ulm university universität
uulm

Statistical Computing 2016

Abstracts der 48. Arbeitstagung

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

Ulmer Informatik-Berichte

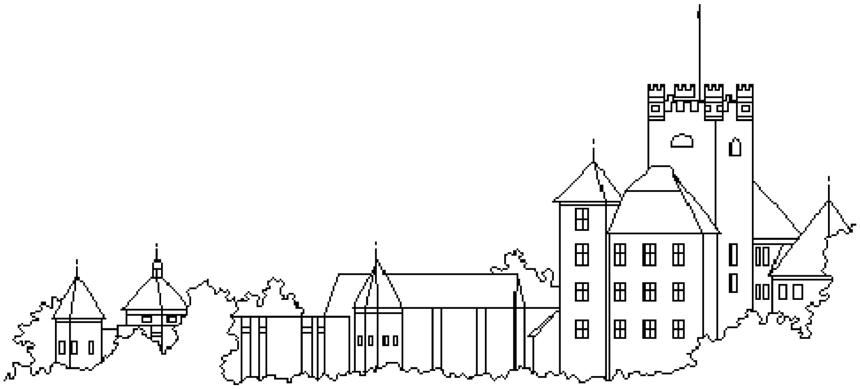
Nr. 2016-04
July 2016

International Graduate School
in Molecular Medicine Ulm

SYSTAR

Ulmer Informatik-Berichte | Universität Ulm | Fakultät für Ingenieurwissenschaften und Informatik

Statistical Computing 2016



48. Arbeitstagung

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

17.07. - 20.07.2016, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, July 17, 2016

18:00-20:00		**Dinner**
20:00-21:00		**Chair: H. A. Kestler**
20:00-21:00	Eyke Hüllermeier (Paderborn)	Superset learning and data imprecisiation

Monday, July 18, 2016

09:15-09:30		**Opening of the workshop: H. A. Kestler, M. Schmid**
09:30-10:30		**Chair: A. Groß**
09:30-09:50	Rainer Schuler (Ulm)	Counting the number of molecules of a single mRNA species
09:50-10:10	Jörn Lötsch (Frankfurt)	Process pharmacology: Using computational functional genomics knowledge to connect drugs with biological processes
10:10-10:30	Sebastian Krey (Köln)	Modeling and Calibration of robust Gas Sensors
10:30-11:00		**Coffee Break**
11:00-15:00		**Chair: M. Schmid**
11:00-12:00	Peter Chronz (Göttingen)	Introduction to Julia
12:00-13:30		**Lunch**
13:30-15:00	Tutorial: Peter Chronz (Göttingen)	Hands-on Introduction to Julia: Simplicity, Expressiveness, and Performance
15:00-15:30		**Coffee Break**
15:30-21:00	Social Program (Ulm)	Nabada (Schwörmontag)

Tuesday, July 19, 2016

09:30-10:30		**Chair: L. Lausser**
09:30-09:50	Janek Thomas (München)	Stability selection for boosted generalized additive models for location scale and shape
09:50-10:10	Tobias Hepp (Erlangen)	Assessing the significance of effects in boosted location and scale models
10:10-10:30	Andreas Mayr (Erlangen)	Boosting distributional regression models for multivariate responses
10:30-11:00		**Coffee Break**
11:00-12:00		**Chair: B. Bischl**
11:00-11:20	Sarah Friedrich (Ulm)	GFD: An R-package for the Analysis of General Factorial Designs - along with a Graphical User Interface
11:20-11:40	Thomas Welchowski (Bonn)	kernDeepStackNet: An R package for tuning kernel deep stacking networks
11:40-12:00	Julian Schwab (Ulm)	Implementation and Simulation of Boolean Networks on FPGAs
12:00-13:30		**Lunch**
13:30-15:00		**Chair: B. Lausen**
13:30-13:50	Moritz Hanke (Bremen)	A small REvolutioN and modified temporal centrality measures are needed for incomplete graph sequences of dynamic networks
13:50-14:10	Andre Burkovski (Ulm)	Performance of ordinal-scaled prototype-based classifiers on microarray datasets
14:10-14:30	Elisabeth Waldmann (Erlangen)	Boosting Joint Models for Longitudinal and Time-to-Event Data
14:30-15:00	Thomas Villmann (Mittweida)	Classification Certainty and Reject Options in Learning Vector Quantization
15:00-15:30		**Coffee Break**
15:30-16:50		**Chair: E. Sträng**
15:30-15:50	Bernd Bischl (München)	Multi-Objective Parameter Configuration of Machine Learning Algorithms using Model-Based Optimization
15:50-16:10	Philipp Probst (München)	On the Hyperparameter Settings of Random Forest
16:10-16:30	Lyn-Rouven Schirra (Ulm)	Feature selecting multi-class classification
16:30-16:50	Laura Beggel (München)	Anomaly Detection with Shapelet-Based Feature Learning for Time Series
17:00-18:00		Working group meeting on **Statistical Computing 2017** and other topics (all welcome)
18:00-20:00		**Dinner**

Wednesday, July 20, 2016

09:30-10:30		**Chair: J. Kraus**
09:30-09:50	Alexander Engelhardt (München)	Implementing an EM algorithm for partially dependent data
09:50-10:10	Pascal Schlosser (Freiburg)	Netboost: Boosting supported network analysis for highdimensional genomic datasets
10:10-10:30	Leonie Weinhold (Bonn)	A Statistical Model for the Analysis of Beta Values in DNA Methylation Studies
10:30-11:00		**Coffee Break**
11:00-12:00		**Chair: E. Hüllermeier**
11:00-11:20	Gunnar Völkel (Ulm)	Automated Design of Search Algorithms for Feature Set Ensembles
11:20-11:40	Werner Adler (Erlangen)	Ensemble Pruning for Glaucoma Detection
11:40-12:00	Berthold Lausen (Colchester)	Ensemble of selected classifiers
12:00-13:30		**Lunch**

Contents

Superset learning and data imprecisiation

Eyke Hüllermeier [1]

Superset learning is a generalization of standard supervised learning, in which training instances are labeled with a superset of the actual outcomes. Thus, superset learning can be seen as a specific type of weakly supervised learning, in which training examples are imprecise or ambiguous. We introduce a generic approach to superset learning, which is motivated by the idea of performing model identification and "data disambiguation" simultaneously. This idea is realized by means of a generalized risk minimization approach, using an extended loss function that compares precise predictions with set-valued observations. Building on this approach, we furthermore elaborate on the idea of "data imprecisiation": By deliberately turning precise training data into imprecise data, it becomes possible to modulate the influence of individual examples on the process of model induction. In other words, data imprecisiation offers an alternative way of instance weighting. Interestingly, several existing machine learning methods, such as support vector regression or semi-supervised support vector classification, are recovered as special cases of this approach. Besides, promising new methods can be derived in a natural way.

[1] Department of Computer Science, Paderborn University, Pohlweg 51, 33098 Paderborn, Germany

eyke@upb.de

Counting the number of molecules of a single mRNA species

Rainer Schuler[1]

Counting individual RNA or DNA molecules is a challenging task, because it is difficult to amplify quantitatively for detection. One approach is to mark the molecules with a *molecular barcode* i.e. a *unique molecular identifier* thereby making them unique. This method turns a quantitative problem of counting the number of molecules having the same sequence into a qualitative problem of detecting the number of different molecules.

The method includes four basic steps: 1) Isolate molecules of interest, 2) turn each molecule into a different molecular species (DMS), 3) amplify the DMS, 4) detect and count the number of DMS.
To implement the second step, oligonucleotides (tags) of a fixed length but with different sequences are used to label the molecules. Since hybridization of DNA molecules with identical sequences is competitive the tags are selected randomly. A collision occurs if two (or more) molecules are labeled with tags of the same sequence. We estimate the expected number of identical tags and consider the relationship between the observed number of different tag sequences (DMS) and the number of molecules.

References

G. K. Fu, J. Hu, P.-H. Wang, and S. P. A. Fodor, *Counting individual DNA molecules by the stochastic attachment of diverse labels*, PNAS, (2011).

B. Hollas and R. Schuler, *A stochastic approach to count RNA molecules using DNA sequencing methods*, in Algorithms in Bioinformatics: WABI 2003. Proceedings, G. Benson and R. D. M. Page, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 55–62.

H. Hug and R. Schuler, *Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation*, J. theor. Biol., (2003).

K. Karlsson, *Counting molecules in cell-free DNA and single cells RNA*, PhD thesis, Dept of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna, 2016.

T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale, *Counting absolute numbers of molecules using unique molecular identifiers*, Nature Methods, (2012).

[1] Institute of Medical Systems Biology, Ulm University, James-Franck-Ring, 89081 Ulm, Germany

rainer.schuler@uni-ulm.de

Process pharmacology: Using computational functional genomics knowledge to connect drugs with biological processes

Jörn Lötsch[1,2] and Alfred Ultsch[3]

Functional genomics investigates the biochemical, cellular, and/or physiological properties of each and every gene product [1] with the goal of understanding the relationship between genome and phenotype on a genome-wide scale. The acquired knowledge about the functions of gene products is provided in publicly accessible databases of which the current gold standard is the Gene Ontology (GO) database [2, 3]. The combination of this information with the acquired knowledge about the interaction of drugs with gene products is the basis of a recently introduced data science approach to pharmacology that links drugs directly with diseases [4]. This is regarded as the result of pathophysiological processes that are captured by the GO category “biological processes”.

A focus on biological processes [4] may provide a phenotypic approach to drug discovery and repurposing based on (i) selecting disease relevant biological processes as therapeutic targets of novel drugs and (ii) evaluating the utility of known drugs for new indications on the basis of their effects on disease-relevant biological processes. This requires a functional genomics based criterion of drug classification with equivalent performance of the genomics respectively drug target based criterion. Based on prior hints that the functional genomics based criterion provides reasonable drug classification accommodating current pharmacological concepts [4], the present analysis pursued the hypothesis that both criteria provide similarly correct drug classifications. In a comparative assessment using machine-learned techniques for computational drug classification, both criteria were compared in order to provide support for the utility of a functional genomics based criterion for drug development.

[1] Institute of Clinical Pharmacology, Goethe - University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany

[2] Fraunhofer Institute of Molecular Biology and Applied Ecology - Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany

[3] DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany

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

References

1 Gibson G, Muse SV. A Primer of genome science. Sunderland, Massachusetts: Sinauer Associates; 2009.

2 Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004;32(Database issue):D258-61.

3 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

4 Lötsch J, Ultsch A. Process Pharmacology: A Pharmacological Data Science Approach to Drug Development and Therapy. CPT Pharmacometrics Syst Pharmacol. 2016.

5 Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36(Database issue):D901-6.

Modeling and Calibration of robust Gas Sensors

Sebastian Krey[1], Margarita Alejandra Rebolledo Coy[1], Jörg Stork[1], Thomas Bartz-Beielstein[1]

In our society's pursuit to reduce air pollution and counteract global warming the reduction of carbon dioxide emission and other air pollutants is an important aim of all industrial sectors. Especially the automotive segment with its problems to reach the rising standards for carbon dioxide and nitrogen dioxide reduction has recently been in the news. To reach the ecological stricter standards the in-situ measurement of the different exhaust gas components gains importance. Based on these measurements a better control of combustion processes allows a more efficient and cleaner operation.

While the in-situ measurement of the oxygen percentage in exhaust gases is industrial standard, the development of long term durable sensors for the selective in-situ measurement of other gas components has been an unsolved problem. The usage of experimental design and linear models helped our industrial research partners to develop durable sensors with different sensitivities and optimize the sensor output for a higher signal quality. These sensors still measure only gas mixtures, but based on an array of different sensors individual readings for the different gas components can be calculated.

In this work we present how (frequentist and bayesian) linear models compare in this scenario with very noisy data against more computational intensive methods (Support Vector Regression, Kriging, Symbolic Regression using Genetic Programming) in modeling the sensors response as well as in creating a calibration model for the prediction of the different gas concentrations. Additionally we present a method to allow an easy and efficient recalibration of the measurement system if the working conditions change or to adapt the system to the effects of sensor aging.

[1] Technische Hochschule Köln, Fakultät für Informatik und Ingenieurwissenschaften, Steinmüllerallee 1, 51643 Gummersbach

sebastian.krey@th-koeln.de, margarita.rebolledo@th-koeln.de, joerg.stork@th-koeln.de, thomas.bartz-beielstein@th-koeln.de

Hands-on Introduction to Julia: Simplicity, Expressiveness, and Performance

Peter Chronz [1]

This workshop introduces the open source programming language Julia for technical computing. Julia aims to offer a replacement for languages such as C, Fortran, Python, Matlab, and R for technical computing. Julia's development is motivated by the growing gap between processing requirements for data analysis and the capabilities of current programming languages and computing platforms. Currently, data processing is often performed with languages that roughly fall into two categories. The first category consists of languages, such as Fortran and C, that offer high performance but are tedious to program. The second category contains languages, such as Python and R, that offer convenient abstractions, but poor performance. Julia overcomes this dichotomy by merging convenient, high-level abstractions and high performance at the same time. This combination empowers programmers to quickly prototype and achieve excellent performance with the same code. To combine high-level abstraction and high performance, Julia's developers leverage the modern features of the LLVM compiler chain. One important feature leveraged by Julia is just-in-time compilation. Compiling code on the fly allows Julia to perform type inference at runtime and in effect to achieve high performance even if the code is dynamically typed. Concretely, Julia reaches nearly the performance of C for a set typical data processing algorithms. The workshop aims to teach the participants enough to start coding productively afterwards. The workshop covers the language fundamentals in 5 sections. First, basics, such as strings, functions, control flow, and multi-dimensional arrays are presented. Second, an introduction to plotting packages, which are similar to R's ggplot2 and Python's matplotlib, are introduced. Third, data management with DataFrames is covered. Fourth, Julia's package for generalized linear models is demonstrated. Finally, we will cover parallel programming with built-in methods. To learn as much as possible about Julia for productive use, the workshop offers exercises on all topics. The second phase of the workshop consists mostly of hands-on experience.

[1] Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

`peter.chronz@gwdg.de`

Stability selection for boosted generalized additive models for location scale and shape

Janek Thomas[1]

In this project, a new algorithm to incorporate stability selection for boosting generalized additive models for location, scale and shape (GAMLSS) is presented. In one motivating application, a negative binomial hurdle model was fitted to handle excess zeros, overdispersion, non-linearity and spatiotemporal structures to investigate the abundance of the common eider in Massachusetts. The number of birds was estimated via boosting GAMLSS, allowing both mean and overdispersion to be regressed on covariates and incorporating variable selection.An increasingly popular way to obtain stable sets of covariates while controlling the false discovery rate (FDR) is stability selection. The model is fitted repeatedly to subsampled data and variables with high selection frequencies are extracted.Currently, this leads to a fundamental problem with boosted GAMLSS, where in every boosting iteration, the algorithm sequentially selects the best fitting effect for each distribution parameter. Thus, it is currently not possible to stop fitting individual parameters as soon as they are sufficiently modeled.In order to solve this problem, we developed a new approach to fit boosted GAMLSS. Instead of updating all distribution parameters in each iteration, only the update of the parameter which leads to the biggest reduction in loss is performed. With this modification, the stability selection framework can be applied. Furthermore, optimizing the tuning parameters of boosting is reduced from a multi-dimensional to a one-dimensional problem.The performance of the algorithm is evaluated in a simulation study and the application is demonstrated for the seabirds data, selecting stable predictors while controlling the FDR.

[1] Department of statistics, Ludwig-Maximilians-University Germany

janek.thomas@stat.uni-muenchen.de

Assessing the significance of effects in boosted location and scale models

Tobias Hepp[1], Matthias Schmid[2], Andreas Mayr[1,2]

While providing many advantages like automatic variable selection and feasibility in high-dimensional settings with more predictors than observations, the implicit regularization of gradient boosting algorithms that allows the shrinkage of effect estimates prevents the computation of standard errors As a result, the construction of confidence intervals or significance tests is problematic. To overcome this problem, Mayr et al.[1] proposed the use of permutation tests to derive p-values for the effect estimates of boosted location and scale models [2], which rely on the independence of the predictor of interest with all other covariates in the model.

In this talk, we therefore discuss the performance of two alternative approaches to extend the scope of possible application scenarios. In order to remove the correlations with other predictors, the first option is based on the replacement of the variable of interest with regression residuals, which can then be equally used for the permutation tests [2]. Another alternative is to draw new samples from the conditional distribution of the constrained model without the variable of interest. Then, the differences in the quality of fit between the full model and the constrained one are attributable only to the randomness of these parametric bootstrap samples and should be less distinctive than for the original data.

All models are fitted via the R add-on package `gamboostLSS`. In addition, the results are compared to the parametric Wald-type tests implemented in the `gamlss`-package [4].

References

1 Mayr, A., Schmid, M., Pfahlberg, A., Uter, W. and Gefeller, O. (2015). A permutation test to analyse systematic bias and random measurement errors of medical devices via boosting location and scale models. *Statistics in Medicine.* 24(5), 693-708.

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.

3 Potter, DM. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Statistics in Medicine.* 24(5), 693-708.

4 Rigby, RA. and Stasinopoulos, DM. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* 54(3), 507–554.

[1] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg

[2] Department of Medical Biometrics, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn

tobias.hepp@uk-erlangen.de

Boosting distributional regression models for multivariate responses

Andreas Mayr[1,2], Janek Thomas[3], Matthias Schmid[2], Nadja Klein[4]

Over the last few years, statistical modelling approaches that go beyond the classical regression of the conditional mean have gained more and more attention. One of the most popular approaches in this context are generalized additive models for location, scale and shape (GAMLSS, [1]). The main idea of GAMLSS is that each parameter of the conditional distribution – not only the expected value – is modelled by its own additive predictor.

We extend this approach towards multivariate responses [2] and present a statistical boosting algorithm that is able to estimate the unknown quantities of these complex models in potentially high-dimensional settings by circling through the different parameter and outcome dimensions [3].

Our approach will be illustrated by an epidemiological study where the factors influencing children's growth regarding their height and weight are analysed simultaneously in a longitudinal setting.

References

1 Rigby, R. A. and Stasinopoulos, D. M. (2005): Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C - Applied Statistics.* 54(3), 507–554.

2 Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015): Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society Series C - Applied Statistics.*, 64(4), 569–591.

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 Medizininformatik, Biometrie und Epidemiologie, FAU Erlangen-Nürnberg

[2] Institut für Medizinische Biometrie, Informatik und Epidemiologie, Rheinische Friedrich-Wilhelms-Universität Bonn

[3] Institut für Statistik, LMU München

[4] Chair of Statistics, Georg-August-Universität Göttingen

andreas.mayr@fau.de

GFD: An R-package for the Analysis of General Factorial Designs - along with a Graphical User Interface

Sarah Friedrich[1] *,Frank Konietschke*[2] *and Markus Pauly*[1]

Factorial designs are widely used tools for modeling statistical experiments in all kinds of disciplines, e.g., biology, psychology, econometrics and medicine. For testing null hypotheses formulated in terms of means, analysis-of-variance (ANOVA) methods are well known, and preferred for making statistical inference. ANOVA methods are implemented in R within the functions anova and lm in the R-package stats, along with clearly arranged ANOVA tables in the output. The corresponding F tests, however, are only valid for normally distributed data with equal variances, two assumptions which are often not met in practice. The R-package GFD provides implementation of the Wald-type statistic (WTS), the ANOVA-type statistic (ATS) and a studentized permutation version of the WTS as in [1]. Both the WTS and the permuted WTS do not require normally distributed data or variance homogeneity, whereas the ATS assumes normality. All methods are available for general crossed or nested designs and all main and interaction effects can be plotted. Additionally, the package is equipped with an optional graphical user interface to facilitate application for a wide range of users. We illustrate the implemented methods for a range of different designs.

References

1 Pauly, M., Brunner, E. and Konietschke, F. (2015). Asymptotic Permutation Tests in General Factorial Designs. Journal of the Royal Statistical Society - Series B, 77, 461–473.

2 Friedrich, S., Konietschke, F. and Pauly, M. (2015). GFD: An R-package for the Analysis of General Factorial Designs - along with a Graphical User Interface. Submitted preprint.

[1] Institute of Statistics, Ulm University, Ulm, Germany
[2] Department of Statistics, University of Texas at Dallas

`sarah.friedrich@uni-ulm.de, fxk141230@utdallas.edu, markus.pauly@uni-ulm.de`

kernDeepStackNet: An R package for tuning kernel deep stacking networks

Thomas Welchowski[1] and Matthias Schmid[1]

Kernel deep stacking networks [1] (KDSNs) are a novel method for supervised learning in biomedical research and belong to the class of deep learning methods. Deep learning uses multiple layers of non-linear transformations to derive higher abstractions of the input features [2]. These can more efficiently represent complex dependencies of joint distributions [3]. Training of deep artificial neural networks is a non-convex optimization problem, which may result in local optima and slow convergence. Kernel deep stacking networks are an computational faster alternative, which are based on solving multiple convex optimization problems by combined kernel ridge regressions with random Fourier transformations.

Tuning of KDSNs is a challenging task, as there are multiple hyper parameters to tune. We propose a new data-driven tuning strategy for KDSNs using model based optimization (MBO) [4]. The performance criterion is RMSE on cross validation samples, and noisy Kriging is used as surrogate model. New design points are choosen by maximisation of the expected improvement criterion.

Numerical studies show, that the MBO approach is substantially faster than traditional grid search strategies. Further analysis of real data sets demonstrate, that tuned KDSNs are competetive to other state-of-art machine learning techniques in terms of prediction accuracy. The fitting and tuning procedures are implemented in the R package *kern-DeepStackNet.*

References

1 Huang P-S, Deng L, Hasegawa-Johnson M, He X. , 2013, Random features for kernel deep convex network. In: Ward R, Deng L, editors. Proceedings of the IEEE international conference on acoustics, speech, and signal processing, NewYork, USA; pages 3143–7.

2 Li Deng, Dong Yu, 2014, Deep learning: Methods and Applications, Foundations and Trends®in Signal Processing, Volume 7, Issues 3-4

3 Bengio Y, Delalleau O. On the expressive power of deep architectures. In: Kivinen J, Szepesvari C, Ukkonen E, Zeugmann T, editors. Algorithmic learning theory. Berlin: Springer; 2011. p. 18–36.

4 Welchowski, Schmid, 2016, A framework for parameter estimation and model selection in kernel deep stacking networks, Artificial Intelligence in Medicine, Volume 70, Pages 31–40

[1] Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universitat Bonn, Sigmund-Freud-Str. 25, 53127 Bonn,Germany

welchow@imbie.meb.uni-bonn.de,matthias.schmid@imbie.uni-bonn.de

Implementation and Simulation of Boolean Networks on FPGAs

Julian Schwab[1], Andre Burkovski[1], Johann M. Kraus[1], Hans A. Kestler[1]

In Systems Biology mathematical models are often used to gain insights into cellular pathways and regulatory networks. If only qualitative knowledge is available, Boolean networks can provide important insights into the dynamic behavior of complex regulatory systems.

Boolean network models are described as a set of Boolean variables $X = \{x_1, ..., x_n\}$ and a set of Boolean functions $F = \{f_1, ..., f_n\}$, one variable x_i and one corresponding Boolean function $f_i : \mathbb{B}^n \to \mathbb{B}$ for each regulatory factor and its interaction within the system. The successor state of a regulatory factor x_i at time t is determined by the function $x(t+1) = f_i(\mathbf{x}(t))$ with $\mathbf{x}(t) = (x_1(t), ..., x_n(t))$ [1]. In synchronous Boolean networks a state transition is performed by updating each regulatory factor at the same time.

State transitions from each state in the network eventually lead to a recurrent cycle of states. These stationary cycles are called attractors. Attractors describe the longterm behavior of Boolean networks and can often be linked to biological phenotypes [2].

To search a network for attractors exhaustively, the search algorithm has to be executed from each of the 2^n possible states in a network with n regulatory factors. This leads to an exponential growth $\mathcal{O}(2^n)$ of computation time and memory consumption for an exhaustive attractor search [3]. Due to this fact, exhaustive attractor search is limited to a comparatively small number of regulatory factors.

Aim of this work is to simulate Boolean networks on field programmable gate arrays (FPGAs). We use FPGAs, specialized hardware chips, to accelerate the attractor search. The integration of FPGAs makes attractor search faster and allows simulation of larger networks than in conventional software solutions. We developed an algorithm to search for attractors on FPGAs using VHDL, a hardware description language. The implementation on the FPGA has a modular design, which allows to adapt the design to new Boolean networks readily.

Our first results showed that, due to the acceleration of the FPGA, we achieved a reduction of computation time in orders of magnitude.

[1] Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm

julian.schwab@uni-ulm.de

References

1 Müssel, C., Hopfensitz, M., and Kestler, H. A. Boolnet an R package for generation, reconstruction and analysis of boolean networks. *Bioinformatics 26*, 10 (2010), 1378–1380.

2 Kauffman, S. The origins of order. Self-organization and selection in evolution. *Journal of Evolutionary Biology 7*, 4 (1994), 518–519.

3 Hopfensitz, M., Müssel, C., Maucher, M., and Kestler, H. A. Attractors in boolean networks: a tutorial. *Computational Statistics 28*, 1 (2013), 19–36.

A small REvolutioN and modified temporal centrality measures are needed for incomplete graph sequences of dynamic networks

Moritz Hanke[1], Ronja Foraita[1]

Statistical analysis of networks is gaining importance in areas like social, computer and life sciences. Different centrality measures (e.g. betweenness, closeness) exist for static networks to capture different aspects of a single vertex' importance within the network. However, in reality many processes defined on networks are rather dynamic than as static (e.g. spread of disease, gene cell regulation). Most approaches represent such dynamic networks as sequences of static graphs (so called graph snapshots). Based on these snapshots, for example Tang et al. [1] and Kim and Anderson [2] extended classical vertice centrality measures to take the network dynamics over time into account. For this purpose they assumed to have complete information about the dynamic network, i.e. that the sequence of snapshots is containing all dynamics within the network over time. In real world applications this assumption could often be untenable due to limited access to raw network data which might bias the observed centrality values.

To account for this incompleteness we propose the idea of adding extra snapshots to the observed graph sequence. Based on this idea we extended two of the original temporal centrality metrics of Kim and Anderson [2] by cloning observed snapshots as a first and easy implementation. We show for different simulated scenarios of incomplete graph sequences that our approach increases the accuracy of detecting important vertices in dynamic networks compared to the original methods. Furthermore, by proposing a new algorithm called REvolutioN (Reversed Evolution Network) we address the challenging calculation of temporal centrality measures which depends on the number of vertices, the edge density and the number of snapshots. Due to our algorithms linear computational effort regarding the number of snapshots it makes both feasible, the calculation of the original methods over long snapshot sequences as well as the calculation of methods based on extra snapshots. Additionally, the algorithm benefits from sparse or very dense temporal networks and can be parallelized up to the number of vertices. To illustrate our method we use an age-related gene expression data set from the human brain, consisting of 1128 genes and 55 samples [3].

[1] Leibniz Institute for Prevention Research and Epidemiology - BIPS

hanke@leibniz-bips.de

References

1 J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia, in Proceedings of the 3rd Workshop on Social Network Systems, SNS '10 (ACM, New York, NY, USA,2010) pp. 3:1–3:6.

2 H. Kim and R. Anderson, Phys. Rev. E 85, 026107 (2012).

3 F. E. Faisal and T. Milenkovic, Bioinformatics 30, 1721 (2014).

Performance of ordinal-scaled prototype-based classifiers on microarray datasets

Andre Burkovski[1]*, Lyn-Rouven Schirra*[1]*, Ludwig Lausser*[1]*, Hans A. Kestler*[1]

Comparative microarray studies provide whole genome expression profiles for different phenotypes. These phenotypes, assuming a suitable correspondence between phenotype and gene expression, can be represented by a prototypical gene expression pattern or signature. Such a signature can then be determined via feature selection techniques and be further used to create suitable classification models. Here, prototype-based classification is of special interest as they allow a direct biological interpretation.

However, the measurements of gene expression levels can be noisy and the training on real-valued profiles might be misguided by potential outliers. In order to reduce these effects, we propose to operate on ordinal scaled signatures. These signatures are known to be invariant to a wide range of data transformations. Standard prototype-based classifiers can be adapted for processing the ordinal-scaled data in various ways. Both instance-based and centroid-based classifiers can rely on distances developed for rankings, i.e. ordinal data, and rank-aggregation procedures in order to compute centroids.

In this study we analyse and compare the performance of ordinal-scaled prototype-based classifiers against their real-valued counterparts. They are examined in experiments with different feature selection methods on a number of publicly available microarray datasets. We show that the performance of ordinal-scaled prototype-based classifiers in some cases can improve the classification performance and therefore should be incorporated in classification experiments with microarrays.

[1] Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm

Boosting Joint Models for Longitudinal and Time-to-Event Data

Elisabeth Waldmann[1], David Taylor-Robinson[2], Nadja Klein[3], Thomas Kneib[3], Matthias Schmid[4], Andreas Mayr[1,4]

Joint Models for longitudinal and time-to-event data have gained a lot of attention in the last few years as they are a helpful technique to approach a data structure very common in life sciences. The two outcomes are modeled by predictors which are composed of individual as well as shared sub predictors. The shared sub predictor is scaled by a so called association parameter which quantifies the relationship between the two parts of the model. Commonly Joint Models are estimated in likelihood based expectation maximization approaches or in a Bayesian framework[1]. The main drawbacks of the classical estimation procedures for joint model for modern biomedical settings is (i) the lack of a clear variable selection strategy and (ii) that they are unfeasible in high-dimensional data situations where the number of candidate variables p exceeds the number of observations n. In this work we propose a new inference scheme for joint model based on gradient boosting [2] that was particular designed to overcome these issues.

References

1 Faucett, C. and Thomas, D. (1996). Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: a Gibbs Sampling Approach. *Statistics in Medicine*, 15, pp 1663 – 1685.

2 Bühlmann, P. and Hothorn, T. (2006). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science*, 22, pp 477 – 522.

[1] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg

[2] Department of Public Health and Policy, University of Liverpool

[3] Chair of Statistics and Econometrics, Georg-August-Universität Göttingen

[4] Department of Medical Biometrics, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn

elisabeth.waldmann@fau.de

Classification Certainty and Reject Options in Learning Vector Quantization

Thomas Villmann[1], A. Bohnsack[1,2], and M. Kaden[1]*

Classification learning by means of prototype-based classifiers like learning vector quantization (LVQ, [1]) or support vector machines (SVM, [2]) is one of the most successful paradigms in machine learning. The prototypes in SVM are the support vectors, which are data vectors determining the class borders. The aim of LVQ is distribute a preselected number of prototypes such that the classes are represented. After training it realizes a nearest prototype classifier (NPC) based on (differentiable) distances or dissimilarities. An energy function based variant of LVQ was proposed in [3] denoted as generalized LVQ (GLVQ). The main idea is to use a classifier function

$$\mu\left(\mathbf{v}\right)=\frac{d\left(\mathbf{v},\mathbf{w}^{+}\left(\mathbf{v}\right)\right)-d\left(\mathbf{v},\mathbf{w}^{-}\left(\mathbf{v}\right)\right)}{d\left(\mathbf{v},\mathbf{w}^{+}\left(\mathbf{v}\right)\right)+d\left(\mathbf{v},\mathbf{w}^{-}\left(\mathbf{v}\right)\right)} \tag{1}$$

in this energy function, where $\mathbf{w}^{+}\left(\mathbf{v}\right)$ and $\mathbf{w}^{-}\left(\mathbf{v}\right)$ are the nearest prototypes with correct and incorrect class, respectively, regarding a presented data vector $\mathbf{v}$ depending on the dissimilarity measure d in use.[3] Hence, the classifier function $\mu\left(\mathbf{v}\right)\in\left[-1,1\right]$ becomes negative, iff $\mathbf{v}$ correctly classified. The respective energy function $E=\sum_{\mathbf{v}}\mu\left(\mathbf{v}\right)$ to be minimized approximates the classification error and is optimized by stochastic gradient descent learning [6]. However, as it is shown in [7], classification combined with data representation requires additional constraints to achieve class representative prototypes. These constraints can be realized in GLVQ by an additional penalty term based on the mean squared error [8].

One challenging problem in classification learning is the so-called classification certainty, i.e. the estimation of the evidence of a classification decision for an unknown data object. For SVM, a quantity related to secure classification decisions is the separation margin, which is maximized during the model learning. For GLVQ, the hypothesis margin

$$m_{h}\left(\mathbf{v}\right)=d\left(\mathbf{w}^{-}\left(\mathbf{v}\right),\mathbf{w}^{+}\left(\mathbf{v}\right)\right) \tag{2}$$

is optimized describing the robustness of the GLVQ regarding model shifts [9]. Yet, these quantities cannot be used to judge the decision certainty for unknown data. For NPC-classifiers the distance relation between best matching prototypes for each class can be used to estimate class assignment probabilities [10].

If classification decisions are related to costs, those classifiers come into play, which optimize Bayes decisions regarding optimum costs [11]. The performance of these classifiers can be further improved, reject options are incorporated [12], i.e. objects can be

[1] University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, DE

[2] FOS Kaufbeuren, Josef-Fischer-Str. 5, 87600 Kaufbeuren, DE

[3] For a dissimilarity measure d we require at least $d\left(\mathbf{v},\mathbf{v}\right)=0$ and $d\left(\mathbf{v},\mathbf{w}\right)\geq 0$ [4, 5]. For stochastic gradient learning in GLVQ we further suppose the differentiability in the second argument.

[*] *corresponding author*, email: thomas.villlmann@hs-mittweida.de

rejected to push the classification performance but paying additional reject cost. Alternatively, the model can be used to indicate insecure classification decisions. Yet, the methods require the precise determination of the class distributions, which might be difficult [13]. GLVQ provides a robust model approach to estimate class distributions and, hence, it may serve as an approximated Bayes classifier with reject option in the working phase [14]. Recently, a GLVQ variant was proposed, which takes into account this knowledge about reject options in recall already during learning, i.e. the prototype learning is influenced by the perspective reject options in the application phase [15]. In this approach, cost dependent (geometric) dissimilarity thresholds are considered indicating a reject decision. Analogously, a similar approach can be derived, if costs for outlier detection are introduced [16]. After model adaptation, insecure data at the class borders or outlier can be rejected or indicated as decisions with high uncertainty. Both concepts of reject option are explained in detail during the conference presentation.

Additionally to these models, we provide an alternative outlier detection strategy for GLVQ, which takes explicitly the hypothesis margin m_h from (2) into account. According to this idea, a prototype rejects a data vector $\mathbf{v}$ because of uncertainty, if the distance value $d\left(\mathbf{v}, \mathbf{w}\left(\mathbf{v}\right)\right)$ is greater than the hypothesis margin $m_h\left(\mathbf{v}\right)$. In this way, classification certainty knowledge determines the dissimilarity range for secure classification decisions. We denote this strategy as an exploration horizon based outlier reject option (EHBORO). Particulalrly, we show that the knowledge about this post-learning reject option can be integrated into GLVQ adaptation to obtain an respectively optimized model. As before, this EHBORO can also be used to indicate classification decisions with high uncertainty.

References

1 Teuvo Kohonen. Improved versions of Learning Vector Quantization. In *Proc. IJCNN-90, International Joint Conference on Neural Networks, San Diego*, volume I, pages 545–550, Piscataway, NJ, 1990. IEEE Service Center.

2 B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, 2002.

3 A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

4 E. Pekalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications.* World Scientific, 2006.

5 T. Villmann, M. Kaden, D. Nebel, and A. Bohnsack. Similarities, dissimilarities and types of inner products for data analysis in the context of machine learning - a mathematical characterization. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, and J.M. Zurada, editors, *Proceedings of the 15th International Conference Artificial Intelligence and Soft Computing - ICAISC, Zakopane*, volume 2 of *LNAI 9693*, pages 125–133, Berlin Heidelberg, 2016. Springer International Publishing, Switzerland.

6 M. Kaden, M. Riedel, W. Hermann, and T. Villmann. Border-sensitive learning in generalized learning vector quantization: an alternative to support vector machines. *Soft Computing*, 19(9):2423–2434, 2015.

7 K.L. Oehler and R.M. Gray. Combining image compressing and classification using vector quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):461–473, 1995.

8 B. Hammer, D. Nebel, M. Riedel, and T. Villmann. Generative versus discriminative prototype based classification. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of 10th International Workshop WSOM 2014, Mittweida*, volume 295 of *Advances in Intelligent Systems and Computing*, pages 123–132, Berlin, 2014. Springer.

9 K. Crammer, R. Gilad-Bachrach, A. Navot, and A.Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing (Proc. NIPS 2002)*, volume 15, pages 462–469, Cambridge, MA, 2003. MIT Press.

10 F.-M. Schleif, T. Villmann, M. Kostrzewa, B. Hammer, and A. Gammerman. Cancer informatics by prototype networks in mass spectrometry. *Artificial Intelligence in Medicine*, 45(2-3):215–228, 2009.

11 C.K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions in Information Theory*, 16(1):41–46, 1970.

12 C.K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6:247–254, 1957.

13 L. Fischer and T. Villmann. A probabilistic classifier model with adaptive rejection option. *Machine Learning Reports*, 10(MLR-01-2016):1–16, 2016. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_01_2016.pdf.

14 L. Fischer, B. Hammer, and H. Wersing. Efficient rejection strategies for prototype-based classification. *Neurocomputing*, 169:334–342, 2015.

15 T. Villmann, M. Kaden, A. Bohnsack, S. Saralajew, J.-M. Villmann, T. Drogies, and B. Hammer. Self-adjusting reject options in prototype based classification. In E. Merényi, M.J. Mendenhall, and P. O'Driscoll, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of 11th International Workshop WSOM 2016*, volume 428 of *Advances in Intelligent Systems and Computing*, pages 269–279, Berlin-Heidelberg, 2016. Springer.

16 T. Villmann, M. Kaden, D. Nebel, and M. Biehl. Learning vector quantization with adaptive cost-based outlier-rejection. In G. Azzopardi and N. Petkov, editors, *Proceedings of 16th International Conference on Computer Analysis of Images and Pattern, CAIP 2015, Valetta - Malta*, volume Part II of *LNCS 9257*, pages 772 – 782, Berlin-Heidelberg, 2015. Springer.

Multi-Objective Parameter Configuration of Machine Learning Algorithms using Model-Based Optimization

Daniel Horn[1],Bernd Bischl [2]

The performance of many classification algorithms heavily depends on the setting of their respective hyperparameters. Many different tuning approaches exist, from simple grid or random search approaches to evolutionary algorithms and sequential model-based optimizers. Often, these algorithm are used to optimize only a single performance criterion. However, in some practical situations a single criterion may not be sufficient to adequately characterize the behaviour of the machine learning method under consideration and the Pareto front of multiple criteria has to be considered. We propose to use multi-objective model-based optimization to efficiently approximate these Pareto fronts.

Furthermore, the parameter sets of many classifiers do not only consist of numeric, but also categorical and integer parameters. Moreover, the ultimate goal in parameter tuning is the automatic selection of the best classifier. Instead of tuning each classifier individually, one after the other, our method operates on the joint space of all considered machine learning algorithms. Therefore a control parameter that selects the currently active base classifier is introduced. All other parameters of the machine learning algorithms are made dependent / subordinate on this model choice parameter. Our model-based optimization approach can efficiently handle these hierarchical, mixed parameter spaces in a multi-objective setting.

Our optimization method is readily available as part of the mlrMBO R package on Github. We compare its performance against the TunePareto package and regular random search in a pure numerical setting of SVM parameter tuning and a hierarchical, mixed setting where we optimize over multiple model spaces at once.

References

1 Horn D., Wagner .T, Biermann D., Weihs C., Bischl B. (2015) Model-Based Multi-objective Optimization: Taxonomy, Multi-Point Proposal, Toolbox and Benchmark. In: Evolutionary Multi-Criterion Optimization, Lecture Notes in Computer Science, vol 9018, Springer International Publishing, pp 64–78

2 Müssel, C., Lausser, L., Maucher, M., Kestler, H.A.: Multi-objective parameter selection for classi ers. Journal of Statistical Software 46(i05) (2012)

3 mlrMBO - model-based optimization with mlr, see `https://github.com/mlr-org/mlrMBO`

[1] TU Dortmund, Computational Statistics, 44227 Dortmund, Germany
[2] LMU München, Computational Statistics, 80539 München, Germany

`daniel.horn@tu-dortmund.de, bernd.bischl@stat.uni-muenchen.de`

On the Hyperparameter Settings of Random Forest

Philipp Probst[1], Anne-Laure Boulesteix[1] and Bernd Bischl[2]

Due to their good predictive performance, simple applicability and flexibility, random forests are getting increasingly popular for building prediction rules. Unfortunately, not much knowledge is available about the ideal hyperparameter settings of random forests. Some important hyperparameters are the number of trees, the number of randomly drawn features at each split, the number of randomly drawn samples in each tree and the minimal number of samples in a node. Common modern strategies for tuning are grid search, random search iterated F-racing or bayesian optimization. This can be too complicated for users without expertise on random forests, computationally costly or even infeasible in case of too big datasets.

In our empirical study, we study the influence of a diverse range of hyperparameter settings of random forest algorithms / implementations of many different R packages on more than 200 different regression and classification problems from the OpenML platform. We use out-of-bag predictions and different performance measures for evaluation, and simple meta-learning to relate the performance results to data set characteristics.

Our results yield valuable insights into a) parameter sensitivity for different performance measures b) optimal default settings, to be applied without further tuning c) tuning starting points and ranges for less time-consuming model building.

[1] Department of Medical Informatics, Biometry and Epidemiology, LMU Munich, 81377 Munich, Germany

[2] Department of Statistics, LMU Munich, 80539 Munich, Germany

probst@ibe.med.uni-muenchen.de, bouleste@ibe.med.uni-muenchen.de,
bernd.bischl@stat.uni-muenchen.de

Feature selecting multi-class classification

Lyn-Rouven Schirra[1,2], Florian Schmid[1], Ludwig Lausser[1] and Hans A. Kestler[1]

Gene expression profiles are a valuable resource for gaining insight into molecular processes in cells and tissues. The analysis of these profiles gives the opportunity of understanding the development of certain diseases on a cellular level. Nevertheless, interpretation of this data is quite challenging due to their high dimensionality. For instance, although support vector machines achieve fairly good classification performances the resulting classification models are barely interpretable and obscuring possible insight in the molecular processes.

Feature selection methods have proven to be a useful approach handling the high dimensionality of gene expression data [1]. By selecting small subsets of highly informative features, they can provide starting points for new biological hypotheses and experiments. The elementary strategies are using purely data driven filters, selecting features on the basis of scores, which is calculated feature-wise.

Another demanding challenge appears with the introduction of finer grained diagnostic classes. While most of the classifiers and the feature selection methods are designed to be applied to binary classification tasks, we are increasingly facing scenarios with multiple classes. One basic approach is to divide the whole dataset into two-class subsets by choosing every possible pair of classes (One-against-One). Another strategy picks each class in turn as the positive class and subsume the remaining classes as the negative class (One-against-All).

In this study we examined the interaction between four data-driven feature selection methods and the aforementioned multi-class classification strategies and their impact on the accuracy of the classification as well as on the classwise sensitivities and specifities.

References

1 Saeys, Y., Iñza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)

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

[2] Institut of Number Theory and Probability Theory, Ulm University, D-89069 Ulm

lyn-rouven.schirra@uni-ulm.de, florian.schmid@uni-ulm.de, ludwig.lausser@uni-ulm.de, hans.kestler@uni-ulm.de

Anomaly Detection with Shapelet-Based Feature Learning for Time Series

Laura Beggel[1,2], Bernhard X. Kausler[1], Martin Schiegg[1] and Bernd Bischl[2]

The detection of anomalous behavior in temporal data is a very diverse area of research. The concrete applications have focused on detecting anomalous behavior within time series, but the classification of entire time series into normal and anomalous has not received much attention yet. Previous works addressing the latter task typically utilize standard features or require features hand-crafted by an expert with domain knowledge. In response, we propose a novel method based on one-class Support Vector Data Description [1] which jointly detects anomalous time series in a set of sequences and learns features that are highly discriminative for this task. Our algorithm minimizes a learning objective that considers 1.) the quality for anomaly detection of the data representation using shape-based features [2] and 2.) the compactness of the (normal) data under the given feature representation. This objective function is optimized using a block-coordinate descent procedure.
By jointly learning the features and a decision boundary for anomaly detection, our method is able to extract explicit characteristics for domain-specific normal behavior without the requirement of expert domain knowledge. These learned features are thus highly discriminative for the detection of anomalous observations in test data. We demonstrate the effectiveness of our approach on multiple data sets.

References

1 Tax, D. M., & Duin, R. P. (2004). Support vector data description. Machine learning, 54(1), 45-66.

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

[1] Robert Bosch GmbH, Zentralbereich Forschung und Vorausentwicklung, Robert-Bosch-Campus 1, 71272 Renningen
[2] Ludwig-Maximilians-Universität München, Institut für Statistik, Ludwigstraße 33, 80539 München

laura.beggel@bosch.com, bernd.bischl@stat.uni-muenchen.de

Implementing an EM algorithm for partially dependent data

Alexander Engelhardt [1]

The expectation-maximization algorithm (EM algorithm) is a popular method for finding maximum likelihood estimates in the presence of latent variables. We implemented an EM algorithm for a family study of colorectal cancer (CRC), where the data consisted of pedigrees of families, the dependent variable being the age at the onset of CRC. We assume a latent variable Z denoting the presence of an unknown genetic risk factor that increases the hazard ratio for the onset of CRC by a factor alpha. The unobservable proportion p1 of carriers of this risk factor as well as the risk increase alpha were then estimated with the EM algorithm. Genetic inheritance of this risk factor implies dependence within families, which led to computational difficulties for large families. We also show methods to minimize memory usage and computation time of this algorithm.

[1] IBE, LMU München

`engelhdt@ibe.med.uni-muenchen.de`

Netboost: Boosting supported network analysis for highdimensional genomic datasets

Pascal Schlosser[1,2]*, Michael Lübbert*[2] *and Martin Schumacher*[1]

Sequencing and array technologies develop rapidly, even exceeding development in computer science. Thereby, the computational challenge in analysing highdimensional genomic datasets is increasing and a need to identify networks within these large datasets is eminent. Weighted Gene Co-expression Network Analysis (WGCNA) is a versatile framework to do this[1]. But when combining multiple measurement types to interrogate their relations the data has many independent components and we reach computational limits.

We propose to implement a boosting based filtering step to improve the signal-to-noise ratio and obtain sustainable computational requirements. For each feature j likelihood based boosting is applied, using all other features as covariates. The inter-feature distances, defined by the topological overlap measure of convex transformations of the correlations, are then calculated for selected features (i, j). The sparse distances matrix is then hierarchically clustered by MC-UPGMA[2] and the resulting dendrogram separated into modules by DynamicTreeCut[3].

We apply Netboost on The Cancer Genome Atlas data from 192 acute myeloid leukemia patients. Together these methylation and expression arrays cover more than 532,000 features in the human genome. The subset of chromosome 18 features is small enough to apply WGCNA as well. Netboost results in smaller and fewer modules by extracting the most prominent effects. The analysis of the whole genome data comparing Netboost and a scalable adaptation of WGCNA is ongoing. The resulting modules will be benchmarked by their association to the clinical endpoints.

References

1 WGCNA: an R package for weighted correlation network analysis, Langfelder and Horvath, *BMC Bioinformatics*, 2008

2 Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space, Loewenstein et al., *Bioinformatics*, 2008

3 Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R, Langfelder et al., *Bioinformatics*, 2008

[1] Institute of Medical Biometry and Statistics, Medical Center - University of Freiburg, Freiburg, Germany

[2] Department of Hematology, Oncology and Stem Cell Transplantation, Medical Center - University of Freiburg, Freiburg, Germany

schlosser@imbi.uni-freiburg.de

A Statistical Model for the Analysis of Beta Values in DNA Methylation Studies

Leonie Weinhold[1], Simone Wahl[2], Per Hoffmann[3] and Matthias Schmid[1]

The analysis of DNA methylation is a key component in the development of personalized treatment approaches. A common way to measure DNA methylation is the calculation of beta values, which are bounded variables of the form $M/(M+U)$ that are generated by Illumina's 450k BeadChip array. The statistical analysis of beta values is considered to be challenging, as traditional methods for the analysis of bounded variables, such as M-value regression and beta regression, are based on regularity assumptions that are often too strong to adequately describe the distribution of beta values. We develop a statistical model for the analysis of beta values that is derived from a bivariate gamma distribution for the signal intensities M and U. By allowing for possible correlations between M and U, the proposed model explicitly takes into account the data-generating process underlying the calculation of beta values. Using simulated data and a real sample of DNA methylation data from the Heinz Nixdorf Recall cohort study, we demonstrate that the proposed model fits our data significantly better than beta regression and M-value regression. In addition, the proposed model improves the identification of associations between beta values and covariates such as clinical variables and lifestyle factors.

[1] Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Str. 25, D-53127 Bonn, Germany

[2] Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, D-85764 Neuherber, Germany

[3] Human Genomics Research Group, Department of Biomedicine, University Hospital Basel, Hebelstr. 20, CH-4031 Basel, Switzerland

weinhold@imbie.uni-bonn.de, simone.wahl@helmholtz-muenchen.de, per.hoffmann@unibas.ch, matthias.schmid@imbie.uni-bonn.de

Automated Design of Search Algorithms for Feature Set Ensembles

Gunnar Völkel[1], Ludwig Lausser[1], Hans A. Kestler[1]

Feature selection is one of the most important methods for generating interpretable decision rules for high-dimensional profiles. Selecting valuable biomarkers, these techniques point at potential hypotheses on the molecular background of a phenotype or disease. Nevertheless, in the context of low sample sizes these hypotheses will not be unique. Often, several suitable biomarker combinations and explanations exist.

In this work, we present an ensemble feature selection technique aiming at the parallel construction of sparsely overlapping marker combinations. The technique evolves a population of marker combinations using a genetic algorithm with diversity preserving methods. The marker combinations are rated by a correlation-based measure. From the final population, the subset consisting of the best marker combinations is aggregated to a multi-classifier system.

Metaheuristics like the genetic algorithm allow a large degree of customisation. Suitable operators or parameter values are typically not evident. In this situation tuning methods are preferable to manual selection. We utilize the irace tuning package [1] to find good parameter and operator choices for the genetic algorithm. To find a configuration for the genetic algorithm that achieves a good performance for the general marker selection problem, the tuning is carried out on multiple datasets. The tuning is parallelised on remote computation servers via the Sputnik library [2].

References

1 Manuel López-Ibáñez, Jérémie Dubois-Lacoste, Thomas Stützle, and Mauro Birattari. The irace package, Iterated Race for Automatic Algorithm Configuration. Technical Report TR/IRIDIA/2011-004, IRIDIA, Université Libre de Bruxelles, Belgium, 2011.

2 Völkel, G., Lausser, L., Schmid, F., Kraus, J., Kestler, H.: Sputnik: ad hoc distributed computation. Bioinformatics **31**(8), 1298–1301 (2015)

[1] Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm

{gunnar.voelkel,ludwig.lausser,hans.kestler}@uni-ulm.de

Ensemble Pruning for Glaucoma Detection.

Werner Adler[1], *Olaf Gefeller*[1], *Asma Gul*[2], *Folkert K. Horn*[3], *Zardad Khan*[4], *Berthold Lausen*[5]

Glaucoma is a neurodegenerative eye disease leading to blindness when not treated in time. The Heidelberg Retina Tomograph (HRT) is a non-invasive device producing topographical features of the eye-background suitable for glaucoma detection. A Random forest (RF) (Breiman, 2001) is an ensemble of classification trees that has shown to perform very well in detecting glaucoma based on HRT parameters (Adler et al., 2008). A typical RF consists of several hundred trees. To minimize the computational cost, several ensemble pruning techniques have been proposed that reduce the number of trees in the ensemble without loss of classification performance (Tsoumakas et al., 2009). We examine the performance of ensemble pruning based on the Double-Fault similarity and the classification performance of single trees in the field of glaucoma classification using a data set consisting of 309 observations of 102 topographical features. To validate our findings based on this data set, and to examine the influence of the prevalence of glaucoma in the data, we additionally perform a simulation study. Sensitivities, specificities, and the AUCs are reported and the performances of the examined pruning strategies are discussed.

References

Adler, W., Peters, A., Lausen, B. (2008): Comparison of classifiers applied to confocal scanning laser ophthalmoscopy data. Methods of Information in Medicine, 47, 38–46.

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

Tsoumakas, G., Partalas, I., Vlahavas, I. (2009): An ensemble pruning primer. Applications of supervised and unsupervised ensemble methods, 1–13.

[1] Department of Biometry and Epidemiology, University of Erlangen-Nuremberg, Germany
[2] Department of Statistics, Shaheed Benazir Bhutto Women University, Peshawar, Pakistan
[3] Department of Ophthalmology, University of Erlangen-Nuremberg, Germany
[4] Department of Statistics, Abdul Wali Khan University, Mardan, Pakistan
[5] Department of Mathematical Sciences, University of Essex, Colchester, UK

`Werner.Adler@fau.de`

Ensemble of selected classifiers

Berthold Lausen [1]

I discuss recent proposals to improve classification based on data with high dimensional feature space. For example after preprocessing microarray data with 500 000 probes and 22125 features (probesets) which represent genes, I use our proposal to improve feature selection of microarray data based on a proportional overlapping score[1]. Using benchmark data sets I compare random forests and our recent proposals of new classification methods based on ensembles of selected k-nearest neighbours and tree classifiers [2].

References

1 Mahmoud, O., Harrison, A.P., Perperoglou, A., Gul, A., Khan, Z., Metodiev, M., Lausen, B. (2014), A feature selection method for classification within functional genomics experiments based on the proportional overlapping score, BMC Bioinformatics, 15 (1).

2 Khan, Z., Gul, A., Mahmoud, O., Miftahuddin, M., Perperoglou, A., Adler, W., Lausen, B. (2016), An ensemble of optimal trees for class membership probability estimation. In: Wilhelm, A., Kestler, H. (eds.), Proceedings of ECDA2014, Springer, Heidelberg, in press.

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W., Lausen, B. (2016), Ensemble of a subset of kNN classifiers, Advances in Data Analysis and Classification, (online first).

[1] Department of Mathematical Sciences, University of Essex, UK

blausen@essex.ac.uk

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-$\mathrm{P}^c c$ 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

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm