ulm university universität
uulm

Statistical Computing 2019
Abstracts der 51. Arbeitstagung

HA Kestler, M Schmid, L Lausser,
A Fürstberger (eds)

Ulmer Informatik-Berichte

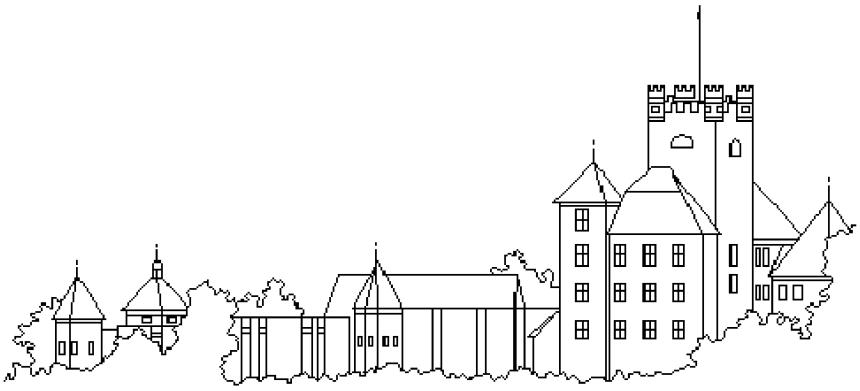
Nr. 2019-01
June 2019

International Graduate School
in Molecular Medicine Ulm

Sponsored by

International Graduate School in Molecular Medicine Ulm (one invited speaker)

Statistical Computing 2019



51. Arbeitstagung

der Arbeitsgruppen Statistical Computing (GMDS/IBS-DR), Klassifikation und Datenanalyse in den Biowissenschaften (GfKl).

30.06. - 03.07.2019, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, June 30, 2019

		Introduction: H. A. Kestler
20:00 – 21:00	Joachim M. Buhmann (Zürich)	Can I believe what I see? - Information theoretic algorithm validation

Monday, July 01, 2019

08:50		**Opening of the workshop: H. A. Kestler**
09:00 – 10:20		**Chair: L. Lausser**
09:00 – 09:20	Alfred Ultsch (Marburg)	ESOM Sampling as a Tool for Detection of Needles in the Haystack of Big Data in Medical Diagnostic Technologies
09:20 – 09:40	Marcus Vollmer (Greifswald)	A Convolutional Neural Network for ECG Annotation as the Basis for the Classification of Cardiac Rhythms
09:40 – 10:00	Shekoufeh Gorgi Zadeh (Bonn)	Uncertainty-Guided Semi-Automated Editing of CNN-based Retinal Layer Segmentations in Optical Coherence Tomography
10:00 – 10:20	Dila Ram Bhandari (Kathmandu)	Big data: Challenges, Tools and Techniques in SAARC Region
10:20 – 10:50		**Coffee Break**
		Introduction: A. Mayr
10:50 – 11:30	Annika Hoyer (Düsseldorf)	A bivariate time-to-event model for the meta-analysis of full ROC curves
09:00 – 10:20		**Chair: A. Mayr**
11:30 – 11:50	Jan Feifel (Ulm)	Enough is as good as a feast: Comparing different subsampling designs for time-to-event data
12:00 – 13:30		**Lunch**
		Introduction: M. Vollmer
13:40 – 14:20	Nadja Klein (Berlin)	A Novel Spike-and-Slab Prior for Effect Selection in Distributional Regression Models
14:20 – 15:00		**Chair: M. Vollmer**
14:20 – 14:40	Jörn Lötsch (Frankfurt am Main)	Machine learning supported hypothesis and pattern finding in pain related phenotype data
14:40 – 15:00	Patrick Thiam (Ulm)	Pain Intensity Recognition via Deep Physiological Models
15:00 – 15:30		**Coffee Break**

Monday, July 01, 2019

15:30 – 16:30		**Chair: A. Ultsch**
15:30 – 15:50	Sebastian Krey (Köln)	Graphical User Interfaces for Surrogate Model-Based Optimization in Practice and Teaching
15:50 – 16:10	Robin Szekely (Ulm)	Correlation-based feature selection utilizing foreign classes
16:10 – 16:30	Roman Hornung (München)	Improved outcome prediction across data sources through robust parameter tuning
16:30 – 16:50		**Break**
16:50 – 17:50		Working group meeting on **Statistical Computing 2020**
18:00 – 20:00		**Dinner**

Tuesday, July 02, 2019

09:00 – 10:20		**Chair: J. Kraus**
09:00 – 09:20	Florian Pfisterer (München)	Towards Human-Centered AutoML
09:20 – 09:40	Christian Staerk (Bonn)	Boosting with random selection of weak learners for variable selection in high-dimensional models
09:40 – 10:00	Steffen Moritz (Köln)	imputeTS: Tidy Univariate Time Series Imputation
10:00 – 10:20	Colin Griesbach (Erlangen)	Joint Modelling approaches to survival analysis via likelihood-based boosting techniques
10:20 – 10:50		**Coffee Break**
		Introduction: M. Schmid
10:50 – 11:50	Christiane Fuchs (Bielefeld)	Tackling leukemia through computational statistics
12:00 – 13:30		**Lunch**

Tuesday, July 02, 2019

		Introduction: S. Krey
13:40 – 14:20	Nikolaus Umlauf (Innsbruck)	bamlss: A Lego toolbox for flexible regression models
14:20 – 15:00		**Chair: S. Krey**
14:20 – 14:40	Xudong Sun (München)	Automatic Machine Learning and (Deep) Reinforcement Learning with rlR package
14:40 – 15:30		**Poster teaser + session & Coffee break**
15:30 – 16:30		**Chair: A. Fürstberger**
15:30 – 15:50	Jörn Lötsch (Frankfurt an Main) Alfred Ultsch (Marburg)	Generative artificial intelligence based algorithm to increase the predictivity of preclinical studies while keeping sample sizes small
15:50 – 16:10	Hryhorii Chereda (Göttingen)	Utilizing molecular network information via Graph Convolutional Neural Networks to predict metastatic event in breast cancer
16:10 – 16:30	Lisa Schäfer (Ulm)	Extracting ordinal class sequences from high-dimensional datasets
16:30 – 16:50		**Break**
		Introduction: A. Fürstberger
16:50 – 17:50	Erin LeDell (California)	Scalable Automatic Machine Learning with H2O
18:00 – 20:00		**Dinner**
20:00 – 21:00	Erin LeDell (California)	Hands-on Tutorial: Scalable Automatic Machine Learning with H2O

Wednesday, July 03, 2019

09:00 – 10:20		**Chair: R. Schuler**
09:00 – 09:20	Tobias Hepp (Bonn)	Adaptive step-lengths in model-based gradient boosting algorithms for distributional regression
09:20 – 09:40	Sina Mews (Bielefeld)	A continuous-time capture-recapture model for the annual movement of bottlenose dolphins
09:40 – 10:00	Ralph Brinks (Düsseldorf)	Simulation of trajectories in the illness-death model for chronic diseases: discrete event simulation and Doob-Gillespie algorithm
10:00 – 10:20	Thomas Welchowski (Bonn)	A Classification Tree Approach for the Modeling of Competing Risks in Discrete Time
10:20 – 10:50		**Coffee Break**
10:50 – 11:50		**Chair: W. Adler**
10:50 – 11:10	Lea Siegle (Ulm)	Semantic feature selection for multi-class classifier systems
11:10 – 11:30	Carlo Maj (Bonn)	Modeling strategies to dissect the variable genetic architecture in the computation of polygenic risk score
11:30 – 11:50	Christoph Molnar (München)	Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition
12:00 – 13:30		**Lunch**
13:30		**Departure**

Contents

Can I believe what I see? - Information theoretic algorithm validation

Joachim M. Buhmann[1]

Data Science promises us a methodology and algorithms to gain insights in ubiquitous Big Data. Sophisticated algorithmic techniques seek to identify and visualize non-accidental patterns that may be (causally) linked to mechanisms in the natural sciences, but also in the social sciences, medicine, technology, and governance. When we use machine learning algorithms to inspect the often high-dimensional, uncertain, and high-volume data to filter out and visualize relevant information, we aim to abstract from accidental factors in our experiments and thereby generalize over data fluctuations. Doing this, we often rely on highly nonlinear algorithms.
This talk presents arguments advocating an information theoretic framework for algorithm analysis, where an algorithm is characterized as a computational evolution of a posterior distribution on the output space with a quantitative stopping criterion. The method allows us to investigate complex data analysis pipelines, such as those found in computational neuroscience, neurology, and molecular biology. I will demonstrate this concept for the validation of algorithms using the example of a statistical analysis of diffusion tensor imaging data. In addition, on the example of gene expression data, I will demonstrate how different spectral clustering methods can be validated by showing their robustness to data fluctuations and yet sufficient sensitivity to changes in the data. All in all, an information-theoretical method is presented for validating data analysis algorithms, offering the potential of more trustful results in Visual Analytics.

[1] Institute for Machine Learning, D-INFK, ETH Zurich

jbuhmann@inf.ethz.ch

ESOM Sampling as a Tool for Detection of Needles in the Haystack of Big Data in Medical Diagnostic Technologies

Alfred Ultsch[1], Jörg Hoffman[2] and Cornelia Brendel[2]

In particular, within the context of molecular medical research data sets become larger and larger. High data volumes obtained with flow cytometric analyses of blood and tissue samples with real time multiparameter measurements were always a challenge for computer hard and software designers. Today, a regular Flow Cytometry [1] data sat for one single patient typically contains d ($10 < d < 100$) variables for $n > 1,000,000$ single blood cells (counts) [2]. A training period of many years is therefore prerequisite for biologists or physicians who perform the clinical data interpretation. It is, however, clear, that diagnostic structures in these files may be captured by an appropriate sampling procedure. In this work, we compare the advantages and disadvantages for three different sampling strategies producing a dataset consisting of $n_s < 5,000$ as a subset of the n original data: simple random [3], Learning Vector Quantization (LVQ) [4] and a novel proposal based on emergent self-organizing feature maps (ESOM) [5]. For a short overview on sampling strategies, see [3]. The approach is tested on different artificial and experimental datasets. Moreover, we validate our method by performing automated diagnosis of lymphomas employing diagnostic files from original flow cytometric patient lymphoma samples [6].

[1] Data Bionics Research Group, Philipps-University Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany
[2] Dept. of Hematology, Oncology and Immunology, Philipps-University Marburg, Baldingerstrasse; 35043 Marburg, Germany

ultsch@Mathematik.Uni-Marburg.de

References

1 Aghaeepour, Nima, et al. "Critical assessment of automated flow cytometry data analysis techniques." Nature methods 10.3 (2013): 228.

2 Köhnle T., Bücklein, Veit. "Anleitung LAIP Gating Strategie AML" Universities of Munich and Erlangen V 1.2 (2018).

3 Elfil, Mohamed, and Ahmed Negida. "Sampling methods in clinical research; an educational review." Emergency 5.1 (2017).

4 Kohonen, Teuvo. "Improved versions of learning vector quantization." 1990 IJCNN International Joint Conference on Neural Networks. IEEE, 1990.

5 Ultsch, Alfred. "Maps for the visualization of high-dimensional data spaces." Proc. Workshop on Self organizing Maps. 2003.

6 Hoffmann, J. et al "Determination of CD43 and CD200 surface expression can improve diagnostic accuracy of mature B-cell neoplasms" Jahrestagung der Deutschen, Österreichischen und Schweizerischen Gesellschaften für Hämatologie und Medizinische Onkologie (2018).

A Convolutional Neural Network for ECG Annotation as the Basis for the Classification of Cardiac Rhythms

Marcus Vollmer[1,2], *Philipp Sodmann*[1,2] , *Neetika Nath* [1] *and Lars Kaderali*[1,2]

Objective: Electrocardiography is the most common tool to diagnose cardiovascular diseases. Annotation, segmentation and rhythm classification of ECGs are challenging tasks, especially in the presence of atrial fibrillation and other arrhythmias. Our aim was to increase the accuracy of heart rhythm classification by the use of extreme gradient boosting trees and the development of a deep convolutional neural network for ECG segmentation.

Methods: We trained a convolutional neural network with waveforms from PhysioNet databases to annotate QRS complexes, P waves, T waves, noise and interbeat ECG segments that characterize the essences of normal and irregular heart beats. We checked

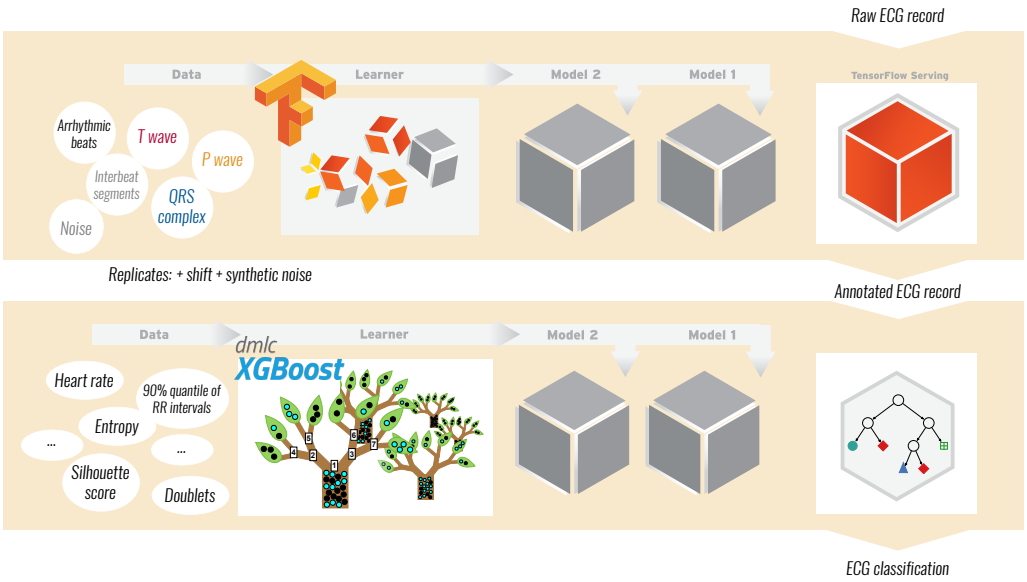


Figure 1: A schematic representation of our workflow. A CNN was trained with labeled 1.5 s ECG segments using TensorFlow. Features were extracted from the annotated ECG and eXtreme Gradient Boosting trees were used to classify the heart rhythm. TensorFlow logo by Wikimedia/FlorianCassayre (CC-BY-SA 4.0), TensorFlow serving chart adapted from www.tensorflow.org/serving/ (CC-BY-SA 3.0), random forest illustration adapted with permission from Sarah Chaudill and the American Physical Society from Figure 1 in [1]

[1] Institute of Bioinformatics, Universtity Medicine Greifswald
[2] German Centre for Cardiovascular Research (DZHK), partner site Greifswald

marcus.vollmer@uni-greifswald.de

the segmentation performance by direct paired comparison to the reference annotations of the QT [2] and MIT-BIH P-wave databases [3]. True positive rates, positive predictive values and the mean absolute difference were calculated. Moreover, we compared the results with standard QRS detectors and Ecgpuwave. Extreme gradient boosting trees [4] were used to determine the heart rhythm based on hancrafted features. Essential features were computed from interval data, including heart rate analysis and noise estimation. Furthermore, we defined particular features based on ECG morphology, appearance of P waves and detection of irregular beats. We examined the feature importance and identified key features for normal sinus rhythm, atrial fibrillation, alternative rhythm and noisy recordings. The classification performance was evaluated externally using F1 scores by applying the algorithm to the hidden test set of the PhysioNet/CinC Challenge 2017 [5].

Results: The true positive rate of the convolutional neural network in detection of manually revised R peaks in the QT database was 98% and the positive predictive value was 99%. The detection of P and T waves reached a true positive rate of 92% and 88% respectively, given a 50 ms tolerance when comparing the reference to the test annotation set. The rhythm classification performance reached an overall F1 score of 0.82 when applying the algorithm to the hidden test set [6,7].

References

1 P. T. Baker, S. Caudill, K. A. Hodge, D. Talukder, C. Capano, and N. J. Cornish, "Multivariate classification with random forests for gravitational wave searches of black hole binary coalescence," Physical Review D, vol. 91, no. 6, pp. 062004–062020, 2015.

2 P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in Computers in Cardiology, vol. 24, pp. 673–676, 1997.

3 Maršánová, Lucie and Němcová, Andrea and Smíšek, Radovan, "Detection of P wave during second-degree atrioventricular block in ECG signals," in Proceedings of the 23rd Conference STUDENT EEICT, pp. 655–659,2017.

4 T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16,pp. 785–794, ACM, 2016.

5 G. Clifford, C. Liu, B. Moody, I. Silva, Q. Li, A. Johnson, and R. Mark, "AF classification from a short single lead ECG recording: the PhysioNet Computing in Cardiology Challenge 2017," in Computing in Cardiology, vol. 44, 2017.

6 M. Vollmer, P. Sodmann, L. Caanitz, N. Nath, and L. Kaderali in 2017 Computing in Cardiology (CinC).

7 P. Sodmann, M. Vollmer, N. Nath, and L. Kaderali, "A convolutional neural network for ecg annotation as the basis for classification of cardiac rhythms," Physiological measurement, vol. 39, no. 10, p. 104005, 2018.

Uncertainty-Guided Semi-Automated Editing of CNN-based Retinal Layer Segmentations in Optical Coherence Tomography

Shekoufeh Gorgi Zadeh[1,2], Maximilian W. M. Wintergerst[3], and Thomas Schultz[4]

Convolutional neural networks (CNNs) have enabled dramatic improvements in the accuracy of automated medical image segmentation. Despite this, in many cases, results are still not reliable enough to be trusted "blindly". Consequently, a human rater is responsible to check correctness of the final result and needs to be able to correct any segmentation errors that he or she might notice. In [1], for a particular use case, segmentation of the retinal pigment epithelium (RPE) and bruch's membrane (BM from optical coherence tomography (OCT), we develop a system that makes this process more efficient by guiding the rater to segmentations that are most likely to require attention from a human expert, and by developing semi-automated tools for segmentation correction that exploit intermediate representations from the CNN. We demonstrate that our automated ranking of segmentation uncertainty correlates well with a manual assessment of segmentation quality, and with distance to a ground truth segmentation. We also show that, when used together, uncertainty guidance and our semi-automated editing tools decrease the time required for segmentation correction by more than a factor of three.

References

1 Gorgi Zadeh, S, Wintergerst, M, Schultz, T. Uncertainty-guided semi-automated editing of cnn-based retinal layer segmentations in optical coherence tomography. In: Proc. Visual Computing for Biology and Medicine. 2018, p. 107–115.

[1] Department of Computer Science, University of Bonn, Endenicher Allee 19a, 53115 Bonn, Germany
[2] Department of Medical Biometry, Computer Science and Epidemiology (IMBIE), University of Bonn, Sigmund-Freud-Str. 25, 53127 Bonn, Germany
[3] Department of Ophthalmology, University of Bonn, Germany
[4] Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Germany

shekoufeh.gorgizadeh@imbie.uni-bonn.de , Maximilian.Wintergerst@ukbonn.de, schultz@cs.uni-bonn.de

Big data: Challenges, Tools and Techniques in SAARC Region

Dila Ram Bhandari[1]

Big data is a term for huge data sets having large, varied and complex structure with challenges, such as difficulties in data capture, data storage, data analysis and data visualizing for further processing for leading to an upsurge of research, as well as industry and government applications. Data is deemed a powerful raw material that can impact multidisciplinary research endeavors as well as government digital revolution and business performance. Exploitation of Big Data platforms and technologies requires both corporate strategies and government policies to be in place much before the results would start pouring in. Digitization also has a significant impact on job creation in the overall economy of big data. The goal of this paper is to describe, review and reflect on big data share the data analytics opinions and perspectives of the authors relating to the new opportunities and challenges brought forth by the big data movement. The irrelevance of statistical significance, the challenges of computational efficiency and the unique characteristics of big data discussed above highlight the need to develop new statistical techniques to gain insights from predictive models. The authors bring together diverse perspectives, coming from different geographical locations with different core research expertise and different affiliations and work experiences.

Keywords: Big data, Government policy, social welfare, Hadoop, Data Privacy

[1] Faculty of Statistics Tribhuvan University, Kathmandu, Nepal

drbhandari2012@gmail.com

A bivariate time-to-event model for the meta-analysis of full ROC curves

Annika Hoyer[1]

Meta-analysis and systematic reviews are the cornerstones of evidence-based medicine and are used to inform treatment, diagnosis or prevention of patients as well as policy decisions in health care. While meta-analytic methods for intervention trials are well-established today, those for diagnostic accuracy studies are still under development in recent years due to the increased complexity of the bivariate outcome of sensitivity and specificity. The situation becomes even more challenging when the single studies report full ROC curves with several pairs of sensitivity and specificity corresponding to different diagnostic thresholds. However, this information is frequently ignored and only a single pair of sensitivity and specificity per study is used to arrive at meta-analytic estimates. Although methods have been proposed which deal with the full information, these have still disadvantages as, for example, allowing only for the same numbers or values of thresholds across studies.

To overcome the disadvantages of previously suggested models, we propose a novel approach for the meta-analysis of full ROC curves using all available information based on bivariate time-to-event models for interval-censored data [1]. The model is illustrated by an example on population-based screening for type 2 diabetes mellitus, where the original reviews include only one pair of sensitivity and specificity from 38 single studies, but an intensified search yields 124 pairs of sensitivity and specificity for 26 different thresholds.

References

1 Hoyer A, Hirt S, Kuss O. Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. Res Synth Methods 2018; 9(1):62-72

[1] German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Duesseldorf, Institute for Biometrics and Epidemiology

annika.hoyer@ddz.de

Enough is as good as a feast: Comparing different subsampling designs for time-to-event data

Jan Feifel[1], Luis Pauler[1] and Jan Beyersmann[1]

Antimicrobial resistance is one of the major burdens not only for today's clinicians, especially in infections resistant to Carbapenem, an antibiotic of last resort. Often the outcome variable is the time to a specific event. Standard procedures like the Cox regression model require covariate information for all individuals within the cohort.
In time-to-event analyses, information on the actual event times is only provided by uncensored patients. If the outcome is rare or if interest lies in evaluating expensive covariates, sub-sampling designs are favorable due to efficient use of limited resources. In our situation, not the outcome is necessarily rare, but interest lies in the impact of a rare time-dependent exposure such as the occurrence of an infection caused by carbapenem-resistance or disease progression.
We introduce and compare two different subcohorting schemes: the nested exposure case-control design[1] and the exposure density sampling[2]. Both account for past time-dependent exposure status to reduce the number of individuals substantially. A simulation study will outlay that a smart utilization of the available information at each point in time can lead to more powerful and simultaneously less expensive designs. Furthermore, a discussion of their relative merits will give a recommendation on when each design is most auspicious.

References

1 Feifel, J., Gebauer, M., Schumacher, M., & Beyersmann, J. (2018). Nested exposure case-control sampling: a sampling scheme to analyze rare time-dependent exposures. Lifetime data analysis, 1-24.

2 Wolkewitz, M., Beyersmann, J., Gastmeier, P., & Schumacher, M. (2009). Efficient risk set sampling when a time-dependent exposure is present. Methods of information in medicine, 48(05), 438-443.

[1] Institute of Statistics, Ulm University, Helmholtzstr. 20, 89081 Ulm, Germany

jan.feifel@uni-ulm.de, jan.beyersmann@uni-ulm.de

A Novel Spike-and-Slab Prior for Effect Selection in Distributional Regression Models

Nadja Klein[1]

We propose a novel spike and slab prior specification with scaled beta prime marginals for the importance parameters of regression coefficients to allow for general effect selection within the class of structured additive distributional regression. This enables us to model effects on all distributional parameters for arbitrary parametric distributions, and to consider various effect types such as non-linear or spatial effects as well as hierarchical regression structures. Our spike and slab prior relies on a parameter expansion that separates blocks of regression coefficients into overall scalar importance parameters and vectors of standardised coefficients. Hence, we can work with a scalar quantity for effect selection instead of a possibly high-dimensional effect vector, which yields improved shrinkage and sampling performance compared to the classical normal-inverse-gamma prior. We investigate the propriety of the posterior, show that the prior yields desirable shrinkage properties, propose a way of eliciting prior parameters and provide efficient Markov Chain Monte Carlo sampling. Using both simulated and three large-scale data sets, we show that our approach is applicable for data with a potentially large number of covariates, multilevel predictors accounting for hierarchically nested data and non-standard response distributions, such as bivariate normal or zero-inflated Poisson.

[1] Humboldt-Universität zu Berlin

nadja.klein@hu-berlin.de

Machine learning supported hypothesis and pattern finding in pain related phenotype data

Jörn Lötsch[1,2], Reetta Sipilä[3], and Eija Kalso[3]

Hypothesis generation in biomedical data can become challenging when a novel research topic is complex and incompletely understood. In the present analysis, we show the utility of a combination of unsupervised and supervised machine-learned techniques for data-driven assessment of pain-related phenotypes.
In a cohort of 1,000 women operated for breast cancer at the Helsinki University Hospital, psychological and sleep related parameters were acquired from a subgroup of n = 373 patients (complete data). In addition, pain intensity during the last week and the impact of pain on daily life activities (interference) were acquired. Unsupervised machine learning, implemented as emergent self-organizing feature map [1, 2], identified a structure in the data space of d = 17 psychological parameters indicating two-clusters. A "low pain but high interference" group was significantly overrepresented in the smaller cluster.
For the membership in this "low pain but high interference" group, feature selection based on fast and frugal tree (FFT [3, 4]) analysis identified d = 4 parameters as the most frequent size of the best performing trees during 1,000 runs on randomly resampled disjoint training and test data subsets. The derived decision rules obtained a balanced classification accuracy of 70% for the assignment to the "low pain but high interference" clinical group, versus all other clinical groups. The feature set was supported in its main parts, when repeating feature selection using random forests analysis followed by computed ABC analysis of the decrease in classification accuracy, when the respective feature was omitted from forest building [5].
The analysis shows, that complexity of clinical facets of persistent pain can be captured by combining unsupervised with supervised methods for of data exploration and analysis. Verifiable hypotheses can be found in complex data acquired with the expectation of interrelationships and subgroups but without clear pre-definitions of hypotheses.

Acknowledgement
Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL) and European Union Seventh Framework Programme (FP7/2013, grant agreement no. 602919, EK, JL, GLORIA).

[1] Institute of Clinical Pharmacology, Goethe - University, Theodor – Stern - Kai 7, 60590 Frankfurt am Main, Germany
[2] Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Project Group Translational Medicine and Pharmacology TMP, Theodor – Stern - Kai 7, 60590 Frankfurt am Main, Germany
[3] Institute of Clinical Medicine, University of Helsinki, Pain Clinic, Helsinki University Central Hospital, Helsinki, Finland

j.loetsch@em.uni-frankfurt.de

References

1 Ultsch A, Lötsch J. Machine-learned cluster identification in high-dimensional data. J Biomed Inform. 2017;66:95-104.

2 Lötsch J, Lerch F, Djaldetti R, Tegeder I, Ultsch A. Identification of disease-distinct complex biomarker patterns by means of unsupervised machine-learning using an interactive R toolbox (Umatrix). BMC Big Data Analytics. 2018(in press).

3 Phillips N, Neth H, Woike J, Gaissmaer W. FFTrees: Generate, Visualise, and Evaluate Fast-and-Frugal Decision Trees. R package version 1.4.0. 2018.

4 Gigerenzer G, Todd PM. Fast and frugal heuristics: The adaptive toolbox. Simple heuristics that make us smart. Evolution and cognition. New York, NY, US: Oxford University Press; 1999. p. 3-34.

5 Lötsch J, Ultsch A, editors. Random forests followed by computed ABC analysis as a feature selection method for machine-learning in biomedical data. Conference of the International Federation of Classification Societies IFCS-2017; 2017 August 8-10; Tokyo: Springer; 2018.

Pain Intensity Recognition via Deep Physiological Models

Patrick Thiam[1,2], Hans A. Kestler[1], and Friedhelm Schwenker[2]

The common inference model optimization process involves the design, assessment and selection of measurable descriptors based on some expert knowledge in the domain of application. The final inference model is subsequently trained, based on the set of selected descriptors. Recent works in the field of deep learning have shown that this whole process can be effectively and efficiently replaced by a neural network architecture that integrates feature engineering, feature selection and inference model optimization into a single learning process. Such approaches have been successfully applied in the domains of image and audio processing, with significantly improved overall performances in comparison to approaches based on traditional inference models such as support vector machines (SVMs) or decision trees.
In the following work, several deep learning approaches based on convolutional neural networks are designed and applied on measurable physiological channels in order to perform an accurate classification of different levels of artificially induced pain intensities. The aim of the current work is to achieve state-of-the-art pain intensity classification performances without the need of specific domain expert knowledge for the generation of relevant descriptors. The assessment of the designed classification architectures is based on the *BioVid Heat Pain Database (Part A)* and the conducted experimental validation demonstrates the relevance of the proposed approaches. Based on a *Leave-One-Subject-Out* (LOSO) cross-validation evaluation relative to the binary classification task consisting of the discrimination between the baseline and the pain tolerance level ($T_0 vs. T_4$), new state-of-the-art classification performances could be achieved, in particular with the electrodermal activity (EDA) and the designed deep fusion approach with respective classification rates of 85.03% and 84.74%.
Keywords: Pain Intensity Classification, Deep Neural Networks, Information Fusion, Signal Processing

[1] Institute of Medical Systems Biology,Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany
[2] Institute of Neural Information Processing, Ulm University, James-Franck-Ring, 89081 Ulm, Germany

patrick.thiam@uni-ulm.de, hans.kestler@uni-ulm.de, friedhelm.schwenker@uni-ulm.de

Graphical User Interfaces for Surrogate Model-Based Optimization in Practice and Teaching

Frederik Rehbach[1], Sebastian Krey[1] and Thomas Bartz-Beielstein[1]

State-of-the-art Surrogate Model-Based Optimization, Evolutionary algorithms and related search heuristics are well suited to solve complex industrial and scientific problems. Unfortunately, easy to use graphical user interfaces (GUI) are not available for many algorithms. Especially tools for the parallel computation of resource intensive methods often require long familiarisation phases. These difficulties prevent a wide adoption of these methods.

We claim that the availability of well-designed GUIs increases the interest for these technologies in industry and even for students without a background in optimization, statistics or computer science. In consequence they lead to a wider adoption in industrial practice.

The spotGUI R-package provides a GUI, based on Shiny, for the already well-established SPOT package. SPOT is a toolbox for model-based optimization, focusing on sequentially updated surrogate models for efficient optimization using state-of-the-art algorithms and modeling techniques that can be used without the requirement of optimization or programming knowledge. Additionally, it allows offloading computationally intensive tasks to an HPC cluster in a completely transparent way, allowing to easily solve extremely complex problems even for people who have never used an HPC system before.

We are successfully using the spotGUI for teaching design of experiments and statistical modeling to engineering students and for solving industrial optimization problems. One such problem is the electrostatic precipitator. It is a large scale electrical filtering/separation device, used to remove solid particles from gas streams, for example from the exhaust gases of coal-burning power plants. The configuration of these devices is a highly complex discrete optimization problem based on a computationally expensive computational fluid dynamics models of the exhaust gas system. In this work, we show how this optimization process can be controlled from a spotGUI interface and how the used methods improved the performance of the particle separation.

[1] Technische Hochschule Köln, Institut für Data Science, Engineering and Analytics, Steinmüllerallee 1, 51643 Gummersbach

frederik.rehbach@th-koeln.de, sebastian.krey@th-koeln.de,
thomas.bartz-beielstein@th-koeln.de

Correlation-based feature selection utilizing foreign classes

Robin Szekely[1,2], Ludwig Lausser[1] and Hans A. Kestler[1]

Datasets obtained from high-throughput experiments are typically characterized by their extremely high dimensionality compared to a small amount of samples ($n \gg m$). Feature profiles usually comprise several thousand molecular markers where sample sets rarely contain more than a hundred samples. Obtaining more samples is difficult due to ethical and economical reasons. To cope with this high-dimensional setting, sparse classification models or the aqusition of additional data resources might become relevant.
In this work, we supplement samples of an original binary classification task by samples of foreign but related classes [1,2]. We assume that these additional classes occur in multi-class datasets collected for a common research question. To be specific, samples from foreign classes are utilized for feature selection. We use an indirect selection strategy between original and foreign classes based on the Pearson correlation.
The feature selection strategy is evaluated in classification experiments on multi-class microarray and RNA-Sequencing datasets using linear support vector machines, random forests and k nearest neighbour classifiers. All experiments are designed as 10×10 cross-validation experiments. Our results show that certain constellations including foreign classes imply a high quality measure for the original classification task. In our evaluations this strategy outperformed the original feature selection over all datasets in up to 48.42%.

References

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

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

[1] Institute of Medical Systems Biology, Ulm University, 89081 Ulm, Germany
[2] International Graduate School in Molecular Medicine Ulm, Ulm University, 89081 Ulm, Germany
[3] Institute of Neural Information Processing, Ulm University, 89081 Ulm, Germany

Improved outcome prediction across data sources through robust parameter tuning

Nicole Schueller[1], Anne-Laure Boulesteix[1], Bernd Bischl[2], Kristian Unger[3,4], Roman Hornung[1]

In many application areas, prediction rules trained based on high-dimensional data are subsequently applied to make predictions for observations from other sources, but they do not always perform well in this setting. This is because data sets from different sources can feature (slightly) differing distributions, even if they are, in principle, similar in terms of population and definitions of the variables. In the context of high-dimensional data and beyond, most prediction methods involve one or several tuning parameters. Their values are commonly chosen by maximizing the cross-validated prediction performance within the training data. This procedure, however, implicitly presumes that the data to which the prediction rule will be ultimately applied, follow the same distribution as the training data. If this is not the case, less complex prediction rules that slightly underfit the training data may be preferable. Indeed, a tuning parameter does not only control the degree of adjustment of a prediction rule to the training data, but also, more generally, the degree of adjustment to the *distribution of* the training data. On the basis of this idea, we compare various approaches including new procedures for choosing tuning parameter values that lead to better generalizing prediction rules than those obtained based on cross-validation. Most of these tuning approaches use an external validation data set. In our extensive comparison study based on a large collection of 15 transcriptomic real data sets, tuning on external data and robust tuning with tuned robustness parameter are the two approaches leading to better generalizing prediction rules.

[1] Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich, 81377, Germany
[2] Department of Statistics, University of Munich, Ludwigstrasse 33, 80539 Munich, Germany
[3] Research Unit Radiation Cytogenetics, Helmholtz Zentrum Munich, German Research Center for Environmental Health GmbH, Neuherberg, Germany
[4] Department of Radiation Oncology, University Hospital, University of Munich, Munich, Germany

nschueller@ibe.med.uni-muenchen.de, boulesteix@ibe.med.uni-muenchen.de,
bernd.bischl@stat.uni-muenchen.de, unger@helmholtz-muenchen.de,
hornung@ibe.med.uni-muenchen.de

Scalable Automatic Machine Learning with H2O

Erin LeDell[1]*, Jo-fai Chow*[1]

Course Description:

H2O.ai will be presenting an overview of the H2O machine learning library with a focus on the AutoML (Automatic Machine Learning) functionality. This is an opportunity to learn about the field of Automatic Machine Learning and get hands-on with a popular, open source AutoML tool which can be used to speed-up and/or augment your machine learning modeling workflow. The tutorial will focus on the AutoML function inside the **h2o** R library, though equivalent materials will also be available for the **h2o** Python module. The H2O platform provides fast, scalable implementations of a variety of popular machine learning algorithms. No prior knowledge of H2O is required, though a familiarity with the basic concepts of supervised machine learning will be helpful.

In recent years, the demand for machine learning experts has outpaced the supply, despite the surge in people entering the field. To address this gap, there have been big strides in the development of user-friendly software for AI that can be used by non-experts. Although these tools have made it easier for non-experts to experiment with machine learning, there is still a fair bit of experience that is required to produce high-performing, production-ready machine learning models. A user of these tools must understand which algorithms to use under what circumstances, as well as how to tune the models to get good results on any particular dataset. A nascent subfield of AI called Automatic Machine Learning or simply, "AutoML" is rapidly growing to address this issue.

This tutorial will provide a history and overview of the field of Automatic Machine Learning and introduce H2O's approach to AutoML. The presentation will be followed by a hands-on demonstration of how to use H2O's open source AutoML software. R and Python code will be provided so that attendees can walk away with a practical understanding of how to automatically train and tune production-ready machine learning models in a single line of code or with press of a button on their own datasets.

Required Hardware / Software:

- Attendees will need their **laptops**.
- H2O R or Python package.
- Java JRE or JDK installed (requirement of H2O).
- Code will be on GitHub (thus internet access is required).

[1] H2O.ai, Mountain View, California, USA

erin@h2o.ai, joe@h2o.ai

Towards Human-Centered AutoML

Florian Pfisterer [1], Janek Thomas [1], Bernd Bischl [1]

Building models from data is an integral part of the majority of data science work flows. While data scientists are often forced to spend the majority of the time available for a given project on data cleaning and exploratory analysis, the time available to practitioners to build actual models from data is often rather short due to time constraints for a given project. AutoML systems such as *auto-sklearn* [2] are currently rising in popularity, as they can build powerful models without human oversight and knowledge. We propose to modify the AutoML process in two ways by 1) allowing to incorporate multiple criteria in the AutoML workflow, and 2) opening up systems to allow for human intervention and adaption during the search process.

In the scientific community, those systems are compared in several AutoML challenges[3] organized at top machine-learning conferences. These challenges focus solely on the predictive performance of models built by the AutoML systems, mainly because it is easy to compare and rank the systems in this way. The humans role in current AutoML processes is to choose data sets, validation protocols, performance measures to optimize and to define the pipeline search space, i.e., which preprocessing and modeling steps to consider. After that, the systems usually do not require human intervention and returns an optimal model after a prespecified amount of time. This often drastically speeds up the process of obtaining well working models as technical optimization is left to the machine and has not to be dealt with in a manual trial-and-error process. Furthermore, this process can be scaled up to run on massively parallel systems.

Besides predictive performance, we consider sparseness, model size, prediction speed and interpretability, as well as other measures such as fairness and robustness important criteria we might want to integrate in the AutoML process. All those criteria and their respective importance vary between projects and have inherent trade-offs, meaning that not all of them can be optimized equally well. We want to point out some criteria that are currently ignored in many AutoML systems, and showcase an implementation based on [4], that allows for the selection and tuning of models according to user-defined preferences. While some of the criteria mentioned can not be easily assessed in an automatic manner, others can already be integrated in existing AutoML frameworks. This problem can be circumvented by either finding working proxies that allow to assess models wrt. some criteria, such as fairness, or by allowing humans to rate whether models suit their choice and incorporating this feedback in the process. The latter can be achieved by getting the human-in-the-loop. This does not only increase adaptivity of the system, but would also help to increase trust in those systems because parts of the process is driven by the data scientist. Thus, systems should make intermediate results available to

[1] Ludwig-Maximilians Universität, München

florian.pfisterer@stat.uni-muenchen.de, janek.thomas@stat.uni-muenchen.de, bernd.bischl@stat.uni-muenchen.de

the practitioner, which can then be evaluated and played back to the AutoML system. An important approach to making this complex process more accessible to humans was proposed in ATMSeer[1]. This can especially help in situations, where user preferences are not easily quantifyable, or where relevant criteria are not known a-priori. The field of AutoML promises great enhancements to the current data science workflow, but to harness its full potential, it needs to be extended to be more accommodating towards multiple criteria and human intervention.

References

1 Qianwen el al. 2019, ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning, Human Factors in Computing Systems Proceedings (CHI 2019)

2 Feurer et al, 2015, Efficient and Robust Automated Machine Learning, Advances in Neural Information Processing Systems 28

3 Guyon et al., 2019, Analysis of the AutoML Challenge series, Springer series on Challenges in Machine Learning

4 Thomas et al., 2018, Automatic Gradient Boosting, International Workshop on Automatic Machine Learning at ICML

Boosting with random selection of weak learners for variable selection in high-dimensional models

Christian Staerk[1] *and Andreas Mayr*[1]

Statistical boosting is a promising alternative to popular regularization methods such as the Lasso [1] for fitting high-dimensional models with many possible explanatory variables: early stopping of the algorithm leads to implicit regularization and variable selection, enhancing the interpretability of the final models. Traditionally, the class of possible weak learners is fixed for all iterations of Boosting and usually consists of simple learners including only one explanatory variable at a time. Furthermore, the choice of the number of Boosting iterations is typically guided by optimizing the predictive performance of the resulting models, leading to final models which often include unnecessarily large numbers of noise variables with small effects.

We propose modifications of L_2Boost [2] for variable selection in high-dimensional linear models which aim at addressing the potential issues described above. The modifications are based on an adaptive random selection of different classes of weak learners in each Boosting iteration, where the adaptation of the weak learners is motivated by the recently proposed Adaptive Subspace (AdaSub) method [3, 4]. The considered classes include weak learners with several variables so that multiple coefficients can be updated at a single iteration. Furthermore, the proposed modifications of L_2Boost can impose an automatic stopping of the algorithm, leading to a reduced number of selected noise variables. The performance of the new approach is investigated in a simulation study, comparing it with the original version of L_2Boost as well as with other approaches such as Stability Selection for Boosting [5], which also aim at controlling the number of falsely selected variables.

References

1 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1), 267–288.

2 Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462), 324–339.

3 Staerk, C., Kateri, M. and Ntzoufras, I. (2016). An adaptive subspace method for high-dimensional variable selection. *Proceedings of the 31st International Workshop on Statistical Modelling*, Rennes, France, ed. by J. Dupuy and J. Josse. 295–300.

4 Staerk, C. (2018). Adaptive subspace methods for high-dimensional variable selection. *Ph.D. thesis, RWTH Aachen University*. URL `http://doi.org/10.18154/RWTH-2018-226562`

5 Hofner, B., Boccuto, L. and Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, 16(1), 144.

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn

`staerk@imbie.uni-bonn.de, mayr@uni-bonn.de`

imputeTS: Tidy Univariate Time Series Imputation

Steffen Moritz[1]*, Thomas Bartz-Beielstein*[1]

While nowadays more and more time series data are generated, missing values quite naturally remain a pervasive problem. Those missing values can compromise subsequent processes, thus replacing missing values (imputation) often benefits the quality of further data analysis.
Univariate time series imputation thereby is a special sub-field. Imputation techniques for cross-sectional data, time series cross-sectional data or multivariate time series altogether at least partially rely on inter-variable correlations to estimate the missing data. In comparison, univariate time series imputation solely can employ inter-temporal correlations.
We present **imputeTS**, an R package for time series imputation [1]. It provides several different state-of-the-art imputation and visualization functions for univariate, equispaced time series, $X = \{x_1, x_2, ..., x_n\}$. Additionally, the package works seamlessly with current tidy data workflows.
Favorable use cases hereby do not only include originally univariate time series. Univariate time series imputation is also helpful for multivariate time series with only uncorrelated variables. Another use case are multivariate time series, where whole observations at some points in time are completely missing. Especially the latter case is quite common for sensor data, if the data recording or transmission fails it often does for all variables at once. Meaning no inter-variable correlations can be employed for these points in time.

References

1 Moritz, S., Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. The R Journal, 9(1), 207-218. doi: https://doi.org/10.32614/RJ-2017-009

[1] Institute for Data Science, Engineering, and Analytics, TH Köln, Steinmüllerallee 1, 51643 Gummersbach

steffen.moritz10@gmail.com, thomas.bartz-beielstein@th-koeln.de

Joint Modelling approaches to survival analysis via likelihood-based boosting techniques.

Colin Griesbach[1], *Andreas Groll*[2], *Elisabeth Waldmann*[1]

When analyzing data where event-times are recorded alongside a longitudinal outcome, one commonly used approach in practice is separate modelling of the two outcomes without considering any interaction effects. Especially in survival analysis one main interest is incorporating time-varying covariates into the model. This however is quite a challenge, since popular methods like the extended cox regression produce biased results. Joint modelling on the other hand combines a longitudinal and a survival submodel in one single joint likelihood and thus accounts for interactions like time-varying covariates measured with error, which can be often found in follow-up studies. Previous works proposed algorithms to fit joint models via component-wise gradient boosting techniques which focus on minimizing the predictive risk, offer advantages like variable selection and also work with high dimensional data. However, gradient boosting leads to problems in the survival part of the model, since effects of time-varying covariates can not be estimated so easily. Likelihood-based boosting approaches on the other hand are, as verified in various literature, capable of handling time-dependent covariates in survival analysis, since likelihood-based boosting directly optimizes the likelihood by using newton algorithms with a component-wise updating procedure.

References

1 Wulfsohn, M., Tsiatis, A. (1997): A joint model for survival and longitudinal data measured with error. *Biometrics*. 53: 330-339.

2 Tutz, G., Binder, H. (2006): Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting. *Biometrics*. 62: 961-971.

3 Waldmann, E., Taylor-Robinson D., Klein N., Kneib T., Pressler T., Schmid, M. and Mayr, A. (2017): Boosting Joint Models for Longitudinal and Time-to-Event Data. *Biometrical Journal*. 59: 1104–1121.

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Technische Universität Dortmund

colin.griesbach@fau.de, groll@statistik.tu-dortmund.de, elisabeth.waldmann@fau.de

Tackling leukemia through computational statistics

Christiane Fuchs[1,2,3], *Lisa Amrhein*[2,3]

Acute Myeloid Leukemia (AML) is the most common acute leukemia affecting adults. Even after complete remission, leukemic cells likely remain in numbers below detection limit. Without further postremission or consolidation therapy, most AML patients will eventually relapse and die. An essential step towards successful treatment of AML is to understand the evolution of the genetic, epigenetic, and functional properties of clonally growing tumor cells.

The advent of single-cell RNA sequencing technologies holds enormous potential for research into this field. In recent years, technological advances have led to the development of large complex data. However, this data encounters uncertainty on both experimental and biological levels. Despite several breakthrough, measured values are still affected by substantial noise and missing values. Even under identical experimental conditions, data from different runs suffer from systematic differences including natural biological variation within and between individuals, especially regarding diseases like AML. In addition, AML patients frequently carry mixtures of different cancer cell types, so-called subclones, which evolve over time, so that the mixture at relapse is different from the one at diagnosis. Understanding clonal evolution and identifying rare subclones is still an open challenge.

In this talk, we will present computational modeling and estimation techniques that can contribute to the understanding of evolutionary processes in AML: the derivation of appropriate probability distributions to describe single-cell mRNA counts [1]; the identification of differently regulated cells from heterogeneous populations using mixture models [2]; and the modeling of time-continuous evolutions of clonal compositions.

References

1 Amrhein, A., Harsha, K., and Fuchs, C. (2019). A mechanistic model for the negative binomial distribution of single-cell mRNA counts. BioRxiv, doi: https://doi.org/10.1101/657619.

2 Bajikar*, S., Fuchs*, C., Roller, A., Theis, F., and Janes, K. (2014). Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. PNAS 111, E626–E635.

[1] Bielefeld University, of Business Administration and Economics, Universitätsstraße 25, 33615 Bielefeld
[2] Helmholtz Zentrum München, Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg
[3] Department of Mathematics, Technical University of Munich, Boltzmannstraße 3, 85747 Garching

christiane.fuchs@uni-bielefeld.de , lisa.amrhein@helmholtz-muenchen.de

bamlss: A Lego toolbox for flexible regression models

Nikolaus Umlauf[1]

During the last decades there has been an increasing interest in distributional regression models that allow to model all distributional parameters, such as location, scale and shape and thereby the entire data distribution conditional on covariates. In particular, the framework of structured additive distributional regression models enables to specify different types of effects such as linear, non-linear or interaction effects on all the distribution parameters hence providing a very flexible and generic framework suited for many complex real data problems. However, the implementation of new models is usually time-consuming and complex, especially using Bayesian estimation algorithms. We propose an unified modeling architecture that makes it possible to embed many different approaches suggested in literature and software. We show that implementing (new) algorithms, or the integration of already existing software, is relatively straightforward in this setting. An implementation is provided in the R package bamlss (https://cran.r-project.org/package=bamlss). We illustrate the usefulness of the approach by implementing neural network distributional regression models and evaluate the performance on simulated data and an application in survival analysis.

[1] Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck

Nikolaus.Umlauf@uibk.ac.at

Automatic Machine Learning and (Deep) Reinforcement Learning with rlR package

Xudong Sun[1] *,Jiali Lin*[1]*,Bernd Bischl*[1]

Reinforcement Learning especially Deep Reinforcement Learning has gained increasing attention recently. Deep reinforcement learning has gained great attention in the past years with its rapid advances in solving complex scenarios like Atari Games, the game of Go and continous robotic control, etc. The idea of using deep neural network as a general function approximator is relatively new to both the Computer Science and Statistics field. The R package rlR https://github.com/smilesun/rlR is developed to bridge the reinforcement learning with statistics by incorporating several state of art deep reinforcement learning algorithms into the package. Algorithms like Frozen Target Deep Q learning, Actor Critic Method, Deep Deterministic Policy Gradient, Trust Region Policy Optimization, Soft Q learning and Stein Variational Inference will be covered. The talk will cover the basic usage of rlR, and how it can be extended for Automatic Machine Learning (AutoML) with Bayesian Optimization. The new algorithm is called ReinBo. Its mechanism will be explained and Benchmarks with state of art AutoML softwares will be presented which showed a considerable improvement.

[1] LMU Munich, Ludwig Strasse 33, 80539, Munich

smilesun.east@gmail.com, linjialideu@gmail.com, bernd.bischl@gmx.net

Additional sparsity and enhanced variable selection for statistical boosting algorithms

Annika Strömer[1,2], Jan Speller[1], Christian Staerk[1] and Andreas Mayr[1]

An alternative approach for fitting linear models is component-wise gradient boosting. It is very flexible and useful for fitting high-dimensional data. Furthermore, statistical boosting includes variable selection, which is controlled by the main tuning parameter the number of boosting iterations [1,2].

While being very flexible and also relatively easy to extend, in some practical applications the algorithm shows the tendency towards selecting too many variables, including false positives. This seems to take place particularly for rather low-dimensional data ($p < n$) when one can generally observe a slower overfitting behavior. Due to the slow overfitting, the resulting stopping iteration (m_{stop}), tuned with cross-validation, gets larger and more variables are effectively included in the model. In practice, many of the false positives are incorporated with very small coefficients as the estimates are heavily shrinked towards zero. They hence do not have a larger impact on prediction accuracy, but lead to larger models with difficult interpretation.

We try to fix this problem by giving the algorithm the chance to de-select those variables that have (i) either been updated in very few iterations or have (ii) been estimated with very small coefficients. We analyze the impact of these fixes on both variable selection and prediction accuracy while comparing them to other methods for enhanced variable selection in the context of boosting.

References

1 Mayr, A. and Hofner, B. (2018): Boosting for statistical modelling – a non-technical introduction. *Statistical Modelling*. 18(3-4): 365-384.

2 Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014). Model-based Boosting in R: A Hands-on Tutorial Using the R Package mboost. *Computational Statistics*, 29:3-35.

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn
[2] Hochschule Koblenz, RheinAhrCampus Remagen

Annika.Stroemer@ukbonn.de, speller@imbie.uni-bonn.de, staerk@imbie.uni-bonn.de, mayr@uni-bonn.de

Adaptive robust boosting for high-dimensional regression models

Jan Speller[1], Christian Staerk[1] and Andreas Mayr[1]

A promising, application-oriented alternative to the already established methods for computing regression models is statistical boosting [1,2]. This approach is particularly helpful for high-dimensional data incorporating implicit variable selection via an iterative gradient-based descent in function space. The basic idea is to minimize the empirical risk by fitting simple base-learners (these are regression type functions in case of statistical boosting) to the negative gradient of the loss function. For the classical L_2 loss, this basically leads to re-fitting the residuals from the previous iteration.

While the classical mean regression via the L_2 loss is quite sensitive to extreme observations or outliers, optimizing the L_1 loss leads to the more robust median regression. We analyse the Huber loss [3,4] as a robust mixture between L_1 and L_2 and propose an adaptive approach via a self-regulating quantile-based parameter which ensures that the same amount of observations are incorporated in both parts of the loss in all boosting iterations. Therefore, we investigate in simulation and an application whether setting this quantile-based self-regulating parameter has an effect on the robustness of the final model.

References

1 Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014). Model-based Boosting in R: A Hands-on Tutorial Using the R Package mboost. *Computational Statistics*, 29:3-35.

2 Mayr, A. and Hofner, B. (2018): Boosting for statistical modelling – a non-technical introduction. *Statistical Modelling*. 18(3-4): 365-384.

3 Maronna, R. A., Martin, R. D., Yohai, V. J. and Salibián-Barrera, M. (2018). Robust Statistics: theory and methods (with R). *Wiley*.

4 Fox, J. and Weisberg, S. (2018). Robust Regression in R: An Appendix to An R Companion to Applied Regression, third edition.

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn

speller@imbie.uni-bonn.de, staerk@imbie.uni-bonn.de, mayr@uni-bonn.de

Correlation between quality of life and gait parameters in Parkinson's Disease

Isabelle Kaiser[1], Heiko Gassner[2], Jochen Klucken[2] and Werner Adler[1]

Gait deficits are common symptoms in Parkinson's disease (PD) [1], which leads to the restriction of mobility and thus the quality of life. A wearable sensor-based gait analysis system enables to objectively assess gait parameters without the need of having a specialized gait laboratory [2]. The correlation between the quality of life, which is represented by the widely used and standardized health survey SF-12 [3], and gait parameters is analyzed in this work.
Our study population consists of 163 Patients with Parkinson's disease visiting the movement disorder outpatient center at the University Hospital Erlangen, Germany, as wells as 95 controls. After an examination of pairwise correlations, we analyze the ability of gait parameters and common clinical variables like age and sex to predict quality of life with several different regression models, e.g. multiple linear regression, support vector machine [4] and random forest [5]. The performances of these models are estimated via cross-validation.
Results show that there are several significant correlations between gait parameters and the physical component of the quality of life. Nevertheless, reliable prediction of quality of life using only variables available in our study is not possible.

References

1 Gaßner, H. et al. Gait and Cognition in Parkinson's Disease: Cognitive Impairment Is Inadequately Reflected by Gait Performance during Dual Task. Frontiers in Neurology. 2017.

2 Schlachetzki, J. et al. Wearable sensors objectively measure gait parameters in Parkinson's disease. PLOS ONE. 2017.

3 Bullinger, M. Erfassung der gesundheitsbezogenen Lebensqualität mit dem SF36 Health Survey. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz. 2000.

4 Vapnik, V. The Nature of Statistical Learning Theory. Springer Verlag. 1995.

5 Breiman, L. Random Forests. Machine Learning. 45. 5-32. 2001.

[1] Department of Biometry and Epidemiology, University of Erlangen-Nuremberg, Erlangen
[2] Department of Molecular Neurology, University Hospital Erlangen, University of Erlangen-Nuremberg, Erlangen

isabelle.kaiser@fau.de, heiko.gassner@uk-erlangen.de, jochen.klucken@uk-erlangen.de, werner.adler@fau.de

Generative artificial intelligence based algorithm to increase the predictivity of preclinical studies while keeping sample sizes small

Jörn Lötsch[1,2], and Alfred Ultsch[3]

The translation of basic science into new clinical effective compounds seems to be often unsatisfactory. This has been attributed to inappropriate data quality standards and to the often small sample sizes. As solutions of this problem, combinations of several studies or otherwise-increased sample sizes have been proposed to obtain adequate statistical power, facilitating the validity of the results and their predictivity of clinical drug effects. We propose the development of an artificial intelligence (AI) -based method, which can generate valid, i.e. which have the same structure and properties, additional data from available data sets. If the data is acquired only in small numbers, the sample size for data analysis is enhanced, without increasing the number of laboratory animals included in the preclinical experiments. Specifically, generative models (GM) will be used that try to solve the problem of generating valid data from a nontrivial, possibly high dimensional distributions, that are either unknown or can be hardly described analytically. In the case of success, GM can solve the problems of sparse data, rare cases or small sample sizes. Several statistical models for GM have been proposed, including rejection sampling, the Metropolis-Hasting algorithm and the so-called inverse transform method. More recent developments include Generative Adversarial Networks (GAN), which have the advantage that a comparison of the distribution of generated data with the targeted non-trivial distribution can be obtained by means of a learned classifier implemented as a supervised neuronal network. GAN proofed as particularly useful in image processing and related tasks.

An alternative structure detecting neuronal network based algorithm is provided by emergent self-organizing maps (ESOM), which have the advantage of a vast body of successful applications on non-image related biomedical data, both experimental and clinical. Using the so-called "U-matrix" it can be judged whether or not structure in the data exists [1-5]. Based on the valid data structure detection by means of the ESOM/U-matrix artificial intelligence, generative neuronal networks offer the possibility to generate valid synthetic examples in the data space. Following investigation of the structure of data acquired in small samples or small subgroups, the probability matrix ("P-matrix") can be obtained that represents the joint distribution p(x,c), where x is a feature vector of a data example and c the label of a subgroup [6]. Using both, the structure detection properties of the SOM algorithm and the probability matrix of the data, new data can

[1] Institute of Clinical Pharmacology, Goethe - University, Theodor – Stern - Kai 7, 60590 Frankfurt am Main, Germany
[2]Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Project Group Translational Medicine and Pharmacology TMP, Theodor – Stern - Kai 7, 60590 Frankfurt am Main, Germany
[3] Databionics Research Group, Faculty of Mathematics and Computer Science, University of Marburg, Hans - Meerwein - Straße, 35032 Marburg, Germany

j.loetsch@em.uni-frankfurt.de, ultsch@Mathematik.Uni-Marburg.de

be generated that artificially increase the sample size.
Thus, to increase the predictive power of preclinical studies, rather than an increase in the number of cases, directly or indirectly via merging multiple studies, we propose an innovative, AI-based, valid, and probabilistic generation of experimental data that, together with the original data, provides the high data density necessary to draw valid conclusions from preclinical experiments. First proof-of-concept data will be presented; a working implementation into preclinical model derived data is subject of ongoing and future research.

Acknowledgement
Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL).

References

1 Lötsch J, Geisslinger G, Heinemann S, Lerch F, Oertel BG, Ultsch A. QST response patterns to capsaicin- and UV-B-induced local skin hypersensitization in healthy subjects: a machine-learned analysis. Pain. 2017.

2 Ultsch A, Lötsch J. Machine-learned cluster identification in high-dimensional data. J Biomed Inform. 2017;66:95-104.

3 Lötsch J, Schiffmann S, Schmitz K, Brunkhorst R, Lerch F, Ferreiros N, et al. Machine-learning based lipid mediator serum concentration patterns allow identification of multiple sclerosis patients with high accuracy. Sci Rep. 2018;8(1):14884.

4 Lötsch J, Lerch F, Djaldetti R, Tegeder I, Ultsch A. Identification of disease-distinct complex biomarker patterns by means of unsupervised machine-learning using an interactive R toolbox (Umatrix). BMC Big Data Analytics. 2018;3(5):https://doi.org/10.1186/s41044-018-0032-1.

5 Lötsch J, Hummel T, Ultsch A. Machine-learned pattern identification in olfactory subtest results. Sci Rep. 2016;6:35688.

6 Ultsch A. Maps for Visualization of High-Dimensional Data Spaces. WSOM. 2003:225-30.

Utilizing molecular network information via Graph Convolutional Neural Networks to predict metastatic event in breast cancer

Hryhorii Chereda[1], Annalen Bleckmann[1, 2], Frank Kramer[3], Andreas Leha[1], and Tim Beissbarth[1]

Gene expression data is commonly available in cancer research and provides a snapshot of the molecular status of a specific tumor tissue. This high-dimensional data can be analyzed for diagnoses, prognoses, and to suggest treatment options. Machine learning based methods are widely used for such analysis.
In recent years deep learning was applied to a wide range of problems in various areas. Deep learning methods are aimed at the automatic learning of data representations (features) needed for machine learning task. These methods demonstrated state-of-the-art performance in visual object recognition, object detection, speech recognition as well as other domains such as drug discovery and genomics [1]. One of the most popular methods of deep learning are Convolutional Neural Networks (CNN). They show cutting edge results for data that are spatially structured. The main property of CNNs is a capability of capturing local spatial patterns in natural signals and merging them into high-level abstractions. Nowadays, deep learning is extending to Non-Euclidean domains. Essentially such an extension is based on generalization of CNNs [2] to graphs. Molecular networks are commonly represented as graphs detailing interactions between molecules. Gene expression data can be assigned to the vertices of these graphs, and the edges can depict interactions, regulations and signal flow. In other words, gene expression data can be structured by utilizing molecular network information as prior knowledge.
Here, we applied graph CNN to gene expression data of breast cancer patients to predict the occurrence of metastatic events. To structure the data we utilized a protein-protein interaction network. We show that the graph CNN exploiting the prior knowledge is able to provide classification improvements for the prediction of metastatic events compared to existing methods.

References

1 Y. LeCun, Y. Bengio, G. Hinton, Deep Learning, Nature 521 (2015), 436-444. doi: 10.1038/nature14539

2 F. Monti et al, Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5115-5124. doi: 10.1109/CVPR.2017.576

[1] Medical Bioinformatics, University Medical Center Göttingen, 37099 Göttingen
[2] Hematology and Medical Oncology, University Medical Center Göttingen, 37099 Göttingen
[3] IT Infrastructure for Translational Medical Research, University of Augsburg, 86159 Augsburg

Hryhorii.Chereda@med.uni-goettingen.de, Tim.Beissbarth@ams.med.uni-goettingen.de

Extracting ordinal class sequences from high-dimensional datasets

Lisa M. Schäfer[1,2], Ludwig Lausser[1] and Hans A. Kestler[1]

Ordinal classifier cascades have been shown to be suitable for detecting embeddings of total ordinal relations in high-dimensional feature spaces [1]. The performance of these classifier cascades depends on the assumed class order. A performance drop indicates that an assumed order is not reflected within the analysed feature representation. By screening through all possible class permutations it is possible to distinguish between reflected and not reflected classes.
However, the assumption of one total order comprising all class labels might not be valid. We therefore extended this screening procedure to ordinal relations in any subgroups of classes.

Within an explorative analysis we applied the extended screening procedure to various datasets. We analysed the best performing longest cascades, as those imply the most and accurate information about a dataset. Furthermore, we observed that the found cascades often allow for a compact assignment of the remaining classes.
The screening procedure does not only allow the confirmation of known ordinal relations in alternative feature representations, but also to reveal new relations and thereby to get more insight about the neighbourhood structure of classes.

References

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.

[1] Institute of Medical Systems Biology, Ulm University, 89081 Ulm, Germany
[2] International Graduate School in Molecular Medicine Ulm, Ulm University, 89081 Ulm, Germany

lisa-1.schaefer@uni-ulm.de, ludwig.lausser@uni-ulm.de, hans.kestler@uni-ulm.de

Adaptive step-lengths in model-based gradient boosting algorithms for distributional regression

Tobias Hepp[1,3],Boyao Zhang[2],Andreas Mayr[1],Elisabeth Waldmann[3]

Gradient boosting methods are iterative updating schemes, in which the gradient of the loss-function in the current step is fitted to the data. The best fitting variable is then selected and updated by a fracture ν of the parameter suggested. Technically speaking, this means that the algorithm must be tuned with two parameters: the overall number of iterations and this fracture, also called step-length, but current practice is to fix the latter at $\nu = 0.1$. While this has been shown to work well in a variety of scenarios, those findings are based on models with only one gradient [1]. However, using gradient boosting algorithms for more complex model classes such as generalized additive models for location, shape and scale involves fitting multiple gradients derived from a single global loss function [2]. Depending on the updating scheme in the iterations, using the same step-length for all base-leaners may result in unfair comparisons in the selection process. Balancing the step-lengths with the potential contribution of the corresponding base-learners to the global loss function might therefore help to improve the performance of the algorithm regarding variable selection accuracy and overall model sparsity.

References

1 Matthias Schmid and Torsten Hothorn. (2008). Boosting additive models using component-wise P-splines as base-learners. *Computational Statistics & Data Analysis,* 53(2), 298-311.

2 Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 61(3), 403-427.

[1] Institut für Medizinische Biometrie, Informatik und Epidemiologie, Rheinische Friedrich-Wilhelms-Universität Bonn
[2] Institut für Statistik, Ludwig-Maximilians-Universität München
[3] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg

tobias.hepp@uk-erlangen.de

A continuous-time capture-recapture model for the annual movement of bottlenose dolphins

Sina Mews[1],Roland Langrock[1],Ruth King[2],Nicola Quick[3]

Our modelling approach is motivated by individual sighting histories of bottlenose dolphins off the east coast of Scotland. Due to ongoing offshore development, conservation managers seek to better understand the temporal movement patterns of the dolphin population between different sites. Typically, the Arnason-Schwarz model is fitted to such multi-state capture-recapture data[1], assuming a first-order Markov chain in discrete time for the state process, which here corresponds to the location (site) of a given individual (in addition to alive and dead). In our case, however, the capture occasions are not regularly spaced in time, leading to the problem that standard capture-recapture methods are not readily applicable as they address the more commonly found regular sampling protocols. Therefore, we consider a continuous-time model formulation instead.

The capture-recapture setting can be regarded as a special case of a (partially) hidden Markov model (HMM), with the observed capture history of an individual as the state-dependent process and an underlying, partially observed state process related to the movement of the individual between different sites. In particular, we can exploit the convenient and efficient HMM-based forward algorithm for evaluating the likelihood and hence for parameter estimation[2]. Further inferential tools that become applicable by embedding the capture-recapture setting in the HMM framework include the Viterbi algorithm and the forward-backward algorithm, which can be used to decode the underlying states (sites).

The main aim of the present analysis was to investigate how the dolphins' movement rates

[1] Bielefeld University, Germany, Faculty of Business Administration and Economics, Universitätsstraße 25, 33615 Bielefeld
[2] University of Edinburgh, UK, School of Mathematics, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD Z
[3] Duke University Marine Lab, United States, Nicholas School of the Environment, Duke Marine Lab, 135 Duke Marine Lab Road, Beaufort, NC 28516

sina.mews@uni-bielefeld.de, roland.langrock@uni-bielefeld.de, Ruth.King@ed.ac.uk, nicola.quick@duke.edu

between two sites, expressed as state transition intensities in our model, depend on the time of year. However, incorporating such time-varying covariates into the continuous-time Markov state process is rather challenging as the corresponding likelihood function then becomes intractable. We suggest an approximation using piecewise constant state transition intensities, which renders the likelihood evaluation feasible[3]. The approximation can be made arbitrarily accurate by using an increasingly fine resolution of the approximating step function.

The suggested approach is applied to investigate the annual movement of bottlenose dolphins between their main sites on the east coast of Scotland, revealing seasonal patterns which can help to inform conservation management. Our modelling approach can easily be transferred to other scenarios and is hence a general method for irregularly sampled capture-recapture data subject to switches in underlying states.

References

1 Schwarz, C.J., Schweigert, J.F., and Arnason, A.N. (1993). Estimating migration rates using tag-recovery data. *Biometrics*, **49**(1), 177–193.

2 Jackson, C.H., Sharples, L.D., Thompson, S.G., Duffy, S.W., and Couto, E. (2003). Multistate Markov models for disease progression with classification error. *The Statistician*, **52**(2), 193–209.

3 Langrock, R., Borchers, D.L., and Skaug, H.J. (2013). Markov-modulated nonhomogeneous Poisson processes for modeling detections in surveys of marine mammal abundance. *Journal of the American Statistical Association*, **108**(503), 840–851.

Simulation of trajectories in the illness-death model for chronic diseases: discrete event simulation and Doob-Gillespie algorithm

Ralph Brinks[1], *Annika Hoyer*[1]

This work is about simulation of populations transiting through the stages of the illness-death model for chronic diseases. First, we compare the commonly used discrete event simulation [1] with the Doob-Gillespie algorithm [2,3] in terms of computational speed. The comparison is accomplished in a test example motivated from diabetes in the German population [4]. It turns out that the current implementation of the discrete event simulation is slower than the Doob-Gillespie algorithm by a factor of about 75. Second, we use the Doob-Gillespie algorithm to explore the coverage probability of the 95% Wald confidence intervals of the binomial distribution for different population sizes (n) and success probabilities (p). The success probabilities p in the illness-death model corresponds to the prevalences of the chronic disease. Coverage is examined by 5000 simulation runs for population sizes from n = 50 to n = 500,000 and prevalences from p = 1% to p = 35%. The prevalences p are obtained from the solving an ordinary differential equation that is closely related to the illness-death model [5]. Irrespective of the tested simulation settings, the coverage probability of the 95% Wald confidence intervals is at least 94.3% and reaches up to 97.4%. Thus, in the tested settings the Wald confidence interval is a reasonable approximation to the 95% confidence interval of the binomial distribution .

References

1 Gill RD, Johansen S. A survey of product-integration with a view toward application in survival analysis, The Annals of Statistics 18(4): 1501-1555, 1990

2 Doob JL. Topics in the Theory of Markoff Chains, Transactions of the American Mathematical Society 52(1): 37-64, 1942

3 Gillespie DT. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions, Journal of Computational Physics 22 (4): 403–434, 1976

4 Brinks R. Illness-Death Model in Chronic Disease Epidemiology: Characteristics of a Related, Differential Equation and an Inverse Problem, Computational and Mathematical Methods in Medicine, Article ID 5091096, 2018. https://doi.org/10.1155/2018/5091096.

5 Brinks R, Hoyer A. Illness-death model: statistical perspective and differential equations, Lifetime Data Analysis, 2018

[1] Institute of Biometrics and Epidemiology, German Diabetes Center, Duesseldorf

Ralph.Brinks@ddz.de, Annika.Hoyer@ddz.de

A Classification Tree Approach for the Modeling of Competing Risks in Discrete Time

Moritz Berger[1], Thomas Welchowski[1], Steffen Schmitz-Valckenberg[2], Matthias Schmid[1]

Cause-specific hazard models are a popular tool for the analysis of competing risks data. The classical modeling approach in discrete time consists of fitting parametric multinomial logit models. A drawback of this method is that the focus is on main effects only, and that higher order interactions are hard to handle. Moreover, the resulting models contain a large number of parameters, which may cause numerical problems when estimating coefficients. To overcome these problems, a tree-based model is proposed that extends the survival tree methodology developed previously for time-to-event models with one single type of event. The performance of the method, compared with several competitors, is investigated in simulations. The usefulness of the proposed approach is demonstrated by an analysis of age-related macular degeneration among elderly people that were monitored by annual study visits.

Keywords: Discrete time-to-event data, Competing Risks, Recursive Partitioning, Cause-Specific Hazards, Regression modeling

[1] Rheinische Friedrich-Wilhelms-Universität, Department of Medical Biometry, Informatics and Epidemiology, Sigmund-Freud-Str. 25, 53127 Bonn, Germany
[2] Rheinische Friedrich-Wilhelms-Universität, University Eye Hospital Bonn, Ernst-Abbe-Str. 2, D-53127 Bonn, Germany

Moritz.Berger@imbie.uni-bonn.de, welchow@imbie.meb.uni-bonn.de,
steffen.schmitz-valckenberg@ukb.uni-bonn.de, matthias.schmid@imbie.uni-bonn.de

Semantic multi-class classifier systems in precision medicine

Lea Siegle[1], Ludwig Lausser[1], Hans A. Kestler[1]

Molecular high-throughput technologies have made the aggregation of large collections of bio-markers possible. The downside: the size of these collections prohibit the direct analysis by human experts. Intermediate representations such as classification models are required for the evaluation of these high-dimensional datasets, be it in diagnostics (for personalised medicine) or in sciences (to create new hypotheses). However, the patterns used to classify samples are often as unintererepretable as the data itself. We developed a new technique named "Semantic multi-class classifier system" (SM-CCS) to train biologically meaningful classifiers and to increase model interpretability. SMCCSs combine (knowledge-based) semantic feature selection (SFS) on the basis of established vocabularies (e.g. GO and KEGG) with a multi-class classification (MCC) method [1]. Here, we test two types of SFS on the well known one-against-one (OaO) and one-against-all (OaA) MCCs [2]: In both strategies features are selected individually for each base classifier. In the first strategy (type I), however, only one of the selected terms is chosen (either the least well, the mean or the best term, called MMM selection) and provided to all of the base classifiers, whereas in the second strategy (type II), each base classifier is trained in their own selected term. The four combinations (with all different MMM selections) are evaluated on a 4-classed breast cancer as well as on a 5-classed liposarcoma dataset and characterised by their accuracy and feature selection stability. We additionally provide evidence on the selected semantic terms. Interestingly, support can even be found for terms selected for individual twoclass comparisons.

References

1 Lausser, L., Schmid, F., Platzer, M., Sillanpaa, M.J., Kestler, H.A.: Semantic multi-classifier systems for the analysis of gene expression profiles. Arch. Data Sci. Ser. A (Online First) 1(1) (2016)

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

[1] Institute of Medical Systems Biology, Ulm University, 89069 Ulm, Germany

lea.siegle@uni-ulm.de, ludwig.lausser@uni-ulm.de, hans.kestler@uni-ulm.de

Modeling strategies to dissect the variable genetic architecture in the computation of polygenic risk score

Carlo Maj[1], Oleg Borisov[1], Christian Staerk[2], Andreas Mayr[2], Peter Krawitz[1]

Many complex (i.e., not monogenic) phenotypes are characterized by a relatively high heritability (e.g., height, common disease susceptibility [1]). However, only a little proportion of the phenotypic variance can be explained by significantly associated genetic variants identified by means of genome-wide association studies [2]. The genetic contribution of these traits can be due to the small effect of several variants leading to an overall genetic load according to an additive model [3]. Such a genetic load is known in the genetic field as "polygenic risk score" and indeed several works revealed strong polygenetic associations for many traits using linear models with quantitative traits (e.g., height [4], weight [5]) or logistic models for binary traits (e.g., case/control status for many disease [6]). The integration of polygenic risk score with other influencing factors proved to improve the performance of phenotype prediction model [7]. However, the overall applicability of these scores in clinical practice is still under investigation due to issues in the modeling of the genetic architecture [8]. In particular, genetic correlation across variants (i.e., linkage disequilibrium) and different allele frequencies across populations (i.e., population stratification) can strongly influence polygenic risk score hampering the generability of these scores across different datasets [9]. In the present work, we analyze different strategies to model linkage disequilibrium and population stratification in order to have more generalizable genetic risk models.

[1] Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn, Germany
[2] Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany

cmaj@uni-bonn.de, olegbor@uni-bonn.de, amayr@uni-bonn.de,
christian.staerk@imbie.uni-bonn.de. amayr@uni-bonn.de. pkrawitz@uni-bonn.de

References

1 Jie Zheng, A Mesut Erzurumluoglu, Benjamin L Elsworth, John P Kemp, Laurence Howe, Philip C Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, Beate St Pourcain, et al. "LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis". In: Bioinformatics 33.2 (2017), pp. 272–279

2 Evan A Boyle, Yang I Li, and Jonathan K Pritchard. "An expanded view of complex traits: from polygenic to omnigenic". In: Cell 169.7 (2017), pp. 1177–1186

3 Frank Dudbridge. "Power and predictive accuracy of polygenic risk scores". In: PLoS genetics 9.3 (2013), e1003348.

4 Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zolt an Kutalik, et al. "Defining the role of common variation in the genomic and biological architecture of adult human height". In: Nature genetics 46.11 (2014), p. 1173.

5 Amit V Khera, Mark Chaffin, Kaitlin H Wade, Sohail Zahid, Joseph Brancale, Rui Xia, Marina Distefano, Ozlem Senol-Cosar, Mary E Haas, Alexander Bick, et al. "Polygenic prediction of weight and obesity trajectories from birth to adulthood". In: Cell 177.3 (2019), pp. 587–596

6 Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, et al. "Genome-wide polygenic scores for com- mon diseases identify individuals with risk equivalent to monogenic mutations". In: Nature genetics 50.9 (2018), p. 1219.

7 Jing Dong, Matthew F Buas, Puya Gharahkhani, Bradley J Kendall, Lynn Onstad, Shanshan Zhao, Lesley A Anderson, Anna H Wu, Weimin Ye, Nigel C Bird, et al. "Determining risk of Barrett's esophagus and esophageal adenocarcinoma based on epidemiologic factors and genetic variants". In: Gastroenterology 154.5 (2018), pp. 1273–1281.

8 Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. "Clinical use of current polygenic risk scores may exacerbate health disparities". In: Nature genetics 51.4 (2019), p. 584.

9 David Curtis. "Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia". In: Psychiatric genetics 28.5 (2018), pp. 85–89

Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition

Christoph Molnar[1], *Giuseppe Casalicchio*[1], *Bernd Bischl*[1]

To obtain interpretable machine learning models, either interpretable models are constructed from the outset – e.g. shallow decision trees, rule lists, or sparse generalized linear models – or post-hoc interpretation methods – e.g. partial dependence or ALE plots – are employed. Both approaches have disadvantages. While the former can restrict the hypothesis space too conservatively, leading to potentially suboptimal solutions, the latter can produce too verbose or misleading results if the resulting model is too complex, especially w.r.t. feature interactions. We propose to make the compromise between predictive power and interpretability explicit by quantifying the complexity / interpretability of machine learning models. Based on functional decomposition, we propose measures of number of features used, interaction strength and main effect complexity. We show that post-hoc interpretation of models that minimize the three measures becomes more reliable and compact. Furthermore, we demonstrate the application of such measures in a multi-objective optimization approach which considers predictive power and interpretability at the same time.

[1] Department of Statistics, LMU Munich, Ludwigstr. 33

christoph.molnar@stat.uni-muenchen.de,
giuseppe.casalicchio@stat.uni-muenchen.de,
bernd.bischl@stat.uni-muenchen.de

List of technical reports published by the University of Ulm

Some of them are available by FTP from ftp.informatik.uni-ulm.de
*Reports marked with * are out of print*

91-01 Ker-I Ko, P. Orponen, U. Schöning, O. Watanabe
Instance Complexity

*91-02** K. Gladitz, H. Fassbender, H. Vogler
Compiler-Based Implementation of Syntax-Directed Functional Programming

*91-03** Alfons Geser
Relative Termination

*91-04** J. Köbler, U. Schöning, J. Toran
Graph Isomorphism is low for PP

91-05 Johannes Köbler, Thomas Thierauf
Complexity Restricted Advice Functions

*91-06** Uwe Schöning
Recent Highlights in Structural Complexity Theory

*91-07** F. Green, J. Köbler, J. Toran
The Power of Middle Bit

*91-08** V.Arvind, Y. Han, L. Hamachandra, J. Köbler, A. Lozano, M. Mundhenk, A. Ogiwara, U. Schöning, R. Silvestri, T. Thierauf
Reductions for Sets of Low Information Content

*92-01** Vikraman Arvind, Johannes Köbler, Martin Mundhenk
On Bounded Truth-Table and Conjunctive Reductions to Sparse and Tally Sets

*92-02** Thomas Noll, Heiko Vogler
Top-down Parsing with Simulataneous Evaluation of Noncircular Attribute Grammars

92-03 Fakultät für Informatik
17. Workshop über Komplexitätstheorie, effiziente Algorithmen und Datenstrukturen

*92-04** V. Arvind, J. Köbler, M. Mundhenk
Lowness and the Complexity of Sparse and Tally Descriptions

*92-05** Johannes Köbler
Locating P/poly Optimally in the Extended Low Hierarchy

*92-06** Armin Kühnemann, Heiko Vogler
Synthesized and inherited functions -a new computational model for syntax-directed semantics

*92-07** Heinz Fassbender, Heiko Vogler
A Universal Unification Algorithm Based on Unification-Driven Leftmost Outermost Narrowing

*92-08** Uwe Schöning
On Random Reductions from Sparse Sets to Tally Sets

*92-09** Hermann von Hasseln, Laura Martignon
Consistency in Stochastic Network

92-10 Michael Schmitt
A Slightly Improved Upper Bound on the Size of Weights Sufficient to Represent Any Linearly Separable Boolean Function

92-11 Johannes Köbler, Seinosuke Toda
On the Power of Generalized MOD-Classes

92-12 V. Arvind, J. Köbler, M. Mundhenk
Reliable Reductions, High Sets and Low Sets

92-13 Alfons Geser
On a monotonic semantic path ordering

*92-14** Joost Engelfriet, Heiko Vogler
The Translation Power of Top-Down Tree-To-Graph Transducers

93-01 Alfred Lupper, Konrad Froitzheim
AppleTalk Link Access Protocol basierend auf dem Abstract Personal Communications Manager

93-02 M.H. Scholl, C. Laasch, C. Rich, H.-J. Schek, M. Tresch
The COCOON Object Model

93-03 Thomas Thierauf, Seinosuke Toda, Osamu Watanabe
On Sets Bounded Truth-Table Reducible to P-selective Sets

93-04 Jin-Yi Cai, Frederic Green, Thomas Thierauf
On the Correlation of Symmetric Functions

93-05 K.Kuhn, M.Reichert, M. Nathe, T. Beuter, C. Heinlein, P. Dadam
A Conceptual Approach to an Open Hospital Information System

93-06 Klaus Gaßner
Rechnerunterstützung für die konzeptuelle Modellierung

93-07 Ullrich Keßler, Peter Dadam
Towards Customizable, Flexible Storage Structures for Complex Objects

94-01 Michael Schmitt
On the Complexity of Consistency Problems for Neurons with Binary Weights

94-02 Armin Kühnemann, Heiko Vogler
A Pumping Lemma for Output Languages of Attributed Tree Transducers

94-03 Harry Buhrman, Jim Kadin, Thomas Thierauf

On Functions Computable with Nonadaptive Queries to NP

94-04 Heinz Faßbender, Heiko Vogler, Andrea Wedel
Implementation of a Deterministic Partial E-Unification Algorithm for Macro Tree Transducers

94-05 V. Arvind, J. Köbler, R. Schuler
On Helping and Interactive Proof Systems

94-06 Christian Kalus, Peter Dadam
Incorporating record subtyping into a relational data model

94-07 Markus Tresch, Marc H. Scholl
A Classification of Multi-Database Languages

94-08 Friedrich von Henke, Harald Rueß
Arbeitstreffen Typtheorie: Zusammenfassung der Beiträge

94-09 F.W. von Henke, A. Dold, H. Rueß, D. Schwier, M. Strecker
Construction and Deduction Methods for the Formal Development of Software

94-10 Axel Dold
Formalisierung schematischer Algorithmen

94-11 Johannes Köbler, Osamu Watanabe
New Collapse Consequences of NP Having Small Circuits

94-12 Rainer Schuler
On Average Polynomial Time

94-13 Rainer Schuler, Osamu Watanabe
Towards Average-Case Complexity Analysis of NP Optimization Problems

94-14 Wolfram Schulte, Ton Vullinghs
Linking Reactive Software to the X-Window System

94-15 Alfred Lupper
Namensverwaltung und Adressierung in Distributed Shared Memory-Systemen

94-16 Robert Regn
Verteilte Unix-Betriebssysteme

94-17 Helmuth Partsch
Again on Recognition and Parsing of Context-Free Grammars: Two Exercises in Transformational Programming

94-18 Helmuth Partsch
Transformational Development of Data-Parallel Algorithms: an Example

95-01 Oleg Verbitsky
On the Largest Common Subgraph Problem

95-02 Uwe Schöning
Complexity of Presburger Arithmetic with Fixed Quantifier Dimension

95-03 Harry Buhrman,Thomas Thierauf
The Complexity of Generating and Checking Proofs of Membership

95-04 Rainer Schuler, Tomoyuki Yamakami
Structural Average Case Complexity

95-05 Klaus Achatz, Wolfram Schulte
Architecture Indepentent Massive Parallelization of Divide-And-Conquer Algorithms

95-06 Christoph Karg, Rainer Schuler
Structure in Average Case Complexity

95-07 P. Dadam, K. Kuhn, M. Reichert, T. Beuter, M. Nathe
ADEPT: Ein integrierender Ansatz zur Entwicklung flexibler, zuverlässiger kooperierender Assistenzsysteme in klinischen Anwendungsumgebungen

95-08 Jürgen Kehrer, Peter Schulthess
Aufbereitung von gescannten Röntgenbildern zur filmlosen Diagnostik

95-09 Hans-Jörg Burtschick, Wolfgang Lindner
On Sets Turing Reducible to P-Selective Sets

95-10 Boris Hartmann
Berücksichtigung lokaler Randbedingung bei globaler Zieloptimierung mit neuronalen Netzen am Beispiel Truck Backer-Upper

95-11 Thomas Beuter, Peter Dadam
Prinzipien der Replikationskontrolle in verteilten Systemen

95-12 Klaus Achatz, Wolfram Schulte
Massive Parallelization of Divide-and-Conquer Algorithms over Powerlists

95-13 Andrea Mößle, Heiko Vogler
Efficient Call-by-value Evaluation Strategy of Primitive Recursive Program Schemes

95-14 Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
A Generic Specification for Verifying Peephole Optimizations

96-01 Ercüment Canver, Jan-Tecker Gayen, Adam Moik
Formale Entwicklung der Steuerungssoftware für eine elektrisch ortsbediente Weiche mit VSE

96-02 Bernhard Nebel
Solving Hard Qualitative Temporal Reasoning Problems: Evaluating the Efficiency of Using the ORD-Horn Class

96-03 Ton Vullinghs, Wolfram Schulte, Thilo Schwinn
An Introduction to TkGofer

96-04 Thomas Beuter, Peter Dadam
Anwendungsspezifische Anforderungen an Workflow-Mangement-Systeme am Beispiel der Domäne Concurrent-Engineering

96-05 Gerhard Schellhorn, Wolfgang Ahrendt
Verification of a Prolog Compiler - First Steps with KIV

96-06 Manindra Agrawal, Thomas Thierauf
Satisfiability Problems

96-07 Vikraman Arvind, Jacobo Torán
A nonadaptive NC Checker for Permutation Group Intersection

96-08 David Cyrluk, Oliver Möller, Harald Rueß
An Efficient Decision Procedure for a Theory of Fix-Sized Bitvectors with Composition and Extraction

96-09 Bernd Biechele, Dietmar Ernst, Frank Houdek, Joachim Schmid, Wolfram Schulte
Erfahrungen bei der Modellierung eingebetteter Systeme mit verschiedenen SA/RT–Ansätzen

96-10 Falk Bartels, Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
Formalizing Fixed-Point Theory in PVS

96-11 Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
Mechanized Semantics of Simple Imperative Programming Constructs

96-12 Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
Generic Compilation Schemes for Simple Programming Constructs

96-13 Klaus Achatz, Helmuth Partsch
From Descriptive Specifications to Operational ones: A Powerful Transformation Rule, its Applications and Variants

97-01 Jochen Messner
Pattern Matching in Trace Monoids

97-02 Wolfgang Lindner, Rainer Schuler
A Small Span Theorem within P

97-03 Thomas Bauer, Peter Dadam
A Distributed Execution Environment for Large-Scale Workflow Management Systems with Subnets and Server Migration

97-04 Christian Heinlein, Peter Dadam
Interaction Expressions - A Powerful Formalism for Describing Inter-Workflow Dependencies

97-05 Vikraman Arvind, Johannes Köbler
On Pseudorandomness and Resource-Bounded Measure

97-06 Gerhard Partsch
Punkt-zu-Punkt- und Mehrpunkt-basierende LAN-Integrationsstrategien für den digitalen Mobilfunkstandard DECT

97-07 Manfred Reichert, Peter Dadam
$ADEPT_{flex}$ - Supporting Dynamic Changes of Workflows Without Loosing Control

97-08 Hans Braxmeier, Dietmar Ernst, Andrea Mößle, Heiko Vogler
The Project NoName - A functional programming language with its development environment

97-09 Christian Heinlein
Grundlagen von Interaktionsausdrücken

97-10 Christian Heinlein
Graphische Repräsentation von Interaktionsausdrücken

97-11 Christian Heinlein
Sprachtheoretische Semantik von Interaktionsausdrücken

97-12 Gerhard Schellhorn, Wolfgang Reif
Proving Properties of Finite Enumerations: A Problem Set for Automated Theorem Provers

97-13 Dietmar Ernst, Frank Houdek, Wolfram Schulte, Thilo Schwinn
Experimenteller Vergleich statischer und dynamischer Softwareprüfung für eingebettete Systeme

97-14 Wolfgang Reif, Gerhard Schellhorn
Theorem Proving in Large Theories

97-15 Thomas Wennekers
Asymptotik rekurrenter neuronaler Netze mit zufälligen Kopplungen

97-16 Peter Dadam, Klaus Kuhn, Manfred Reichert
Clinical Workflows - The Killer Application for Process-oriented Information Systems?

97-17 Mohammad Ali Livani, Jörg Kaiser
EDF Consensus on CAN Bus Access in Dynamic Real-Time Applications

97-18 Johannes Köbler,Rainer Schuler
Using Efficient Average-Case Algorithms to Collapse Worst-Case Complexity Classes

98-01 Daniela Damm, Lutz Claes, Friedrich W. von Henke, Alexander Seitz, Adelinde Uhrmacher, Steffen Wolf
Ein fallbasiertes System für die Interpretation von Literatur zur Knochenheilung

98-02 Thomas Bauer, Peter Dadam
Architekturen für skalierbare Workflow-Management-Systeme - Klassifikation und Analyse

98-03 Marko Luther, Martin Strecker
A guided tour through Typelab

98-04 Heiko Neumann, Luiz Pessoa
Visual Filling-in and Surface Property Reconstruction

98-05 Ercüment Canver
Formal Verification of a Coordinated Atomic Action Based Design

98-06 Andreas Küchler
On the Correspondence between Neural Folding Architectures and Tree Automata

98-07 Heiko Neumann, Thorsten Hansen, Luiz Pessoa
Interaction of ON and OFF Pathways for Visual Contrast Measurement

98-08 Thomas Wennekers
Synfire Graphs: From Spike Patterns to Automata of Spiking Neurons

98-09 Thomas Bauer, Peter Dadam
Variable Migration von Workflows in ADEPT

98-10 Heiko Neumann, Wolfgang Sepp
Recurrent V1 – V2 Interaction in Early Visual Boundary Processing

98-11 Frank Houdek, Dietmar Ernst, Thilo Schwinn
Prüfen von C–Code und Statmate/Matlab–Spezifikationen: Ein Experiment

98-12 Gerhard Schellhorn
Proving Properties of Directed Graphs: A Problem Set for Automated Theorem Provers

98-13 Gerhard Schellhorn, Wolfgang Reif
Theorems from Compiler Verification: A Problem Set for Automated Theorem Provers

98-14 Mohammad Ali Livani
SHARE: A Transparent Mechanism for Reliable Broadcast Delivery in CAN

98-15 Mohammad Ali Livani, Jörg Kaiser
Predictable Atomic Multicast in the Controller Area Network (CAN)

99-01 Susanne Boll, Wolfgang Klas, Utz Westermann
A Comparison of Multimedia Document Models Concerning Advanced Requirements

99-02 Thomas Bauer, Peter Dadam
Verteilungsmodelle für Workflow-Management-Systeme - Klassifikation und Simulation

99-03 Uwe Schöning
On the Complexity of Constraint Satisfaction

99-04 Ercument Canver
Model-Checking zur Analyse von Message Sequence Charts über Statecharts

99-05 Johannes Köbler, Wolfgang Lindner, Rainer Schuler

Derandomizing RP if Boolean Circuits are not Learnable

99-06 Utz Westermann, Wolfgang Klas
Architecture of a DataBlade Module for the Integrated Management of Multimedia Assets

99-07 Peter Dadam, Manfred Reichert
Enterprise-wide and Cross-enterprise Workflow Management: Concepts, Systems, Applications. Paderborn, Germany, October 6, 1999, GI–Workshop Proceedings, Informatik '99

99-08 Vikraman Arvind, Johannes Köbler
Graph Isomorphism is Low for ZPP^{NP} and other Lowness results

99-09 Thomas Bauer, Peter Dadam
Efficient Distributed Workflow Management Based on Variable Server Assignments

2000-02 Thomas Bauer, Peter Dadam
Variable Serverzuordnungen und komplexe Bearbeiterzuordnungen im Workflow-Management-System ADEPT

2000-03 Gregory Baratoff, Christian Toepfer, Heiko Neumann
Combined space-variant maps for optical flow based navigation

2000-04 Wolfgang Gehring
Ein Rahmenwerk zur Einführung von Leistungspunktsystemen

2000-05 Susanne Boll, Christian Heinlein, Wolfgang Klas, Jochen Wandel
Intelligent Prefetching and Buffering for Interactive Streaming of MPEG Videos

2000-06 Wolfgang Reif, Gerhard Schellhorn, Andreas Thums
Fehlersuche in Formalen Spezifikationen

2000-07 Gerhard Schellhorn, Wolfgang Reif (eds.)
FM-Tools 2000: The 4th Workshop on Tools for System Design and Verification

2000-08 Thomas Bauer, Manfred Reichert, Peter Dadam
Effiziente Durchführung von Prozessmigrationen in verteilten Workflow-Management-Systemen

2000-09 Thomas Bauer, Peter Dadam
Vermeidung von Überlastsituationen durch Replikation von Workflow-Servern in ADEPT

2000-10 Thomas Bauer, Manfred Reichert, Peter Dadam
Adaptives und verteiltes Workflow-Management

2000-11 Christian Heinlein
Workflow and Process Synchronization with Interaction Expressions and Graphs

2001-01 Hubert Hug, Rainer Schuler
DNA-based parallel computation of simple arithmetic

2001-02 Friedhelm Schwenker, Hans A. Kestler, Günther Palm
3-D Visual Object Classification with Hierarchical Radial Basis Function Networks

2001-03 Hans A. Kestler, Friedhelm Schwenker, Günther Palm
RBF network classification of ECGs as a potential marker for sudden cardiac death

2001-04 Christian Dietrich, Friedhelm Schwenker, Klaus Riede, Günther Palm
Classification of Bioacoustic Time Series Utilizing Pulse Detection, Time and Frequency Features and Data Fusion

2002-01 Stefanie Rinderle, Manfred Reichert, Peter Dadam
Effiziente Verträglichkeitsprüfung und automatische Migration von Workflow-Instanzen bei der Evolution von Workflow-Schemata

2002-02 Walter Guttmann
Deriving an Applicative Heapsort Algorithm

2002-03 Axel Dold, Friedrich W. von Henke, Vincent Vialard, Wolfgang Goerigk
A Mechanically Verified Compiling Specification for a Realistic Compiler

2003-01 Manfred Reichert, Stefanie Rinderle, Peter Dadam
A Formal Framework for Workflow Type and Instance Changes Under Correctness Checks

2003-02 Stefanie Rinderle, Manfred Reichert, Peter Dadam
Supporting Workflow Schema Evolution By Efficient Compliance Checks

2003-03 Christian Heinlein
Safely Extending Procedure Types to Allow Nested Procedures as Values

2003-04 Stefanie Rinderle, Manfred Reichert, Peter Dadam
On Dealing With Semantically Conflicting Business Process Changes.

2003-05 Christian Heinlein
Dynamic Class Methods in Java

2003-06 Christian Heinlein
Vertical, Horizontal, and Behavioural Extensibility of Software Systems

2003-07 Christian Heinlein
Safely Extending Procedure Types to Allow Nested Procedures as Values (Corrected Version)

2003-08 Changling Liu, Jörg Kaiser
Survey of Mobile Ad Hoc Network Routing Protocols)

2004-01 Thom Frühwirth, Marc Meister (eds.)
First Workshop on Constraint Handling Rules

2004-02 Christian Heinlein
Concept and Implementation of C+++, an Extension of C++ to Support User-Defined Operator Symbols and Control Structures

2004-03 Susanne Biundo, Thom Frühwirth, Günther Palm(eds.)
Poster Proceedings of the 27th Annual German Conference on Artificial Intelligence

2005-01 Armin Wolf, Thom Frühwirth, Marc Meister (eds.)
19th Workshop on (Constraint) Logic Programming

2005-02 Wolfgang Lindner (Hg.), Universität Ulm , Christopher Wolf (Hg.) KU Leuven
2. Krypto-Tag – Workshop über Kryptographie, Universität Ulm

2005-03 Walter Guttmann, Markus Maucher
Constrained Ordering

2006-01 Stefan Sarstedt
Model-Driven Development with ACTIVECHARTS, Tutorial

2006-02 Alexander Raschke, Ramin Tavakoli Kolagari
Ein experimenteller Vergleich zwischen einer plan-getriebenen und einer leichtgewichtigen Entwicklungsmethode zur Spezifikation von eingebetteten Systemen

2006-03 Jens Kohlmeyer, Alexander Raschke, Ramin Tavakoli Kolagari
Eine qualitative Untersuchung zur Produktlinien-Integration über Organisationsgrenzen hinweg

2006-04 Thorsten Liebig
Reasoning with OWL - System Support and Insights –

2008-01 H.A. Kestler, J. Messner, A. Müller, R. Schuler
On the complexity of intersecting multiple circles for graphical display

2008-02 Manfred Reichert, Peter Dadam, Martin Jurisch,l Ulrich Kreher, Kevin Göser, Markus Lauer
Architectural Design of Flexible Process Management Technology

2008-03 Frank Raiser
Semi-Automatic Generation of CHR Solvers from Global Constraint Automata

2008-04 Ramin Tavakoli Kolagari, Alexander Raschke, Matthias Schneiderhan, Ian Alexander
Entscheidungsdokumentation bei der Entwicklung innovativer Systeme für produktlinien-basierte Entwicklungsprozesse

2008-05 Markus Kalb, Claudia Dittrich, Peter Dadam
Support of Relationships Among Moving Objects on Networks

2008-06 Matthias Frank, Frank Kargl, Burkhard Stiller (Hg.)
WMAN 2008 – KuVS Fachgespräch über Mobile Ad-hoc Netzwerke

2008-07 M. Maucher, U. Schöning, H.A. Kestler
An empirical assessment of local and population based search methods with different degrees of pseudorandomness

2008-08 Henning Wunderlich
Covers have structure

2008-09 Karl-Heinz Niggl, Henning Wunderlich
Implicit characterization of FPTIME and NC revisited

2008-10 Henning Wunderlich
On span-P$^c c$ and related classes in structural communication complexity

2008-11 M. Maucher, U. Schöning, H.A. Kestler
On the different notions of pseudorandomness

2008-12 Henning Wunderlich
On Toda's Theorem in structural communication complexity

2008-13 Manfred Reichert, Peter Dadam
Realizing Adaptive Process-aware Information Systems with ADEPT2

2009-01 Peter Dadam, Manfred Reichert
The ADEPT Project: A Decade of Research and Development for Robust and Fexible Process Support Challenges and Achievements

2009-02 Peter Dadam, Manfred Reichert, Stefanie Rinderle-Ma, Kevin Göser, Ulrich Kreher, Martin Jurisch
Von ADEPT zur AristaFlow® BPM Suite – Eine Vision wird Realität "Correctness by Construction" und flexible, robuste Ausführung von Unternehmensprozessen

2009-03 Alena Hallerbach, Thomas Bauer, Manfred Reichert
Correct Configuration of Process Variants in Provop

2009-04 Martin Bader
On Reversal and Transposition Medians

2009-05 Barbara Weber, Andreas Lanz, Manfred Reichert
Time Patterns for Process-aware Information Systems: A Pattern-based Analysis

2009-06 Stefanie Rinderle-Ma, Manfred Reichert
Adjustment Strategies for Non-Compliant Process Instances

2009-07 H.A. Kestler, B. Lausen, H. Binder H.-P. Klenk. F. Leisch, M. Schmid
Statistical Computing 2009 – Abstracts der 41. Arbeitstagung

2009-08 Ulrich Kreher, Manfred Reichert, Stefanie Rinderle-Ma, Peter Dadam
Effiziente Repräsentation von Vorlagen- und Instanzdaten in Prozess-Management-Systemen

2009-09 Dammertz, Holger, Alexander Keller, Hendrik P.A. Lensch
Progressive Point-Light-Based Global Illumination

2009-10 Dao Zhou, Christoph Müssel, Ludwig Lausser, Martin Hopfensitz, Michael Kühl, Hans A. Kestler

Boolean networks for modeling and analysis of gene regulation

2009-11 J. Hanika, H.P.A. Lensch, A. Keller
Two-Level Ray Tracing with Recordering for Highly Complex Scenes

2009-12 Stephan Buchwald, Thomas Bauer, Manfred Reichert
Durchgängige Modellierung von Geschäftsprozessen durch Einführung eines Abbildungsmodells: Ansätze, Konzepte, Notationen

2010-01 Hariolf Betz, Frank Raiser, Thom Frühwirth
A Complete and Terminating Execution Model for Constraint Handling Rules

2010-02 Ulrich Kreher, Manfred Reichert
Speichereffiziente Repräsentation instanzspezifischer Änderungen in Prozess-Management-Systemen

2010-03 Patrick Frey
Case Study: Engine Control Application

2010-04 Matthias Lohrmann und Manfred Reichert
Basic Considerations on Business Process Quality

2010-05 HA Kestler, H Binder, B Lausen, H-P Klenk, M Schmid, F Leisch (eds):
Statistical Computing 2010 - Abstracts der 42. Arbeitstagung

2010-06 Vera Künzle, Barbara Weber, Manfred Reichert
Object-aware Business Processes: Properties, Requirements, Existing Approaches

2011-01 Stephan Buchwald, Thomas Bauer, Manfred Reichert
Flexibilisierung Service-orientierter Architekturen

2011-02 Johannes Hanika, Holger Dammertz, Hendrik Lensch
Edge-Optimized À-Trous Wavelets for Local Contrast Enhancement with Robust Denoising

2011-03 Stefanie Kaiser, Manfred Reichert
Datenflussvarianten in Prozessmodellen: Szenarien, Herausforderungen, Ansätze

2011-04 Hans A. Kestler, Harald Binder, Matthias Schmid, Friedrich Leisch, Johann M. Kraus (eds):
Statistical Computing 2011 - Abstracts der 43. Arbeitstagung

2011-05 Vera Künzle, Manfred Reichert
PHILharmonicFlows: Research and Design Methodology

2011-06 David Knuplesch, Manfred Reichert
Ensuring Business Process Compliance Along the Process Life Cycle

2011-07 Marcel Dausend
Towards a UML Profile on Formal Semantics for Modeling Multimodal Interactive Systems

2011-08 Dominik Gessenharter
Model-Driven Software Development with ACTIVECHARTS - A Case Study

2012-01 Andreas Steigmiller, Thorsten Liebig, Birte Glimm
Extended Caching, Backjumping and Merging for Expressive Description Logics

2012-02 Hans A. Kestler, Harald Binder, Matthias Schmid, Johann M. Kraus (eds):
Statistical Computing 2012 - Abstracts der 44. Arbeitstagung

2012-03 Felix Schüssel, Frank Honold, Michael Weber
Influencing Factors on Multimodal Interaction at Selection Tasks

2012-04 Jens Kolb, Paul Hübner, Manfred Reichert
Model-Driven User Interface Generation and Adaption in Process-Aware Information Systems

2012-05 Matthias Lohrmann, Manfred Reichert
Formalizing Concepts for Efficacy-aware Business Process Modeling

2012-06 David Knuplesch, Rüdiger Pryss, Manfred Reichert
A Formal Framework for Data-Aware Process Interaction Models

2012-07 Clara Ayora, Victoria Torres, Barbara Weber, Manfred Reichert, Vicente Pelechano
Dealing with Variability in Process-Aware Information Systems: Language Requirements, Features, and Existing Proposals

2013-01 Frank Kargl
Abstract Proceedings of the 7th Workshop on Wireless and Mobile Ad-Hoc Networks (WMAN 2013)

2013-02 Andreas Lanz, Manfred Reichert, Barbara Weber
A Formal Semantics of Time Patterns for Process-aware Information Systems

2013-03 Matthias Lohrmann, Manfred Reichert
Demonstrating the Effectiveness of Process Improvement Patterns with Mining Results

2013-04 Semra Catalkaya, David Knuplesch, Manfred Reichert
Bringing More Semantics to XOR-Split Gateways in Business Process Models Based on Decision Rules

2013-05 David Knuplesch, Manfred Reichert, Linh Thao Ly, Akhil Kumar, Stefanie Rinderle-Ma
On the Formal Semantics of the Extended Compliance Rule Graph

2013-06 Andreas Steigmiller, Birte Glimm, Thorsten Liebig
Nominal Schema Absorption

2013-07 Hans A. Kestler, Matthias Schmid, Florian Schmid, Dr. Markus Maucher, Johann M. Kraus (eds)
Statistical Computing 2013 - Abstracts der 45. Arbeitstagung

2013-08 Daniel Ott, Dr. Alexander Raschke

Evaluating Benefits of Requirement Categorization in Natural Language Specifications for Review Improvements

2013-09 Philip Geiger, Rüdiger Pryss, Marc Schickler, Manfred Reichert
Engineering an Advanced Location-Based Augmented Reality Engine for Smart Mobile Devices

2014-01 Andreas Lanz, Manfred Reichert
Analyzing the Impact of Process Change Operations on Time-Aware Processes

2014-02 Andreas Steigmiller, Birte Glimm, and Thorsten Liebig
Coupling Tableau Algorithms for the DL SROIQ with Completion-based Saturation Procedures

2014-03 Thomas Geier, Felix Richter, Susanne Biundo
Conditioned Belief Propagation Revisited: Extended Version

2014-04 Hans A. Kestler, Matthias Schmid, Ludwig Lausser, Johann M. Kraus (eds)
Statistical Computing 2014 - Abstracts der 46. Arbeitstagung

2014-05 Andreas Lanz, Roberto Posenato, Carlo Combi, Manfred Reichert
Simple Temporal Networks with Partially Shrinkable Uncertainty (Extended Version)

2014-06 David Knuplesch, Manfred Reichert
An Operational Semantics for the Extended Compliance Rule Graph Language

2015-01 Andreas Lanz, Roberto Posenato, Carlo Combi, Manfred Reichert
Controlling Time-Awareness in Modularized Processes (Extended Version)

2015-03 Raphael Frank, Christoph Sommer, Frank Kargl, Stefan Dietzel, Rens W. van der Heijden
Proceedings of the 3rd GI/ITG KuVS Fachgespräch Inter-Vehicle Communication (FG-IVC 2015)

2015-04 Axel Fürstberger, Ludwig Lausser, Johann M. Kraus, Matthias Schmid, Hans A. Kestler (eds)
Statistical Computing 2015 - Abstracts der 47. Arbeitstagung

2016-03 Ping Gong, David Knuplesch, Manfred Reichert
Rule-based Monitoring Framework for Business Process Compliance

2016-04 Axel Fürstberger, Ludwig Lausser, Johann M. Kraus, Matthias Schmid, Hans A. Kestler (eds)
Statistical Computing 2016 - Abstracts der 48. Arbeitstagung

2016-05 Axel Fürstberger, Johann M. Kraus, Hans A. Kestler (eds)
Classification 2016 - Abstracts of the 5th German-Japanese Symposium

2016-06 Vera Künzle, Sebastian Steinau, Kevin Andrews, and Manfred Reichert
An Approach for Modeling and Coordinating Process Interactions

2017-01 Hans A. Kestler, Matthias Schmid, Ludwig Lausser, Johann M. Kraus, Axel Fürstberger (eds)
Statistical Computing 2017 - Abstracts der 49. Arbeitstagung

2018-01 Hans A. Kestler, Matthias Schmid, Ludwig Lausser, Johann M. Kraus, Axel Fürstberger (eds)
Statistical Computing 2018 - Abstracts der 50. Arbeitstagung

2019-01 Hans A. Kestler, Matthias Schmid, Ludwig Lausser, Axel Fürstberger (eds)
Statistical Computing 2019 - Abstracts der 51. Arbeitstagung

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm