assessment of the model accuracy. In particular we assess both the accuracy and the scalability of a Kafka-Spark architecture in a very large number of synthetic prediction tasks.

G. Bontempi, Y. A. L. Borgne and J. de Stefani (2017). A Dynamic Factor Machine Learning Method for Multi-variate and Multi-step-Ahead Forecasting, 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 222-231, Tokyo, Japan

Improving Predictions of Polarities of Entity-Centered Documents using Entity-Centered Multinomial Naive Bayes

Christian Beyer, Uli Niemann, Vishnu Unnikrishnan, Eirini Ntoutsi, Myra Spiliopoulou

We report on our work (Beyer et al., 2018), where we investigate the usefulness of entity-centered, text-ignorant models for the prediction of amazon ratings. In that study, we model each product as an entity and the incoming ratings with regards to a product are considered to form a trajectory belonging to said entity. We divide the incoming stream of product reviews into a training and a test stream ensuring that we have a minimum amount of ratings available for initializing the entity-centric models and also a minimum amount of reviews available for testing the performance of the prediction models. After excluding entities which were to short we end up with 33989 total entities which were are using for evaluation. The entity-centric models are compared against global models which see all the incoming reviews of all entities. Our findings on products from the Tools and Home Improvement category suggest that for some entities simple text-ignorant models could outperform a more complex prediction model based on Multinomial Naive Bayes with Forgetting which had access to the review texts. The MNBF achieved the highest overall prediction performance according to multiple measure including RMSE, (balanced) accuracy and the Kappa+ statistic. Still a simple prediction model based on the prior rating distribution of an entity could achieve a better Kappa+ value in up to 18 percent of the entities. Now our goal is to use more sophisticated entity-centered models in the form of an entity-centered Multinomial Naive Bayes within an ensemble. Additionally, we want to investigate in more depth under what conditions an entity-centric view on a data stream can be of advantage.

C. Beyer, U. Niemann, V. Unnikrishnan, E. Ntoutsi, and M. Spiliopoulou. Predicting polarities of entity-centered documents without reading their contents. In Proceedings of the Symposium on Applied Computing. ACM,2018.

Improving Feature Selection for Multinomial Naive Bayes Classifiers Over Textual Streams

Damianos Melidis, Eirini Ntoutsi

Abstract. Textual streams are more and more evident in nowadays social media life. Consequently, researchers aim to automatically understand feelings and opinions conveyed by these streams. Two critical constraints for learning on streams are the memory efficient and the any-time prediction. A method with an improved feature selection part should meet these constraints because it will create a reduced but representative feature space.

In this work, we propose a feature selection method for the Multinomial Naive Bayes classifier over a textual stream. We extended the method of [3] for incremental feature selection in the sketch proposed to keep the most frequent words [1]. Thus, we have a memory efficient representation of the most frequent words per class. To evaluate our approach we used Twitter data stream monitored over a period of three months [2]. As evaluation measures we applied the accuracy and kappa statistic, κ calculated by prequential evaluation over sliding windows. We are also interested to integrate more complex ideas such as [4] to this approach.

1. Bifet, A., Holmes, G., Pfahringer, B.: Moa-tweetreader: real-time analysis in twit- ter streaming data. In: International Conference on Discovery Science. pp. 46–60. Springer (2011)

2. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant su- pervision. CS224N Project Report, Stanford 1(12) (2009)

3. Katakis, I., Tsoumakas, G., Vlahavas, I.: Dynamic feature space and incremental feature selection for the classification of textual data streams. Knowledge Discovery from Data Streams pp. 107–116 (2006)

4. Tang, B., Kay, S., He, H.: Toward optimal feature selection in naive bayes for text categorization. IEEE transactions on knowledge and data engineering 28(9), 2508– 2521 (2016)

Analysis of Patient Evolution on Time Series of Different Lenghts

Vishnu Unnikrishnan, Rüdiger Pryss, Thomas Probst, Manfred Reichert, Winnfried Schlee, Berthold Langguth, Myra Spiliopoulou

Recent advances in computing have had an increasingly large role in decreasing data storage and networking costs, while increasing computational capabilities of ever-smaller devices. This has enabled the development of applications to monitor heretofore inaccessible environments. One of the prime fields where technology can play a disruptive role is that of medicine, where live patient data can enable preemptive intervention, or even help medical professionals monitor disease manifestation and development in the natural environment.

This study uses data from ’Track your Tinnitus’, a mobile crowd-sensing platform that collects ’Ecologically-valid Momentary Assessments’ (EMAs) regarding the neuro-psychiatric disorder tinnitus. The EMAs form a multivariate time series of 8 questionaire-based variables (most with answers from a Likert scale), of which the variable ’distress’ (current distress of the participant, as caused by tinnitus) is of most interest. The goal of the study is to investigate whether EMAs can be used to predict patient evolution, and if sub-populations of high in-group similarity exist within the people who suffer from the phantom-sound disorder. In addition to the dynamic data, registration-time information is also available like gender, age, etc. The subgroup discovery is performed through the use of Dynamic Time Warping (DTW) to identify users with similiar, but non-aligned time series.

The original dataset contained observations from 1288 participants over 2+ years, but 50% of them were active for no more than two days. All participants with fewer than 10 days of data were removed, shrinking the dataset to just 274 participants. The number of days of data available per participant was still skewed, resulting to time series length between 10-486. In addition, most time series were sparse, with many days for which no responses were recorded. The sparsity problem was solved through mean-imputation (which was found perform better than median, regression, ARIMA, etc. in all but the longest of time sries). In order to ensure that DTW does not noise-match a participant of just 10-days of data against another with 400, k-medoids was applied on the participants' time series lenghts, with DTW computation being restricted to patients within each of the five resulting clusters. The resulting DTW distances were visualised using multiple techniques - line plots of the actual time series, dendrograms, and Force Directed Graphs (FDGs). FDGs were particularly useful in discovering that the underlying data did indeed contain subgroups, though a large number of participants were seen to be dissimilar to everybody else, causing traditional clustering algorithms like k-means to fail. The manually-identified subgroups were tested for significantly differing EMA values at the beginning of their participation compared to the end. However, the tests yielded no statistical significance, suggesting that similar evolution is not what makes their EMAs similar. Further work is needed to determine what static factors impact patient similarity, and to identify if there is a more systematic way to identify patients that do not fall into any sub-population.

Multimodal Data and Cross-modal Relations: Analytics and Search

Combining Textual and Visual Stylometry in the Analysis of Graphic Narrative

Rita Hartel, Alexander Dunst

Textual stylometry has a long tradition in linguistics and literary studies and has only gained in popularity with the digitization of text corpora and tools. Stylometric methods for paintings have been advanced in digital art history but remain at an early stage of development. In contrast, stylometric analyses for visual narratives, such film, TV, and comics, are not yet established. In this paper, we examine the relationship between visual and textual stylometry for graphic narratives - or book-length comics aimed at an adult readership - as representative of multimodal texts. Our analysis is based on a corpus of 209 graphic narratives that has been designed as part of the interdisciplinary research group "Hybrid Narrative: Digital and Cognitive Approaches to Graphic Literature". Unfortunately, a reliable parsing of the retro-digitized comic books is still in its infancy. Therefore, we have to semi-automatically annotate the data with the help of our XML-based "Graphic Narrative Markup Language (GNML)," prior to analyzing the data. GNML annotation covers visual objects (e.g. panels, characters or speech bubbles) as well as the textual objects (that typically are hand lettered and present challenges for automated text recognition systems). For the annotation process, we use our M3-Editor, which stands for multi-modal markup and supports the visual annotation of objects in the form of polygons. The M3-Editor also incorporates semi-automatic annotation-mechanisms such as automatic panel detection, flood-filling for speech-bubbles or a lightwire segmentation technique based on edge detection to efficiently mark characters and other irregular objects.

Our analysis uses well established techniques for textual stylometry, including most frequent words or type-token-ratio. Furthermore, our analysis of visual stylometry employs standard measures for brightness/luminance, colors, shapes, entropy and edge structure and is based on standardized image descriptors, e.g. the color layout and edge histogram descriptors of MPEG-7. Our analysis does not consider textual and visual stylometry in isolation but focuses on identifying correlations between them. With the help of visual and textual stylometry, we try to answer questions like: Do similarities in textual stylometry correlate with similarities in terms of visual stylometry? Or more specifically: Which stylometric textual features correlate with which stylometric visual features?

Applying Frequent Pattern Mining to Multimodal Behavior in Interaction: Visualizing Significant Patterns

Katharina Rohlfing, Marcel Ruland, Sascha Henzgen

Human social interaction can be characterized by multimodal behavior. According to theoretical positions emphasizing that communication is organized by the interaction partners jointly, we identified the challenge of assessing human sequential behavior that is spread across different modalities and co-constructed with a partner. In previous work, we faced this challenge by applying frequent pattern mining in an analysis of a corpus of mother-child dyads.

The application of frequent pattern mining provided some support and initial results for the proposition that human interactive behavior is sequentially organized. Accordingly, verbal and nonverbal behavior are co-constructed by the interaction partners and form a range of patterns. For example, with respect to the occurrence of maternal vocal behavior, some nonverbal framing was notable. Firstly, one pattern with a high confidence suggests an intrapersonal sequence of gazing at the infant, smiling and speaking. In contrast, another pattern suggests an interpersonal sequence of mother gazing at her infant, infant gazing back, followed by vocal behavior of the mother. The analysis reveals patterns emerging between infants as young as 3-months-old and their mothers.

Taken together, frequent pattern mining is a more exploratory analysis as the premises and conclusions emerge as a result of it. While this method yields contingencies and dependencies and provides their frequencies, it is not yet investigated how the significance of these patterns can be calculated for behavioral studies and how it can be visualized. Our paper presents first solutions to the problem of how to discern and visualize the significance of some patterns with respect to other existing patterns in the sample.

Using Voronoi-Cells to Assess Action Efficacy in High-Performance Soccer

Robert Rein, Daniel Memmert

As an invasion game, tactics in soccer address how a team creates and occupies space on the pitch. One way for a team to manipulate its' space control is through passing behaviour. For example, by crossing the ball from one pitch side to the other often new opportunities for attacking play are created due to the changes of player's position on the pitch and resulting changes in space occupancy. Consequently, a soccer player's ability to make an ”effective” pass in a given play situation is considered one of the key skills characterizing successful performance in elite soccer. However, although there is ample evidence in the literature that passing behaviour is important it is much less clear what actually characterizes a ”good” pass. At present, one prevalent approach to investigated passing behaviour in elite soccer is based on the frequency of passing events and their correlation with game performance. Naturally, very little information is gained through this approach about how passing behaviour affects space control in soccer. Here we present an analytical approach to assess passing efficiency in elite soccer using a Voronoi-tessellation of the soccer pitch. The tessellation is performed according to the positioning of individual players such that each player is being assigned a specific area on the pitch which is under his control. However, not all pitch areas are equally important for successful game play. Here, in particular the area in front of the goal, the final attacking third, is regarded as the critical area. Therefore, in the present project we investigated how passing behaviour affects space control of the final third of the pitch. To this end position data from 103 first Bundesliga matches were analysed. Passing events were automatically extracted using a custom software package (SOCCER). The results showed that indeed more successful teams are able to increase their space control in the final third of the pitch due to their passing behaviour. The presented approach provides a first step into assessing passing efficiency in soccer for individual events as opposed to game average metrics previously used. At present, various extensions of this approach are being investigated. Moreover, the present approach demonstrates how a data analytical perspective of team game behavior can provide important insights for current practices in coaching.

Towards Analytics of Relations between Scientific Publications and Related Software Implementations

Anett Hoppe, Jascha Hagen, Helge Holzmann, Günter Kniesel, Ralph Ewerth

Today's research work, especially in technical domains, is seldom fully explained by only one scientific publication. The article might deliver the explanation of the scientist's rationale – a full reproduction of the presented results, however, often demands the review of several resources: processed datasets, own and third-party source code, used computer platforms, and parameter settings. As a consequence, reproducing research results has become a complex and time-consuming task that is hardly ever consequently performed. Indeed, quite recently, Baker et al. (2016) characterise the current situation as a “reproducibility crisis” – stating that false results might not be discovered by common peer review.

In the line of work towards consequent reproducibility of scientific results, there are three main tasks to be tackled: (a) motivate researchers to reproduce past results and give them presentation space in scientific venues; (b) develop novel ways for the integrated presentation of future scientific results; (c) develop tools which allow for easy exploration of existing scientific work.

We present an exploration tool – called SciSoftX – which focuses on the two latter objectives, by linking two possible modalities of research work presentation: software source code and scientific article. We analyse the possible semantic inter-relations between both and develop a hierarchical model of the different levels of abstraction.

The tool's main interface provides the user with both modalities, and avoids the necessity to switch between different software platforms. Moreover, it provides functionalities for the automatic and manual definition of semantic links between them, relying on the above-mentioned hierarchical formalisation. The automatic construction is realized by so-called linkers that use heuristics to discover source code-related text in the article – and then search for the relevant source code fragments. Highlighting in both modalities allows the user to easily swap between the different modalities. Anyhow, the user can manually add semantic relations by simple mouse interactions. Furthermore, the tool provides a number of visualisations which allow users the exploration of code-text relations on different levels of abstractions.

The goal is to provide support for different users. As a tool for readers and reviewers, SciSoftX allows getting to know the code of other authors in an explorative way, in direct relation to its textual description. For authors, it can provide guidance during the writing process: The visualisations give indications which parts of the code are well-covered by the textual description and which should be described more thoroughly. Finally, the tool features export functionalities – the package of article and source code can thus be extended by an additional file which contains their formal inter-relations.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature News 533 (7604), 452.

Multivariate, multi-label and ranking data

Using Multi-Label Logistic Regression to Maximize Macro F-measure

Masaaki Okabe, Jun Tsuchida, Hiroshi Yadohisa

Multi-label classification is a method of supervised learning, in which multiple labels may be assigned to each instance. One typical application of this method is document classification; employing this method allows each document to be assigned to more than one class. However, one of the challenges of multi-label classification lies in the class imbalance problem. When some labels may have much less positive examples than negative one, classifier performance evaluation by accuracy is sometimes inappropriate for minority class detection. The macro F-measure that is the arithmetic mean of F-measure of each label is often used to evaluate a classifier in multi-label classification. This evaluation metric is useful when the desired multi-label classifier performance is based on the overall performance across all sets of data. In this study, we propose a method of macro F-measure maximization for multi-label data. This method is based on the Binary Relevance method. We employ logistic regression that is probabilistic classification model and, maximize macro F-measure. Many F-measure optimization methods use a ratio of the estimators to approximate F-measure. However, this type of optimization method tends to yield a larger bias as compared to a method that directly approximates the ratio. Thus, we estimated the approximate F-measure by estimating the relative density ratio. Then, the usefulness of the proposed method in classification is demonstrated by simulation studies.

R-Vine Mixture Model for Modeling Multivariate Count Data

Marta Nai Ruscone

Finite mixtures are applied to perform model-based clustering of correlated multivariate count data. Existing models do not offer great flexibility for modelling the dependence of this kind of data since they rely on potential undesirable correlation restrictions and strict assumptions on the marginal distribution. Here, we propose a model-based clustering method via R-vine copula. This method allows the relax the correlation restrictions of previous approaches and account for the asymmetry of the data by using blocks of asymmetric bivariate copulas. Simulated data and correlated multivariate count data collected during an ultra-running competition are used to illustrate the applicability of the proposed procedure.

A Representation of the Relationship Between Variables in Quantitative and Qualitative Mixed Data

Mako Yamayoshi, Jun Tsuchida, Hiroshi Yadohisa

Structural equation models have been used extensively in the past to find causal structures in continuous variable data. In such a framework, the Linear Non- Gaussian Acyclic Model (LiNGAM) can find a whole causal model within a stream of continuous variable data. LiNGAM uses a Path diagram to represent acyclic causal structures, which makes it possible to clarify the direction of causality between variables. However, in many disciplines, the data include both quantitative and qualitative variables. In such a case, LiNGAM will lead in the wrong causal direction, as it handles both quantitative and qualitative variables as continuous. Therefore, a graphic representation using Path diagrams may not correctly express the relationship between continuous and discrete variables in some cases. Due to this problem, it is necessary to improve the estimation and representation method for causal structures under such conditions. This study proposes a method to find causal structures for data that include both quantitative and qualitative variables. To overcome the difficulties of the existing method, we use a Link function combined with latent variables. Specifically, we assume that each variable has a relationship through latent variables, which sets a Link function as a function of the structural equation. Furthermore, the relationship between quantitative and qualitative variables is also represented based on such assumptions. Through simulation studies, we prove that our proposed method can represent the direction of causality by clarifying the relationship between variables that are more suitable for mixed data.

Learning to Rank based on Analogical Reasoning

Mohsen Ahmadi Fahandar, Eyke Hüllermeier

Object ranking or ”learning to rank” is an important problem in the realm of preference learning. On the basis of training data in the form of a set of rankings of objects represented as feature vectors, the goal is to learn a ranking function that predicts a linear order of any new set of objects. In this paper, we propose a new approach to object ranking based on principles of analogical reasoning. More specifically, our inference pattern is formalized in terms of so-called analogical proportions and can be summarized as follows: Given objects A,B,C,D, if object A is known to be preferred to B, and C relates to D as A relates to B, then C is (supposedly) preferred to D. Our method applies this pattern as a main building block and combines it with ideas and techniques from instance-based learning and rank aggregation. Based on first experimental results for data sets from various domains (sports, education, tourism, etc.), we conclude that our approach is highly competitive. It appears to be specifically interesting in situations in which the objects are coming from different subdomains, and which hence require a kind of knowledge transfer.

Ranking Distributions based on Noisy Sorting

Adil El Mesaoudi-Paul, Robert Busa-Fekete, Eyke Hüllermeier

We propose a new statistical model for ranking data, i.e., a new parametric family of probability distributions on permutations of a fixed size. Our model is inspired by the idea of a data-generating process in the form of a noisy sorting procedure, that is, the idea that a ranking is produced as the result of a sorting process, in which comparisons are not deterministic but dependant on chance. More specifically, comparisons between pairs of items are modelled as Bernoulli trials, with the Bernoulli parameters representing pairwise preferences. We show that our model can be written in closed form if insertion or quick sort are used as sorting algorithms, and address the problem of maximum likelihood parameter estimation based on sample data. We also introduce a generalization of the model, in which the constraints on pairwise preferences are relaxed, and for which maximum likelihood estimation can be carried out based on a variation of the generalized iterative scaling algorithm. Experimentally, we show that the models perform very well in terms of goodness of fit, compared to existing models for ranking data, on a large number of real-world data sets.

Recommendation and eCommerce

A Segmented Kano Perspective on the User Interface of Online Fashion Shops

Daniel Baier, Alexandra Rese

Curated shopping, scanned shopping, attended shopping, virtual fitting, and so on: Many concepts are discussed that could improve user interface of online fashion shops. But questions arise such as which groups of customers prefer which groups of concepts, which concepts really have the potential to increase conversion rates, sales, and customer satisfaction, in which should be invested (see, e.g., Rese et al. 2014, Baier, Rese 2017b).

To answer such questions from a market segments point of view, a new two-mode clustering approach – the Segmented Kano perspective – is developed and applied in this paper. The approach generalizes Baier et al. (1997)’s two-mode clustering algorithm. The application deals with a major German online fashion retailer who wants to improve the customers' ordering and delivery process. Data are collected using Kano's well-known functional/dysfunctional questioning approach (see Baier, Rese 2017a). A sample of 7,185 customers participated in the survey, 5,253 returned completed questionnaires and formed the basis for the analyses. The Segmented Kano perspective reveals large differences between groups of customers with respect to groups of concepts. The research gives valuable input to the company but is also very interesting for other online retailers from a methodological as well as derived results oriented point of view.

References

Baier, D., Gaul, W., Schader, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring, in: Klar, R., Opitz, O. (Eds.), Classification and knowledge organization. Springer, Berlin, Heidelberg, 557-566.

Baier D, Rese A (2017a): Improving the Ordering and Delivery Processes in Online-Fashion Shops: New Approaches to Integrate the Voice of the Customer. Paper presented at the Conference of the International Federation of Classification Societies (IFCS-2017). Tokyo, August 8-10.

Baier D, Rese A (2017b): Online-Shop Site Engineering Using Eye Tracking, TAM, and A/B-Tests: An Empirical Application, Paper presented at the 4th German-Polish Symposium on Data Analysis and its Applications (GPSDAA-2017), Wroclaw, September 25.

Rese A, Schreiber S, Baier, D (2014): Technology Acceptance Modeling of Augmented Reality at the Point of Sale: Can Surveys be Replaced by an Analysis of Online Reviews? Journal of Retailing and Consumer Services 21 (5): 869-876.

Modeling Customer Journey Value: Predicting Customer Churn and Future Net Value

Dominic Christian Pastoors, Daniel Baier

The digitalization significantly increases the possibilities for companies to interact with their customers, vice versa. As a consequence, customer journeys become increasingly complex but at the same time better traceable. It is apparent that companies should leverage this data-rich environment and take the opportunity to foster sustainable customer relationships by proactively managing journeys and customer experience. However, this includes decisions about investments into diverse marketing actions and channels. Even today, such decisions are primarily based on a marketer's gut feelings, resulting in a perceived lack of credibility of marketing departments (Homburg et al. 2014).

With the objective to enable credible decision-making, a novel approach is introduced that puts a customer's journey at the center of customer valuation. The model is based on the Customer Lifetime Value (CLV) approach. Here, the CLV of an individual customer represents its past and future discounted net cash flows over his entire lifetime with the firm (Dwyer 1997). This forward-looking metric has been proven to be suitable for decision-making purposes including resource allocation (Kumar and Reinartz 2016). Our research extends this approach by focusing on a customer's behavior rather than its lifespan. Hence, the chronological sequences of touchpoints becomes the basis for measuring and predicting a customer's net contributions. This foundation promises to provide more indicators for future conversions and churn.

The here presented paper puts its focus on modeling customer churn probability and predicting future net value. In this context, key challenges are given by the myriad of possible journeys a customer can take as well as the volume of data. Therefore, a variety of existing approaches are examined and evaluated towards their applicability on customer journeys. In addition, this paper provides preliminary results of a case study conducted at an international software company.

Dwyer, F. R. 1997. "Customer Lifetime Valuation to Support Marketing Decision Making," Journal of Direct Marketing, (11:4), Elsevier, pp. 6–13 (doi: 10.1002/(SICI)1522-7138(199723)11).

Homburg, C., Vomberg, A., Enke, M., and Grimm, P. H. 2014. "The loss of the marketing department's influence: is it really happening? And why worry?," Journal of the Academy of Marketing Science, pp. 1–13 (doi: 10.1007/s11747-014-0416-3).

Kumar, V., and Reinartz, W. 2016. "Creating Enduring Customer Value," Journal of Marketing, (0:ja), American Marketing Association, p. JM.15.0414 (doi: 10.1509/JM.15.0414).

Two-mode Overlapping Clustering for Three-mode Data with Applications to Online Shopping and Site Engineering

Atsuho Nakayama, Daniel Baier

Many two-mode clustering approaches are designed to subdivide a two-mode two-way data matrix with data values into row-by-column blocks with a small variation of data values within each block (see, e.g., Schrepen et al. 2017 for a recent overview). A typical application of these approaches is market structuring (see, e.g., Baier et al. 1997), where the data matrix describes to which extent customers judge brands as substitutable and where the approach develops groups of customers that judge groups of brands in a similar way. The clustering scheme in these approaches can be overlapping and/or non-overlapping in order to reflect application-specific challenges.

However, nowadays, even if the goal of the analysis goal still is to simultaneously deliver overlapping and/or non-overlapping groups with respect to two modes (e.g. groups of customers and groups of brands), the available data for developing the two-mode clusterings are of higher complexity: So, e.g., in the above discussed example the relationship between a customer and a brand could be described by ratings with respect to different attributes (e.g. suitability for different purposes), by buying or usage histories (e.g. buying intensities across periods) or by clickstream data (e.g. customer journeys across different information channels).

To deal with this problem, a new two-mode clustering approach basing on Baier et al.'s (1997) algorithm is introduced and discussed. Applications to online shopping and site engineering (see, e.g., Baier et al. 2017) are used for demonstrating the usefulness of the new approach.

References

Baier D, Rese A (2017): Improving the Ordering and Delivery Processes in Online-Fashion Shops: New Approaches to Integrate the Voice of the Customer. Paper presented at the Conference of the International Federation of Classification Societies (IFCS-2017). Tokyo, August 8-10.

Baier, D., Gaul, W., Schader, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring, in: Klar, R., Opitz, O. (Eds.), Classification and knowledge organization. Springer, Berlin, Heidelberg, 557-566.

Schepers, J., Bock, H.H., Van Mechelen, I. (2017): Maximal Interaction Two-Mode Clustering, Journal of Classification 34, 1, 49-75.

Dynamic Prediction of Propensity to Purchase by Landmark Modelling

Ilan Fridman Rojas, Aris Perperoglou, Berthold Lausen, Henrik Nordmark

Recent developments in the analysis of time-to-event data have allowed for methodological and predictive improvements on a number of fronts relative to Cox models. In particular, the landmark modelling framework of Van Houwelingen (2007) allows for the inclusion of several crucial extensions to the Cox model: incorporation of time-varying effects, gross violation of the proportional hazards assumption, and prediction of conditional probabilities given known survival/no re-purchase up to given time points.

In the current study we present a novel application of this landmarking methodology: the modelling of transactional data to predict propensity to purchase. This use case presents a number of challenges, including data sets of considerable size for which many current statistical models and tools no longer scale to, and recurrent events with high frequency and multiplicity, often with time-varying covariates and strongly time-dependent effects. The application of the landmarking approach to this domain yields challenges and results distinct to those faced in epidemiological use cases. We present the results of such an application to subsets of a data set from a retailer with 2m customers and 7 years of collected transactional data.

We compare a Cox model with the Andersen-Gill model to allow for time-varying covariates, and then further extend the analysis to two variations of landmark models, thereby extracting a measure of the time-varying effect of the covariates, and producing dynamic predictions of probability of re-purchase which condition on time elapsed since the last purchase. The resulting model for predicting propensity to purchase by landmark modelling is the first of its kind to the best of the authors' knowledge.

Recommending Travel Itineraries using Social Media

Radhika Gaonkar, Maryam Tavakol, Ulf Brefeld

Planning the next vacations is a complex decision problem. Many variables like the place(s) to visit, how many days to stay, the duration at each location, and the overall travel budget need to be controlled and arranged by the user. Automatically recommending travel itineraries would thus be a remedy to quickly converge to an individual trip that is tailored to a user's interests. Recommending travel itineraries is an instance of a constructive learning task where the structure of the output is not fully determined by the input. Viewing the task as an optimization problem, feasible configurations depend on user preferences and individual constraints such as the time and cost budget. Travel recommendation also relates to preference elicitation, where user preferences are learned interactively over time to personalize the experience. In this paper, we leverage social media, more explicitly photo uploads and their tags, to reverse engineer historic user itineraries. We propose a constructive recommender system that learns from these historic itineraries and suggests new trips that are tailored to a user's preference. Our solution grounds on Markov decision processes that capture the sequential nature of itineraries. Empirically, we observe that the predicted itineraries are more accurate than robust path planning algorithms. Here, the tags attached to the photos provide the factors to generate possible configurations and prove crucial for contextualizing the proposed recommender system.

Using Association Rules for Dynamic Updates of Personalized Recommendations

Nicolas Haubner

In this work we propose a hybrid recommender system which utilizes two well-known techniques to provide dynamic and personalized recommendations to a customer while she is still browsing the on-line platform.

Model-based collaborative filtering (CF) has been shown in research to be the most accurate algorithm family in recommender systems. However, those approaches require a model training phase. This step can take a significant amount of time as the algorithms are typically computationally expensive. Therefore, predictions of customer preferences are often pre-computed on a daily or weekly basis, making the system unable to dynamically react to new purchases or ratings. This can lead to missed opportunities for selling additional products directly after a purchase.

On the other hand, association rule mining (ARM) can make dynamic recommendations based on the current virtual shopping cart or a recent purchase. The technique is often used in market basket analysis to find sets of products which are frequently bought together. However, it is not suited as a stand-alone recommendation engine since the suggestions are not personalized and nothing can be shown in the case of an empty cart.

We propose a system which combines both model-based CF and ARM. The CF component pre-computes predicted preferences for each customer periodically, and the ARM component pre-computes association rules for each product and combinations of products. When a user logs on to the system, she is displayed suggestions from the CF component. When she adds an item to her cart, her recommendation list is updated using a weighted aggregation of the CF recommendations and those association rules which have the current shopping basket as an antecedent.

We expect two advantages in using this combination approach: First, the personalized knowledge about the current customer is not lost, as it would be if the system only displayed association rules after the cart has been updated. Second, the system is still able to react dynamically to the implicit feedback of adding an item to the cart. Those hypotheses are tested in an off-line simulation using a real-world dataset of past purchases.

Statistical Aspects of Machine Learning Methods 1

Consistency and Robustness Properties of Predictors Based on Locally Learned SVMs

Florian Dumpert

Support Vector Machines (SVMs) play a successful role in classification and regression in many areas of science. In the past three decades, much research has been conducted on the statistical and computational properties of support vector machines and related kernel methods. On the one hand, the consistency and robustness of the method are of interest. On the other hand, from a mathematical point of view, there is an interest in a method that can deal with many observations and many features. Since SVMs require a lot of computing power and storage capacity, various possibilities for processing large data sets have been proposed. One of them, called regionalization, divides the space of declaring variables into possibly overlapping domains in a data driven way and defines the function to predict output by the formation of locally learned support vector machines. Another advantage of regionalization: If the generating distribution in different regions of the input space has different characteristics, learning only one ”global” SVM may lead to an imprecise estimate. Locally trained predictors can overcome this problem. It is possible to show that a locally learned predictor holds consistency and robustness results under assumptions that can be checked by the user of this method. We will look at these consistency and robustness properties, such as bounds for the local and global maxbias or the local and global influence function for predictors based on locally learned support vector machines.

Classification in High Dimensions: When Are Rules Beneficial?

Claus Weihs

Theoretical results identify idealized situations where on the one hand the error rate of the linear discriminant analysis (lda) converges to the worst case error rate, i.e. the probability π of the smaller of two classes, for an increasing number of features *p* (Bickel & Levina, 2004). On the other hand, there are situations where the error rate of a fairly large class of classifiers converges to 0 for increasing *p* (Fan et al., 2010). This paper looks for rules to avoid the worst case error rate in practical situations. By means of simulations where relevant parameters like the number of features *p*, the correlation coefficient ρ, the block structure of the covariance matrix, the probability π, and the error rate *f* per dimension, are systematically varied in a compatible way, we show that the following rules of thumb should be followed:

- Avoid “complicated” classifiers: The independence rule (ir) might be adequate, the support vector machine (svm) should only be considered as an expensive alternative, which is additionally sensitive to noise factors.
- Look for stochastically independent dimensions and balanced classes.
- Only take into account features which influence class separation sufficiently. Variable selection might help, though filters might be too rough.
- Compare your result with the result of the data independent rule “Always predict the larger class”.

P. J. Bickel and E. Levina. Some theory for Fisher’s linear discriminant function, ”naive bayes”, and some alternatives when there are many more variables than observations. Bernoulli, 10:989–1010, 2004.

J. Fan, Y. Fan, and Y. Wu. High-dimensional classification. In T. Cai and X. Shen, editors, Highdimensional statistical inference, pages 3–37. World Scientific, New Jersey, 2010.

Data-Driven Robust Control Using Reinforcement Learning

Phuong Ngo, Fred Godtliebsen

This study proposes a robust control design method using reinforcement-learning for controlling partially-unknown dynamical systems under uncertain conditions. The method extends the reinforcement-learning area with a new learning technique that is based on the robust control theory. Unstructured uncertainties in the dynamical systems can be represented from the data as a Euclidean-norm bound in the kernel matrix of the approximated value function. By learning from the measurement data, the algorithm proposes actions that guarantees the stability of the closed loop system within the norm-bounded uncertainties. Control policies can be calculated by solving a set of linear matrix inequalities. The controller was benchmarked using simulations on a blood glucose model for patients with type-1 diabetes. Simulation results show that the proposed methodology is capable of safely regulates the blood glucose within a healthy level under the influence of carbohydrate intakes. The controller has also significantly reduced the post-meal fluctuation of the blood glucose. A comparison between the proposed algorithm and the existing linear quadratic optimal controller shows the improved performance of the closed loop system using our method and provides insights on how insulin doses should be chosen.

Statistical Aspects of Machine Learning Methods 2

Factor Selection and Tests for Independence of Nominal and Metric Variates, Marked Rank Statistics

Ulrich Müller-Funk, Stefanie Weiß

Factor selection implies two tasks. 1) Capture the interaction between target variable and predictors - the latter typically measured on different kind of scales. 2) Find a balance between selection fit and methodological resp. combinatorial complexity. Both tasks are hampered by the fact, that few operating figures and tests exist able to record dependencies among two differently scaled variables - let alone among blocks of factors. Binning all metric variables and dismissing orderings expressed by the ordinal ones, obviously, is not a solution. The other extreme is still worse, i.e. to ignore scaling altogether and treat statistical outcomes just as ”numbers”. Regarding a process model , it seems reasonable, to deal with factor selection in two subsequent steps. First solve task

1) in order to get a starting set of predictors and thereupon try to make good for neclecting interactions among predictors by adding ”boosting” factors. The second step will require domain specific knowledge. To realize the first one, some others suggested to make use of the Benjamini-Yekutieli procedure for multiple testing. We agree with that idea. What seems to be missing so far, are tests for independence dealing with univariate variates X,Y, X metric and Y categorical. A proposal made for binary Y is to split the sample into two - corresponding to the labels- and apply some (permutation- or rank-) two-sample test for equality of the underlying distributions. Statistical tests, however, are not just data-driven but require model assumptions that fail to hold true for that kind of sampling. Instead , we propose a type of nonparametric one - sample tests based on ”marked rank ststistics” generalizing signed rank statistics. The marks catch the label information in Y, the ranks are taken from X, the null-hypothesis of independence of X and Y broadens the hypothesis of symmetry. For a start, Y is assumed to be binary. The familiar distribution theory for signed rank statistics is generalized to the extent needed. Next, we turn to the Benjamini-Yekutieli approach. At the end of the paper, extensions to non-binary Y‘s are discussed and some open questions are outlined.

On the Influence of Margin Conditions on Rates of Localized Algorithms

Ingrid Blaschzyk

In view of large-scale applications learning methods that can handle big datasets efficiently became more interesting in the last years. A solution to address the latter is to consider local learning methods that learn on subsets of the entire datasets. One example are localized support vector machines (SVMs) which learn on small, spatially defined data chunks. These methods improve complexities in time and space compared to global SVMs. To ensure statistical guarantees assumptions which depend on the learning problem are needed. In the problem of binary classification which is, apart from regression, one of the most considered learning problems it turns out that margin conditions are suitable.

In this talk we present finite sample bounds and learning rates for localized SVMs under margin conditions and we show that the resulting rates match those known for global SVMs. Furthermore, we present for the simple histogram rule a partitioning based refinement technique that can be used to improve the statistical analysis of classification algorithms. Here, it turns out that using margin conditions which set the critical noise in relation to the decision boundary make it possible to improve the optimal rates proven for distributions without this margin condition and, in addition, to improve the rates of global SVMs.

Parallelizing Spectral Algorithms

Nicole Mücke, Gilles Blanchard

While kernel-based methods for solving non-parametric (direct or inverse) regression problems are attractive because they attain asymptotically minimax optimal rates of convergence, these methods scale poorly when massive data sets are involved. Large training sets give rise to large computational and storage costs. For example, computing a kernel ridge regression estimate needs inversion of an n-by-n- matrix, with n the sample size. This requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory, which becomes prohibitive for large sample sizes. For this reason, various methods have been developed for saving computation time and memory requirements. Among them is distributed learning:

We partition the given data set D into m disjoint equal-size subsets $D_1, \ldots, D_m$. On each subset D_j we compute a local estimator f_j, using a spectral regularization method coming from a large class (including in particular Kernel Ridge Regression, Gradient Descent and spectral cut-off). The final estimator for the target function f is obtained by simple averaging: $f_D := 1/m(f_1 + ... + f_m)$.

We show that minimax optimal rates of convergence are preserved if 1. the regularization parameter is chosen appropriately according to the global sample size and 2. m grows sufficiently slowly with the sample size, i.e., $m = O(n^a)$, with an upper bound for a strictly smaller than 1.

Both depend on the smoothness assumptions on f and the statistical dimension. Numerical studies on simulated data sets confirm our theoretical findings. These also show that (for fixed m) a hold-out strategy is reasonable for achieving adaptation this setting.

Statistical Learning with Imprecision 1

Learning Range-Query Predictors

Vitalik Melnikov, Eyke Hüllermeier

We study a generalization of the standard setting of supervised learning, in which we seek to learn a model in the form of what we call a range-query predictor (RQP). What we mean by this is a model that accepts partially specified attribute values as an input (query) and produces a set-valued prediction as an output, namely the set of values of the target variable that are possible under any precise input within the range specified by the query. Thus, while the training data is still assumed to be precise, like in standard supervised learning, the model induced from this data is applied and evaluated on data that is only partially specified, typically in the form of interval ranges (hence the term range-query). Apart from predictive accuracy measured by a suitable loss function on sets (that compares predicted with ground-truth output sets), an important aspect is efficiency: given a range query as input, the predictor should be able to compute the output set efficiently. We present a formalization of this problem as well first proposals of methods for tackling it.

As a motivation and application we make use of RQPs for supporting black-box optimization via (heuristic) search. In black-box optimization, the evaluation of the target function in a certain point (e.g., the performance of a machine learning algorithm with a certain parametrization) is often very expensive. Therefore, it is common to train so-called surrogate models, which can be used to produce comparatively cheap predictions of the target function. Replacing a standard surrogate model with an RQP, it is possible to predict the range of values of the target function in a certain part of the search space (range of the parameter space). This information can in turn be used to guide the search process.

Using Polynomial Errors-in-Variables Regression to Analyse Sequential Process Chains

Oliver Meyer, Claus Weihs

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 [1]), this can be studied by using Error-in-Variables models instead of ordinary regression models. In this paper, we want to discuss how to use polynomial Errors-in-Variables Regression 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 to ensure unbiased estimations along the process chain and discuss how the presented methods can be applied to other statistical learning techniques.

Partial Relational Clustering : A Thresholding Approach

Marie-Hélène Masson, Benjamin Quost, Sébastien Destercke

Clustering is a challenging task in machine learning, where equivalence classes are to be computed between objects. We consider here the case of relational data : the information at hand consists in a relational matrix of binary information indicating presence or absence of links between the objects. A hard clustering of the objects can be obtained by re-arranging rows and columns so as to express it as a block matrix. The relational matrix is consistent with a clustering if it is reflexive (diagonal terms are non-zero, each object being in the same cluster as itself), symmetric, and satisfies the transitivity constraint (i.e., if $R(i,j) = R(j,k)$, then $R(i,j) = R(i,k)$ as well). Contrary to the two former conditions, checking (or imposing) the latter is not straightforward.

In our work, we consider matrices of scores instead of mere links. The problem of computing a clustering then amounts to transform these scores into binary relations while ensuring reflexivity, symmetry and transitivity. Rather than determining a single binary matrix, we adopt a cautious behaviour and propose to compute a partial matrix by transforming high scores into ones, low scores into zeros, and considering intermediate scores as unknown relations. A score is deemed intermediate if it lies in a tolerance interval around the neutral score (e.g. 0.5 for probabilities). Our strategy aims at computing the smallest tolerance interval such that the partial matrix is relational.

Checking symmetry and transitivity requires to fill the partial matrix in with zeros and ones, according to the observed relations between objects. However, once this is done, some relations may still be missing (e.g., no relation was observed between an object and any of the others). In this case, the partial relational matrix expresses a partial clustering. Alternatively, once an appropriate number of clusters has been chosen, the set of complete clusterings which are consistent with this partial clustering may be computed. Note that according to the amount of missing relations, this computation may not be reasonable.

We realized experiments in which score matrices are obtained either by averaging probabilities of objects to belong to the same cluster, or the results of various clusterings obtained by resampling. The results show that by controlling the tolerance interval which determines missing relations, a good trade-off between clustering accuracy and information completeness can be achieved. In both cases, our method provides a way of identifying ambiguous objects, for which cluster membership is controversial. Future research directions include developing active learning strategies so as to complete the relational matrix, and applying our strategy to problems where the number of clusters is known beforehand.

Issues in the Context of Missing Values

Martin Spiess, Daniel Salfran

One popular strategy to handle the problem of missing values is the method of multiple imputation (MI) as proposed by Rubin (1987). If the technique to generate MIs is (confidence) proper in the sense of Rubin (1987, 1996) then statistical inference can expected to be valid (see also Tsiatis, 2006), i.e. estimators are approximately unbiased and the actual rejection rates of true null hypotheses are close to (or not substantially larger than) the nominal level. Whether or not an imputation technique is proper depends on properties of the technique itself, like the ability to properly account for all the uncertainty in the imputations, but also on the context. For example, if a parametric imputation model is proper if certain distributional assumptions are met, it may not be proper if these assumptions are violated, although Rubin (2003) claimed a ‘self-correcting’ property of MI in this case. Today many imputation methods are provided by commercial but also open source software packages like R (R Core Team, 2017) assuming that the missing mechanism is ignorable, which basically means that missing values are missing at random (MAR). For many of these techniques, however, it is not known if they are proper at all. In fact, we will present results from a simulation study implying that several available imputation techniques are not (confidence) proper in general. Even worse, the ‘self-correcting’ property, where biases are covered by overestimated variances, may only work in very small data sets. Since relationships beween incompletely observed and completed or completely observed variables are often unknown, flexible imputation techniques are needed. Thus a semi-/nonparametric imputation technique will be proposed that seems to work well in situations were other MI techniques fail. However, these flexible techniques require large data sets and, like all available MI techniques, are not designed to take into account all the available information sometimes needed to make an imputation technique proper. Thus, it will be emphasized that techniques need to be developed that allow consideration of precise and imprecise information even in small data sets.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Rubin, D.B. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91 (434), 473-489.

Rubin, D.B. (2003). Discussion on Multiple Imputation. International Statistical Review, 71(3), 619-625.

Tsiatis, A.A. (2006). Semiparametric Theory and Missing Data. New York: Springer.

Relational Data Analysis for Weakly Structured Information: Utilizing Linear and Binary Programming for Computing Supremum Statistics on Closure Systems

Georg Schollmeyer, Christoph Jansen, Thomas Augustin

In this work we are concerned with the statistical analysis of data with a partial ordinal character. Such type of data appear very naturally for example in the analysis of interval data, where one is interested in an actually unobserved variable x for which one only observes a lower bound l and an upper bound u, and one only knows that the unobserved x lies in the interval [l,u]. Then, one way to analyze such interval data is to look at the associated interval order, which is clearly only partial. Another form of partial ordinal data appears if the variable of interest can be observed precisely, but is genuinely only partially ordered. For example in multidimensional poverty analysis, the dimension education could be seen as only partially ordered when it comes to comparing different educational systems and different educational paths.

Specifically, in our contribution we are dealing with statistical tests in the spirit of generalizations of the Kolmogorov-Smirnov test to more weakly structured partially ordinal data situations. We look at statistical tests that use supremal probability differences over a family of events. One example is stochastic dominance for random variables with values in a partially ordered set (poset), where for checking dominance one has to build the supremum of probability differences over the closure system of all upsets of the underlying poset. Another example is nonparametric analysis of item impact and differential item functioning (DIF) in dichotomous item response data sets. In this setting, one is interested in determining subgroups of persons that are characterized by sets of solved or not solved items and that differ very much in that attribute that is under investigation w.r.t. DIF. This situation is very close to the task of mining interesting subgroups in the field of subgroup discovery (SD).

To compute the supremum statistic, one has to optimize a linear function over a family of sets, which in our case turns out to be a closure system (i.e., contains the whole ground set and is closed under arbitrary intersections) and is thus effectively describable with methods of lattice theory and formal concept analysis (FCA). The optimization can be done by formulating an appropriate binary program. For the case of checking dominance, one can show that the integrality constraints can be relaxed, such that the problem reduces to solving a linear program. In the case of DIF analysis, we use the language of FCA and the duality of formal concept extents and intents to efficiently formulate a binary program, for which much of the integrality constraints can be dropped.

Finally, for statistical inference, we analyze the distribution of the supremum statistic and obtain large deviation bounds by applying Vapnik-Chervonenkis theory, which exactly addresses the statistical analysis of supremum statistics over families of sets.

Statistical Learning with Imprecision 2

Density Estimation with Imprecise Kernels: Application to Classification

Sébastien Destercke, Guillaume Dendievel

When estimating densities using kernel estimation, it is well known that the end-result is highly sensitive to the choice of the kernel bandwidth. Given this, it seems natural to be robust at least against small variations of this value. A too low bandwidth will capture very local variations, while a too high bandwidth will provide a too smooth density. This is particularly true when the number of samples from which the density is estimated is low.

In this talk, we explore the problem of estimating lower and upper densities from imprecisely defined families of parametric kernels. After giving the general form of the problem, we demonstrate that computing those lower and upper bounds on specific families of kernels (triangular, Epanechnikov) can be done efficiently by relying on the specific features of those kernels.

Such methods then provide easy tools to deal with continuous variables in a non-parametric way when using imprecise probabilistic methods. To demonstrate the usefulness of our approach, we apply it to the imprecise version of naive Bayes classification methods, i.e., naive credal classifier. The results are promising, in the sense that they allow one to indentify hard-to-classify instances, without letting the imprecision epxlodes too quickly.

Estimation of an Imputation Model for Non-Ignorable Binary Missing Data

Angelina Hammon

Standard applications of multiple imputation (MI) techniques are based on the assumption that the data are Missing at Random (MAR). However, in many situations it seems very realistic that the missing values follow a Missing Not at Random (MNAR) mechanism. In this case, usual implementations of MI are not sufficient either and may lead to biased estimates. Therefore, the sensitivity of inferences based on multiply imputed data should ideally be assessed by additionally applying plausible MNAR models which reflect the missing data mechanism and incorporate it into the imputation model. We will present a selection modeling approach to multiply impute non-ignorable binary missing data in the framework of Fully Conditional Specification (FCS). The suspected MNAR mechanism will be considered and modeled during the imputation process by applying a censored bivariate probit model as imputation model. For allowing the consideration of hierarchical structures in the data during the imputation process, the model is expanded by a random intercept term. The focus of this paper lies on the technical details of the actual estimation of this joint model realized by applying different quadrature techniques. To assess the performance of the imputation procedure, different simulation studies with varying data generating processes were conducted. In addition, the robustness towards the choice of a suitable exclusion criterion – a crucial condition for the proper estimation of the model – is investigated more in detail.

Reliable Multi-class Classification based on Pairwise Epistemic and Aleatoric Uncertainty

Vu-Linh Nguyen, Sébastien Destercke, Marie-Hélène Masson, Eyke Hüllermeier

Classification algorithms are usually designed to produce point predictions (in the form of single class labels). In cases of uncertainty, however, it might be more desirable to provide imprecise (or indeterminate) set-valued predictions, e.g., in safety-critical applications (s.t medical diagnosis). Set-valued prediction classifiers have been developed following different approaches, including non-deterministic, conformal prediction and imprecise classifiers, to address the problem. Although the predictive abilities of (set-valued prediction) classifiers have been studied both theoretically and experimentally, giving the reasons for why a class should be included into or discarded from a set-value prediction seems to be challenging.

In this talk, we want to discuss various approaches to producing set-valued predictions, and more particularly to what features of the training setting they are sensitive to, i.e., what will tend to make their predictions more indeterminate? For example, should set-valued predictions be sensitive to the values of the instance to classify? to the number of data in the training set? to the closeness of the instance with respect to the decision boundaries? In particular, we will discuss the benefits of differentiating between the uncertainty that can be reduced by adding further examples (a.k.a. espistemic uncertainty), and the uncertainty that cannot (a.k.a. aleatory uncertainty).

How Valid is MAR Imputation under MNAR: Some Insights from Educational Research

Sabine Zinn

It is common practice that large-scale panel studies generate and provide imputed data sets to the scientific community for easing statistical inference. Examples are manifold such as the German Socio-Economic Panel (SOEP), the German National Educational Panel Study (NEPS), and the British Household Panel Survey (BHPS) providing imputed data sets with respect to missing wealth and income information. The idea of imputing whole data sets with a large number of variables (often far more than one hundred) is justified by the hope that in this way potential missing not at random (MNAR) patterns become approximately missing at random (MAR). Concretely, under MNAR the probability of a missing value in a variable depends on unobserved factors such as a confounder or the variable itself, while under MAR this probability depends on observed variables and variable values. By including as many variables as possible into the missing data model used for imputation those preparing the imputed data sets try to tackle distorting effects due to MNAR by increasing the correlation between observed and unobserved factors. However, though intuitively obvious, this approach lacks statistical foundation. To my knowledge, no study exists assessing its validity. Clearly, there are several studies dealing with the robustness of imputation models designed for MAR when the true pattern is MNAR. However, these studies are not targeted to settings with a lot of variables which are more or less correlated with one or several unobserved factors that are related to the statistics of interest. In this paper, I undertake a first step to close this gap. For this purpose, I study the monetary return to education as a typical research question relevant for educational researchers. I design a prototypical data example by taking data on missing and correlation patterns from the NEPS Starting Cohort on Adults. Here, the missing pattern of the income variable is MNAR since the probability of a missing income value depends demonstrably on the income variable itself. Via simulation I study under which correlation pattern approximating a MNAR model by a MAR one becomes feasible.

Statistical Visualization for Data Science

Visual Support for Imbalanced Classification

Adalbert Wilhelm

Visualisation techniques are widely used to successfully enhance the knowledge discovery in data base processes and a broad spectrum of visualisation approaches for classification problems exist. A core aim of applying visual techniques in the data analysis process is to ease the interaction between algorithms and the user. In general, there are three main stages in the analysis process, namely algorithm selection, model construction, and model evaluation, that can be enhanced by the use of appropriate visualisation techniques. In particular, context knowledge of the data owner is considered to be of vital importance and often provides a substantial input to improve modelling results. For this purpose, visual displays provide an important means of communication between content expert and data analyst. Many real life classification problems are characterised by the class imbalance between the various categories of the response variable, e.g. credit risk scoring, fraud detection, concession cases or military conflicts. These imbalances pose a particular challenge to visualisation that go beyond the mere question of image resolution.

In this talk we investigate how visualisation needs and can be adapted to support the further analysis of class-imbalanced data. Looking at continuous and categorical predictors we evaluate the visual comparability of graphical representations for imbalanced data at all three stages of the analysis process. Using the large tool kit of R graphics, we provide a taxonomy of the effect of class-imbalance on the visual displays for data exploration. Displays of model quality and model comparisons are evaluated with a focus on predictive models derived from ensemble methods. Here, a specific emphasis is put at the interplay between graphical representations and applying data level methods and algorithmic level methods to handle class imbalances.

Visualization of Cluster Detection Based on Hierarchical Structure for Geospatial Data and Its Application

Fumio Ishioka, Shoji Kajinishi, Koji Kurihara

With the remarkable development of GIS in recent years, it has become easier to analyze and visualize a wide variety of geospatial data. However, about an objective method to express the geospatial structure has not been fully discussed. Echelon analysis, which was originally proposed by Myers et al (1997), is an approach to be able to visualize the geospatial data systematically and objectively by a topological hierarchical structure according to the adjacency relationship of each region. As an example of data analysis using the echelon analysis effectively, a detection of spatial cluster can be mentioned. Spatial scan statistic (Kulldorff 1997) has been widely used for spatial cluster detection together with the freely available SaTScan(TM) software and applied in such fields as astronomy, biosurveillance, natural disasters, and forestry. This approach is based on the idea of finding a subset region which maximizes a likelihood function over the whole study region using search function called ”window”. In previous study, various types of windows have been proposed that devised the shape of the detected cluster, the reduction of calculation cost and so on. As an idea of window, we have proposed using the echelon analysis (Kurihara; 2002, Ishioka and Kurihara; 2007). By using the echelon-based window, clusters can be detected according to visual order based on the spatial structure of data. In addition, this method enables us to detect an arbitrary shaped cluster even when large amounts of data are targeted. In this study, we will describe application cases of echelon-based cluster detection and introduce a web application we are developing using R Shiny.

Development of an Interactive Visualization System to Analyze the Influence of Drug Resistance Appearance

Sanetoshi Yamada, Yoshiro Yamamoto, Kazuo Umezawa

In the medical field, we judge that antimicrobial drugs with large doses influence the emergence of drug resistance. And guidance such as suppression of administration is common. However, it is thought that it is difficult to suppress multiple occurrences with the antimicrobial drug administration suppression program by the simple dose. In addition, the antimicrobial drug to be administered is highly likely to be decided by the doctor by the knowledge and experience to the situation of the patient, and the medical practice may be hindered by suppression of antibiotic drug administration. In this paper, with respect to the relationship between the appearance of drug resistance and administered antibacterial drugs, ”Association plot that extracts characteristic specialized to attributes” which is a visualization obtained by combining association rule analysis which is a data mining method and correspondence analysis which is a multivariate statistical data analysis method by forming data proposed in Yamada and Yamamoto (2015) was applied. We analyzed of specific antimicrobial drugs that have great influence on the appearance of drug resistance and administration overview of antibacterial drugs overall and drug resistance appearance by this visualization.

In order to make it easier to interpret the influence of the appearance of drug resistance, we proposed an odds ratio which is easy to evaluate in the medical field as an evaluation index. We created a visualization application that can change the survey target, the lower limit of the association rule parameters and the size of the plot according to the request of the user by Sniny package of R. In order to confirm numerical results, we also displayed a table of plots and association rules at the same time. In order to make it easier for users to find the rules you want to pay attention to, we added a feature to reflect on the plot only the rules selected in the table by the DT package. In addition, we add the feature that can interactively manipulate plot by visNetwork package and propose a system that medical staff can refer to the results.

Empirical Study on Analysis of Unauthorized-Access Log Data and its Visual Output

Hiroyuki Minami

We have recently studied an analysis of unauthorized-access log data. The data themself have grown day after day, and it is too tough to analyze them as it is directly. Tons of studies are reported in the field of computer science and network engineering, however, we are wondering if only few are based on mathematical discussion. Originally, most cyber attacks must be done semi-automatically and it means that we could formalize a statistical model. To employ some domain knowledge on the Internet framework (known as TCP/IP and Classless Inter Domain Routing; CIDR), we have found some specific trends by preliminary analyses and empirical findings. We have an idea that we utilize them to introduce some collective analytical methods based on symbolic data analysis and develop an application and its analytical environment including data handling, analysis itself and accessible output. It is not a statistician but an expert in cyberspace that gives an interpretation to a result of analyses. Then, data representation including visual output is a key of further understanding for the experts, even if it is too common in statisticians. Besides, we should pay attention to real-time performance in ractical analysis, compared to our typical studies. In the study, we offer an empirical application for a practical analysis of unauthorized-access log data and the output representation. We investigate the trends and findings, introduce the way to utilize them and discuss the outputs from the substantial viewpoint.

Statistical and Econometric Methods

Adaptive Confidence Intervals for Kinks in Regression Functions

Viktor Bengs, Hajo Holzmann

Kinks of a regression function are irregularities which indicate for instance an abrupt change of some ongoing trend. Therefore it is of special interest to locate such irregularities or to determine a small interval which entails these kink-locations in order to analyze the cause of these change points. Such kinks correspond to jumps in the derivative of the regression function and there are various ways to estimate such discontinuities, though the literature is sparse concerning construction of confidence sets for kink-locations. In this talk we construct adaptive asymptotic confidence intervals for the location of kinks in a univariate regression setting, which are reasonably narrow and adapt to the particular smoothness of the regression function. To this end, we derive the asymptotic normal distribution of the estimator given by the zero-crossing-time-technique known in the statistical literature. The construction of the adaptive confidence intervals then is based on a Lepski-choice of the bandwidth-resolution in the zero-crossing-time-technique and an appropriate control of the bias. Through the Lepski-choice of the bandwidth-resolution the construction of the confidence sets is mainly data-driven. Its finite-sample performance is investigated in a simulation study with artificial data as well as some common datasets in this realm.

The Non-Gaussian ESEMIFAR Model

Yuanhua Feng, Sebastian Letmathe

The Gaussian ESEMIFAR (exponential SEMIFAR) introduced by Beran et al. (2015) is a suitable approach for simultaneously modelling a deterministic trend as well as short- and long-range dependence in a log-normal nonnegative process. The aim of this paper is to extend the ESEMIFAR to general log-linear processes without the normality requirement on the log-data. Under suitable additional assumptions on the innovation distribution we show that the Appell rank of the exponential transformation is one, analogously to the Hermite rank of a log-normal process. Hence, the long-memory parameter for the original and the log-transformed processes is the same. Asymptotic properties of our model are derived. We show that the data-driven SEMIFAR algorithm is applicable to our proposal. Forecasting based on the proposed model using the bootstrapping technique is developed as well. The usefulness of our proposals is illustrated by the application to different types of financial data.

Further Development of the Double Conditional Smoothing for Nonparametric Surfaces Under a Lattice Spatial Model

Bastian Schäfer, Yuanhua Feng

Nonparametric estimation of high-frequency financial data under a lattice spatial model has high demand on computation, due to the huge size of the data and the bivariate nature of the estimators. The double conditional smoothing offers a way to reduce complexity of estimation. In this paper, we first extend the double conditional smoothing by using boundary kernels and propose a much quicker functional smoothing scheme. Then we obtain the asymptotic formulas for the bias and variance, as well as the optimal bandwidths of this estimator under independent errors. An iterative plug-in algorithm for selecting the optimal bandwidths by double conditional smoothing is developed. Both, the Nadaraya-Watson estimator and local linear approaches are considered. The performance of the proposals is compared to the traditional bivariate kernel smoothing through a simulation study and further confirmed by application to financial data.

Identifying the Most Relevant Information in Skewed Distributions

Alfred Ultsch

Skewed distributions are very common in processes which are associated with concentration effects. Typical observations comprise, for example, wealth [1] income [2] and social justice [3]. Such distributions result from competition and/or marked situations involving humans or other organisms [4]. Skewed distributions are also found in data science where the “critical few” [5] parameter deliver most of the information. The canonical example is the selection of the most relevant dimensions in principal component analysis (PCA) or factor analysis. Typical tools for the analysis of such distributions are the Lorenz Curve and the Gini index. Here a more precise analysis in form of a data science based standardized Gini index as a Lorenz dominance preserving measure of the inequality of distribution is presented [6]. Concepts from economy, in particular cost vs effect considerations leads to a separation of the critical few (A class) parameters from the irrelevant (C-class) parameters. An approach called “Calculated ABC Analysis” was developed which delivers a precise mathematical calculation of the limits for the A and C classes [6] It uses an optimization of cost (i.e. number of items) vs. yield (i.e. sum of their estimated importance). Successful applications are presented for: feature selection for biomedical data [7], business process management [8] and bankruptcy prediction [9].

Bernoulli Mixture Models as a Part of Credibility Theory

Anne Sumpf

In this paper, we connect the Bernoulli mixture models with the Bühlmann-Straub model from actuarial science. With the Bühlmann-Straub model we have a distribution-free estimator for same risk measures like expected loss or Value-at-Risk. Based on four different designed data sets, we analyze the effects of distribution assumptions from the estimators of Bühlmann-Straub model, CreditRisk+, CreditMetrics and CreditPortfolioView for different risk measures like Value-at-Risk and expected shortfall.

Observed versus Unobserved Heterogeneity in Structural Equation Models: Cross-Country Data on Virtual-Try Ons

Alexandra Rese, Eleonora Pantano, Daniel Baier

Augmented reality applications such as virtual try-ons are well-known examples for recent successful improvements of online shops’ user interface (see e.g. Rese et al., 2017): Besides the traditional way of ordering clothing, the customer has the possibility to be photographed and displayed as a model on a (smartphone or tablet) screen. Then, he/she can try-on offered products in a convenient and entertaining way and – if desired – order them directly. A recent cross-country study by Pantano et al. (2017) has shown that many customers across different countries accept the virtual try-on as a useful alternative to their traditional way of ordering, but that country-specific differences have to be taken into account when designing country-specific virtual try-on applications.

However, nowadays, with more and more customers that shop online across borders (see, e.g., the success of Alibaba’s singles’ day among American and European online shoppers or the European Commission’s digital single market initiative to strengthen cross-border online shopping) the question arises whether it would be better to design transnational instead of country-specific virtual try-on applications. To answer this question, differences in acceptance of virtual try-on are examined relying on technology acceptance experiments with 318 customers from

Germany and from Italy. The data rely on Pantano et al. (2017)'s study, but in addition to an analysis of observed heterogeneity (differences with respect to observable variables like the customer's nationality), an analysis of unobserved heterogeneity was accomplished based on the finite mixture partial regression concept. Besides methodological insights into the usefulness of this approach, the paper reflects valuable implications for retailers' internationalization strategy.

Pantano, E., Rese, A., Baier, D. 2017. Enhancing the online decision-making process by using augmented reality: A two country comparison of youth markets. Journal of Retailing and Consumer Services, 38, September, 81-95.

Rese, A., Baier, D., Geyer-Schulz, A., Schreiber, S. 2017. How augmented reality apps are accepted by consumers: A comparative analysis using scales and opinions. Technological Forecasting and Social Change, 124, November, 306-319.

Textual Data Analysis and Digital Humanities

Text Broom: A ML-based Tool to Detect and Highlight Privacy Breaches in Physician Reviews: An Insight into Our Current Work

Frederik Simon Bäumer, Michaela Geierhos

In this paper, we present our current work on Text Broom, a software which aims to detect privacy breaches in user-generated texts and make them visible to prevent accidental disclosure of information. Text Broom combines computational linguistic methods of entity recognition with linguistic rules and patterns as well as gazetteers to detect and remove potential privacy violations in texts. Based on a deeper understanding of linguistic characteristics, Text Broom recognizes explicit and inherent private information (such as names, relationships). While possible threats to privacy are relatively obvious when a photo or location is shared, it is less obvious in cases of texts such as an email or review. Natural language allows it to share information in a variety of different ways, so that information can also be inherently present. This also results from the fact that privacy breaches do not necessarily result from a single word or sentence, but from a number of pieces of information disclosed in the text. Text Broom can be used both by the users and website operators. We understand the detection of privacy breaches as a multi-class text classification problem. As use case, public physician reviews were used, which contain sensitive information of authors, which endanger the anonymity and can have a negative influence on the sensitive physician-patient relationship. This holistic view of user-generated texts from the point of view of data protection is a novelty.

Analyzing the Spectrum of Free Verse Poetry by using Digital Methods

Burkhard Meyer-Sickendiek, Hussein Hussein, Timo Baumann

One of the most important explanations for modern art is the theory of aesthetic pleasure, which claims that the fluency of cognitive processing is the cause for the positive effect of aesthetic experience. On the other hand, many modern artists like Picasso or Schönberg complicated the processability of their works using processes of abstraction in order to prevent such automated or fluid forms of art comprehensibility. We will test these two poles of art experience by focusing on similar prosodic features of (dis)fluence in modern and post-modern poetry, starting from a more fluent to a more disfluent and choppy style. As modern poets overcame rhyme and meter, they oriented themselves in these two opposing directions, creating a free verse spectrum that calls for new analyses of prosodic forms. We will present a digital method for automatically analyzing this spectrum. We define and relate six classes of poetic styles (ranging from a very fluent parlando-style to a very disfluent and choppy kind of lettristic decomposition) by their gradual differentiation. Based on this definition, we present a model for automatic prosodic classification of spoken free verse poetry that uses deep hierarchical attention networks to integrate the source text and audio and predict the assigned class for every poem. We evaluate our model on a large corpus of German author-read post-modern poetry using the lyrikline corpus (www.lyrikline.org) in order to semi-automatically find additional poems that belong into one of our six classes. The manually classified corpus is small and hence the quality of intermediate representations is limited by data sparsity. We mitigate this problem by using pre-training methods to help bootstrap the generation of reasonable intermediate representations (i.e., we teach the model some notion of poetic language as a foundation before teaching it to differentiate styles). We also intend to analyze the corpus for clusters of outliers from our current classification in order to determine further and refine the existing classes using an iterative "human in the loop" approach. With regards to the neural network, we will link the text and speech streams using connectionist temporal classification for it to better relate auditory to textual information, and we will connect the (sequential) line and pause encodings of audio in order for the model to better normalize outspeaker specific (but not style-specific) characteristics. Our goal is the mutual benefit of the model (which requires human input) and the philological expert who will be able to quickly scan, analyze and browse vastly larger collections of poetry than has been possible in the past.

Big Data and Digital Humanities(?)

Jochen Tiepmar

Processing large and complex amounts of data can be a difficult challenge due to hardware and software restrictions. This issue is generally referred to as Big Data. This paper investigates the term Big Data in the context of text oriented digital humanities and in illustrates that in this context this is not necessarily a problem of big data but merely one of a lot of small data. This paper provides a guideline for researchers to relate their work to the concept Big Data, with the goal to potentially show them that they might be working on a task that can be considered as Big Data even if the data set itself is comparatively small. Focusing only on volume related data aspects may result in ignorance against a significant number of potentially interesting use cases. This paper also argues that interoperability is one of the most prominent Big Data issues in text oriented digital humanities.

Definition of Big Data Defining the term Big Data is not trivial. The most obvious defining factor is the size of a data set, but this property can not be applied universally and depends on the domain context and data type specific properties. Text volume can be measured as number of tokens/documents or byte. While a token or document count can often result in impressive and seemingly large numbers, the corresponding byte count is often not in an area that can be considered as large. Yet certain text mining analyses and use case specific circumstances may result in workflows that are already calculation expensive for technically small data sets.

IBM defined Big Data using the 4 V's: 4 data specific properties that can help to estimate the Big Data relevance of a specific problem. These V's are Volume, Veracity, Velocity and Variety.

The FinderApp WiTTFind for Wittgenstein's Nachlass

Maximilian Hadersbeck, Alois Pichler, Sabine Ullrich, Ines Röhrer

We present a new web-based approach to searching and researching Ludwig Wittgenstein's full philosophical Nachlass as made available by the Wittgenstein Archives at the University of Bergen (WAB), on Wittgenstein Source (http://www.wittgensteinsource.org/). The approach uses highly sophisticated web-technology together with methods and tools from the field of computational linguistics that are developed at the Centrum für Informations- und Sprachverarbeitung (CIS) at the LMU Munich. Tools include the full-form lexicon WiTTLex, the "FinderApp" WiTTFind, the symmetric autosuggestion tool SIS, a Facsimile Reader with hit-highlighting and an Investigation Mode with an integrated FeedbackApp.

As of summer 2017, WiTTFind (http://wittfind.cis.uni-muenchen.de/) offers lemmatized search access to the entire Wittgenstein Nachlass available at the Wittgenstein Archives at the University of Bergen (WAB, http://wab.uib.no/). Although the only complete electronic edition of the Nachlass hereto, WAB's Wittgenstein's Nachlass: The Bergen Electronic Edition (Oxford University Press, 2000), did offer a number of search and analysis tools which in part remain unmatched, it did not yet include lemmatized search. WiTTFind represents a significant advance on that and produces, upon entering the lemma of a word (i.e. the dictionary form, "Grundform"), a hit list with all forms of this word that occur in the Nachlass. WiTTFind's search capabilities include various search capabilities based on computational linguistic methods, like syntactic and semantic search in the field of colors and music, or a NLP-based method for similarity search. WiTTFind displays the word searched for within the context of the larger remark ("Bemerkung") and additionally highlights the hit in the corresponding facsimile of the remark. Moreover, WiTTFind is equipped with a separate Facsimile Reader which not only makes paging through Wittgenstein's Nachlass easy, but also contains a function for giving feedback to the editors - be it on the facsimile, the transcription or other parts of the resource - and comes with a text search and hit-highlighting functionality for the facsimile. WiTTFind is the result of more than five years of close cooperation between WAB and the CIS, the former contributing its facsimiles and encoded XML transcriptions of the Wittgenstein Nachlass, the latter programming and computational linguistics skills as well as a grammatically encoded digital lexicon of the German language. In our presentation we will demonstrate and describe WiTTFind and can, upon request, also discuss aspects of the cooperation between philosophers (WAB) and computational linguists (CIS) that may be useful for anyone interested in starting a digitally mediated (philosophical) cultural heritage collaboration, incl. the role of open access and linked data policies, aspects of communication between philosophers, philologists and programmers, work flows or also just the technical infrastructure.

Topic Detection and Classification in Consumer Web Communication Data

Atsuho Nakayama

The purpose of this study is to detect trending topics in web communications among consumers using social media by focusing on topics related to new products. In this study, we examined temporal variation in topics regarding new products by classifying words into clusters based on the co-occurrence of words in Twitter entries. To help identify market trends, analysis of consumer tweet data has received much attention. Twitter is an online social networking and microblog service that enables users to post and read tweets. We collected Twitter entries about new products based on their specific expressions of sentiment or interest. We tokenized each tweet message that was written in sentences or sets of words to detect topics more easily. In most Western languages, words are delimited by spaces and punctuation but words are not separated by spaces in Japanese. One of the most difficult natural language-processing problems in Japanese is tokenization. This is referred to as the wakachigaki problem. We have to do morphological analyses such as tokenization, stemming, and part-of-speech tagging to separate the words. We used the Japanese morphological analyzer ChaSen to separate words in passages and to distinguish all nouns, verbs, and adjectives. Next, we selected keywords representative of our chosen topics. To better understand topic characteristics, it was important to establish criteria to choose appropriate words representing temporal variation. We performed a statistical analysis based on the complementary similarity measure the measure has been widely applied in the area of character recognition, and was originally developed for the recognition of degraded machine-printed characters. To construct appropriate word-set topics each week, we estimated the associations within word pairs. Then, we detected trending topics related to a new product by classifying words into clusters based on the co-occurrence of words in Twitter entries. The matrix obtained from the Twitter entries are sparse and of high dimensionality, so we need to perform a dimensionality reduction analysis. We analyzed the matrix using Non-negative Matrix Factorization to reduce the dimensionality. Similar to principal component analysis, the method consists of positive coefficients in linear combination. The computation is based on a simple iterative algorithm, which is particularly useful for applications involving large, sparse matrices. We found each topic consisted of a small set of core words. The topics of Twitter entries were divided into two categories: those associated with reviews, and those associated with advertising. These topics were further classified by the characteristics of their core words. Next, we clarified temporal variation by using the weight coefficients which show the strength of associations between entries and topics. Personal concerns are influenced by new product strategies, such as marketing communication strategies, and they change over time. So, it is important to consider the temporal variation of these topics when detecting trending topics by classifying words into clusters based on co-occurrence of words. We found the topic transition patterns that were influenced by new product strategies, such as marketing communication strategies, and they changed over time.

Time Series Analysis and Online Algorithms

Time Series Study on Job Market Demand using Co-Word Analysis

Elisa Margareth Sibarani, Simon Scerri

The digitization of the economy and society has rapidly changed the job markets. Consequently, job advertisements are now mainly published electronically online and many new job profiles are emerging. Hence, the corresponding skills of those new job profiles are in high-demand while the previous important skills less relevant. In addition, the evolution in job markets also causes a continuous increase in new skills, especially in a fast-evolving sector like information technology. As a result, it raises new challenges for job seekers and education providers to accessing crucial information of which skills are now in high-demand. Therefore, our goal is to provide job seekers and education providers a timely and thorough overview of the current situation on the job market concerning the required skills, competencies, and technologies. The comprehensive overview is based on an objective and practical review and its tendencies and safe near-future predictions according to specific time periods.

In this paper, we describe efforts to obtain a picture of the demand landscape by understanding the employer needs and identify skill demand composition and its dynamics, so as to guide aspiring job seekers and educators alike. We utilize an Ontology-based Information Extraction (OBIE) method to identify the list of skills requested by the job vacancies. This method relies on the Skills and Recruitment Ontology (SARO) to representing job postings covering skills and competencies context. Skill demand by employers is then analyzed and presented using co-word analysis based on the co-occurrences of skill keywords in the job posts. We further generate the constellation of the correlated skills (skills network) and identifies the position of a particular skills network in the strategic diagram based on the internal strength of the network and the degree of interaction with another network.

Next, we implement time-series analysis by relying on a series of observations over several time periods. Firstly, we identify crossroads sub-network, by taking all networks which have a strong association with other networks. Secondly, we establish series of networks for different time periods as the building block for time-series analysis based on the derived similarity indices, whereby each series contains at least one crossroads. Finally, on top of series of networks, we further analyze skills theme evolution through the changes in the strategic diagram over different periods. We present some significant network examples in order to illustrate the method and its interest. The goal is to determine which skills theme played a role based on the observation at time periods; and which skill networks that are at no point central to the structure of the general network, or that is continuously less relevant over some periods.

Therefore based on the assessment of the demanded technical skills, we provide our targeted users with a useful guidance because: (1) these findings could direct job seekers toward striving for the skills they should acquire to raise their competitiveness; and (2) the consequences of this new insight could also support the educators to adequately prepare courses in respect to the skills that are in high demand.

Online Adaptable Time Series Anomaly Detection with Discrete Wavelet Transforms and Multivariate Gaussian Distributions

Markus Thill, Wolfgang Konen, Thomas Bäck

Up till today, anomaly detection in general and especially for time series is remaining a challenging task. One reason for this lies in the fact, that anomaly detection algorithms typically have the goal to identify anomalies in an unsupervised or self-supervised manner since the possible anomalous conditions or patterns are generally not known beforehand. In order to achieve this goal, suitable features are required to learn to distinguish between normal and anomalous behavior in the data.

In this paper we present an unsupervised time series anomaly detection algorithm, which is based on the discrete wavelet transform (DWT) operating fully online. Given streaming data or time series, the algorithm can iteratively compute the (causal and decimating) discrete wavelet transform. For individual frequency scales of the current DWT, the algorithm estimates the parameters of a multivariate Gaussian distribution. These parameters are adapted in an online fashion. Based on the multivariate Gaussian distributions, unusual patterns can then be detected in each frequency scale. The unified view detects unusual patterns appearing simultaneously in different frequency scales. This signifies an anomaly in the original time series and will cause the algorithm to flag the current data point accordingly.

The algorithm is tested on a diverse set of 425 time series taken from five datasets (each containing 58–100 time series) of the Yahoo Webscope S5 Anomaly Benchmark and the Numenta Anomaly Benchmark. A comparison to several other online state-of-the-art anomaly detectors shows that our algorithm can mostly produce results similar to the best algorithm on each dataset. It produces the highest average F1-score across the five datasets with one standard parameter setting. That is, it works more stable on high- and low-frequency-anomalies than all other algorithms. We believe that the wavelet transform is an important ingredient to achieve this.

Comparison of Machine Learning Approaches for Time-Series Based Quality Monitoring of Resistance Spot Welding

Baifan Zhou, Tim Pychynski, Markus Reischl, Ralf Mikut

In automatic manufacturing, enormous amounts of data are generated everyday. However, labeled production data useful for data analysis are difficult to acquire. Resistance spot welding (RSW), widely applied in automobile production, is a typical automatic manufacturing process with inhomogeneous data structures, statistical and systematic dynamics. In resistance spot welding, an electrical current goes through electrodes and the materials in between. The materials are heated and melted down. After that, the materials congealed, forming a contact area, namely the welding nugget. The nugget size is an important quality measure, but can only be precisely obtained by costly destructive methods. This paper strives to address the issue of scarcity of labeled data using simulation data generated with a verified finite element model. Simulation is advantageous in that it enables producing large amounts of labeled data with fewer limits on sensors and cost. Based on the simulation data, this paper explored and compared multiple machine learning methods, achieved prediction of the nugget size with high accuracy, and conducted an analysis of the influence of feature numbers and the amount of training data on prediction accuracy.

Music Generation with Long Short Term Memory

Amin Dada, Rolf P. Würtz

Forecasting a time series in a reasonable and creative way is a central task for artificial intelligence. Here we apply it to learning a certain musical style from examples and compose convincing new ones.

We used the framework tf-seq2seq, which implements an encoder-decoder model. Both encoder and decoder are recurrent neural networks of the long short term memory type. The encoder was chosen to be bidirectional, i.e., reading input sequences in both natural and reverse order.

Training was done on MIDI-encoded versions of chorales by Johann Sebastian Bach. All music was quantized to sixteenth notes and transposed to either C major or A minor, depending on an automatic scale analysis, and then converted to text strings. Each chorale was divided into sequences of four bars. Sequences were used as input and the network was trained to produce the directly following sequence.

Training was done using Adam optimization and dropout to avoid overfitting. Further techniques were teacher forcing and gradient clipping. The best performing network architecture regarding test loss had 256 bidirectional LSTM cells in 2 layers.

Like in many AI tasks the results of the algorithm cannot be evaluated automatically but only by inspection through humans. To that end we created an online questionnaire, in which people were presented 5 pairs of an original and one generated piece and had to decide which of both was artificial. Participants could rate themselves as ”music listeners” (group 1), ”players of an instrument” or ”professional musicians” (group 2). Artificial examples were drawn randomly from 50 handselected pieces out of 150. Both groups performed only slightly above chance (51%) on this task. It can be concluded that the network was capable of capturing some of the ”musical spirit” in the training samples

An Experimental Evaluation of Time Series Classification Using Various Distance Measures

Paweł Piasecki, Tomasz Górecki

In recent years, a vast number of distance measures for time series classification has been proposed. Obviously, the definition of a distance measure is crucial to further data mining tasks, thus there is a need to decide which measure should we choose for particular data set. The objective of this study is to provide a comprehensive comparison of over 30 distance measures enriched with extensive statistical analysis. We compare different kind distance measures: shape-based, edit-based, feature-based and structure-based. Experimental results carried out on all benchmark data sets from UCR Time Series Classification Archive (Chen et al. 2015) are provided. We use 1-NN classifier to compare efficiency of examined measures. Computation time is taken into consideration as well.

Using Time Series Analysis for Predicting Hemodynamic Instability in Intensive Care Patients

Daniela Behnam, Mark Last

Hemodynamic instability is a clinical condition indicating abnormal or unstable blood pressure in a critical care patient (Weil, 2005). It is often not recognized and hence not treated on time with resulting damage to organs or even death. Particularly, hemorrhagic shock (often used as alternative term for hemodynamic instability) is associated with nearly two million yearly deaths worldwide (Cannon, 2018). Earlier detection of hemodynamic instability will allow earlier intervention, faster recovery, contribute to better patient outcomes, and save lives. In addition, it will reduce hospitalization costs by shortening hospital stays.

Clinicians believe that physiological signals derived from intensive care monitors can imply clinical deterioration of a patient well before the start of a hemodynamic instability episode. Our goal is to apply machine learning techniques to the monitoring data for predicting a hemodynamic instability episode at least one hour in advance. Feature extraction is based on continuous time series analysis. From each time series, we calculate statistical measures (moments, extremes, and percentiles), linear trend, wavelet transform coefficients, shapelets (Grabocka et al, 2014), etc. We use several physiological variables including three different BP (blood pressure) measurements, HR (heart rate), RR (respiratory rate), and SpO2 (oxygen saturation). The retrospective data was collected from the General ICU (Intensive Care Unit) in Soroka Medical Center located in Beer-Sheva, Israel over the 2008-2016 period. We identified 3,839 admissions of adult trauma patients who had sufficient monitoring data. For each admission record, we extracted several time windows that were labeled as stable / unstable according to the patient condition at the end of the following hour.

We used nested 5-fold cross validation and applied the gradient boosting classification algorithm (Chen et al., 2015). The results show the area under ROC curve of 0.92 ± 0.005, sensitivity of 0.76 ± 0.02, and specificity of 0.90 ± 0.008. Presumably, the proposed methodology can support clinical decision making by providing an early alert of an impending hemodynamic instability.

Cannon, J. W. (2018). Hemorrhagic Shock. New England Journal of Medicine 378:4:370–379.

Chen, T., He, T., & Benesty, M. (2015). XGBoost: Extreme gradient boosting. R package version 0.4-2:1-4.

Grabocka, J., Schilling, N., Wistuba, M., & Schmidt-Thieme, L. (2014, August). Learning time-series shapelets. Proc. of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 392-401). ACM.

Weil, M. H. (2005). Defining hemodynamic instability. Functional hemodynamic monitoring (pp. 9-17). Springer, Berlin, Heidelberg.

Web & Data Science

A Simple and Fast Approach to Knowledge Graph Embedding

Tommaso Soru, Stefano Ruberto, Diego Moussallem, Edgard Marx

Knowledge Graph Embedding methods aim at representing entities and relations in a knowledge base as points or vectors in a continuous vector space. Several approaches using embeddings have shown promising results on tasks such as link prediction, entity recommendation, question answering, and triplet classification. However, distributional semantics techniques are relatively new in the Semantic Web community. Only a few methods can compute low-dimensional embeddings of very large knowledge bases. HolE [2] and the RDF2Vec approach [3] are examples of pioneering research. Nevertheless, the computation of embeddings on large graphs is known to be expensive with regard to runtime and required memory and, to date, RDF2Vec is the only option for learning embeddings on datasets as large as DBpedia [1]. To this end, we devise a simple and fast approach dubbed KG2Vec, which exploits the skip-gram model to create embeddings on large knowledge graphs in a feasible time but still maintaining the quality of state-of-the-art embeddings.

Existing KGE approaches based on the skip-gram model such as RDF2Vec submit paths built on random walks to a word embedding algorithm. Instead, we preprocess the input knowledge base by converting each triple into a small sentence of three words. Our method is thus faster as it allows us to avoid the path generation step.

We investigate the problem of link prediction using the skip-gram model. Current methods rely on analogies to discover new relationships among entities. Instead of using a predefined scoring function, we learn it using Long Short-Term Memories and show that it overperforms the analogy-based method on a real-world dataset. Our preliminary qualitative evaluation shows that KG2Vec might achieve a better quality than state-of-the-art approaches in terms of similarity search and can scale to large graphs, processing more than 100 million triples in less than 6 hours on common hardware.

The KG2Vec source code and data are available online [4].

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In The Semantic Web (pp. 722-735). Springer, Berlin, Heidelberg.

[2] Nickel, M., Rosasco, L., Poggio, T. A. (2016, February). Holographic Embeddings of Knowledge Graphs. In AAAI (pp. 1955-1961).

[3] Ristoski, P., Paulheim, H. (2016, October). RDF2vec: RDF Graph Embeddings for Data Mining. In International Semantic Web Conference (pp. 498-514). Springer, Cham.

[4] URL: https://github.com/AKSW/KG2Vec

CostFed: Cost-Based Query Optimization for SPARQL Endpoint Federation

Alexander Potocki, Muhammad Saleem, Tommaso Soru

The runtime optimization of federated SPARQL query engines is of central importance to ensure the usability of the Web of Data in real-world applications. Two challenges must be addressed when optimizing federated query processing. The first is the generation of efficient query plans: For a given query, there are most likely several possible plans that a federation system may consider executing to gather results. These plans have different costs in terms of the amount of resources they necessitate and the overall time necessary to execute them. Detecting the most cost-efficient query plan for a given query is hence one of the key challenges in federated query processing. The second is the optimized selection which is one of the key steps towards the generation of efficient query plans. A poor source selection can lead to increases in the network traffic, the number of intermediate results and the overall query processing time.

Current cardinality/cost-based SPARQL endpoint federation approaches address the first challenge by making use of average selectivities to estimate the cardinality of triple patterns. Hence, they assume that the resources pertaining to a predicate are uniformly distributed. However, previous work shows that real RDF datasets are not well-structured, i.e., they do not actually abide by uniform frequency distributions.

To address the second challenge, most SPARQL query federation approaches rely on a triple pattern-wise source selection (TPWSS) to optimize their source selection. The goal of the TPWSS is to identify the set of sources that are relevant for each individual triple pattern of a query. However, it is possible that a relevant source does not contribute to the final result set of a query. This is because the results from a particular data source can be excluded after performing joins with the results of other triple patterns contained in the same query. The join-aware TPWSS strategy has been shown to yield great improvement.

In this work, we present CostFed, an open-source, index-assisted SPARQL endpoint federation engine. CostFed addresses the two challenges aforementioned: CostFed's query planning is based on estimating query costs by using selectivity information stored in an index. In contrast to the state of the art, CostFed takes the skew in distribution of subjects and objects across predicates into account. In addition, CostFed includes a novel trie-based source selection approach which is a join-aware approach to TPWSS based on common URI prefixes. Overall, our contributions are as follow:

(1) We present a novel source selection algorithm based on labelled hypergraphs, which makes use of a novel type of data summaries for SPARQL endpoints based on most common prefixes for URIs.

(2) We propose a cost-based query planning approach which makes use of cardinality estimations for (a) triple patterns as well as for (b) joins between triple patterns. We considered the skew in resource frequency distribution by creating resource buckets with different cardinality estimation functions.

(3) Our results show that we outperform state-of-the-art SPARQL endpoint federation engines on LargeRDF-Bench.

CostFed is open-source and available online at https://github.com/AKSW/CostFed.

Benchmarking Cloud Services for the Internet of Things

Kevin Grünberg, Wolfram Schenck

The Internet of Things (IoT) connects sensors, machines, cars, household appliances, and other physical items for quick and worldwide data exchange. IoT devices typically generate and transmit data at regular intervals. This data is integrated, processed, analyzed (e.g., through data mining or machine learning pipelines), and monitored. In turn, commands for corrective actions are generated either manually or automatically and are sent back to the devices. In the context of ”Industry 4.0”, the IoT approach is also used to create networks of machines in assembly lines. Many application scenarios in the IoT rely on a central cloud service, and — especially in the industrial area — on operations taking place in close to real-time: For example, when manufacturing machines are controlled via apps running in the cloud, or when these machines use the cloud as central instance for data exchange to coordinate their work in an assembly line. This requires small round-trip times (RTT) in the communication between device and cloud. For this reason, two of the major IoT cloud services, Microsoft Azure IoT Hub and Amazon Web Services IoT, were benchmarked in the area of North Rhine-Westphalia (Germany) regarding RTT, varying factors like time of day, day of week, location, inter-message interval, and additional data processing in the cloud. The results show significant performance differences between the cloud services and a considerable impact of some of the aforementioned factors, especially regarding the inter-message interval and additional processing. In conclusion, as soon as (soft) real-time conditions come into play, one has to carefully plan the data flow to the cloud and to carry out corresponding benchmarks because it may depend strongly on the cloud service provider if specific real-time conditions for IoT data processing are met in the end.

Sensitivity Analysis with FANOVA Graphs

Sonja Kuhnt

Sensitivity analysis aims at identifying the input variables of a given model or function that have the highest impact on an output of interest. Global sensitivity analysis has broad applications in screening, interpretation and reliability analysis. A well-established approach is the estimation of Sobol indices to quantify the influence of variables, or groups of variables, on the variability of an output. First-order Sobol indices and closed Sobol indices quantify the single influence of variables and groups of variables respectively. We consider the so-called total interaction index (TII), which measures the influence of a pair of variables together with all its interactions. It provides a deeper insight into the interaction structure of the unknown function, displayed in a so-called FANOVA graph. Besides increasing the knowledge about the unknown black-box function, the results of the sensitivity analysis can be used within modeling and optimization. One field of application are Gaussian process emulations of computer experiments.

Fruth, J., Roustant, O., Kuhnt, S. (2014). Total Interaction Index: A Variance-based Sensitivity Index for Second-order Interaction Screening. Journal of Statistical Planning and Inference 147, 212–223.

Fruth, J., Roustant, O., Kuhnt, S. (2017). Support indices: Measuring the effect of input variables over their support. https://hal.archives-ouvertes.fr/hal-01113555.

Mühlenstädt, T., Roustant, O., Carraro, L., Kuhnt, S. (2012). Data-driven Kriging models based on FANOVA-decomposition. Statistics and Computing, 22, 723-738.

Saltelli, A., Chan, K., Scott, E. (2000). Sensitivity analysis. Wiley series in probability and statistics, Wiley, Chichester.

Sobol, I. (1993). Sensitivity estimates for nonlinear mathematical models. Mathematical Modelling and Computational Experiments 1, 407-414.

Using Textual Features for Validating RDF Knowledge Bases

Zafar Habeeb Syed, Michael Röder, Axel-Cyrille Ngonga Ngomo

With the increasing uptake of knowledge graphs in domains as diverse as question answering and geo-spatial information systems comes an increasing need for validated the knowledge contained in these graphs. However, the sheer size and number of knowledge bases used in real-world applications makes manual fact checking impractical. In this paper, we present FactCheck, an automated fact checking approach based on DeFacto [1].

DeFacto takes an RDF statement as input and returns a score expressing the correctness of it by analysing textual evidences. DeFacto relies on a web search engine to retrieve documents providing proof for the fact. The extracted proofs are assigned a numerical score between 0 and 1 using a linear regression classifier. For the document of a proof a trustworthiness score is computed. This is mainly based on topic words extracted for the subject and object of the fact [2]. Based on several features—mainly the scores of all proofs of the fact and the trustworthiness—a classifier decides whether the given fact is true or not.

In FactCheck we extend DeFacto in four ways. Firstly, we replace the web search engine with a local search engine using a reference corpus since web search APIs have limitations on their usage. Secondly, we improve the proof extraction using coreference resolution [3]. Thirdly, we use dependency parse trees to make sure that the extracted proofs are confirming the fact. Fourthly, we use word set coherences described in [4] to determine topic words instead of the approach of [2].

We evaluated our approach on two different benchmark datasets and two different corpora. Our results show that FactCheck outperforms the state of the art by up to 13.3% in F-measure and 19.3% AUC. FactCheck is open-source and available at https://github.com/dice-group/FactCheck.

[1] Gerber, D., Esteves, D., Lehmann, J., Bühmann, L., Usbeck, R., Ngomo, A.C.N., Speck, R.:Defacto—temporal and multilingual deep fact validation. Web Semantics: Science, Services and Agents on the World Wide Web 35, 85–101 (2015)

[2] Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Tezuka, T., Oyama, S.,Tanaka, K.: Trustworthiness analysis of web search results. Research and advanced technology for digital libraries pp. 38–49 (2007)

[3] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The StanfordCoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics:

system demonstrations. pp. 55–60 (2014)

[4] Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. pp.399–408. ACM (2015)

ARCHIVES OF DATA SCIENCE SERIES A

www.ArchivesofDataScience.org