ulm university universität
uulm

Statistical Computing 2015
Abstracts der 47. Arbeitstagung

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

Ulmer Informatik-Berichte

Nr. 2015-04
July 2015

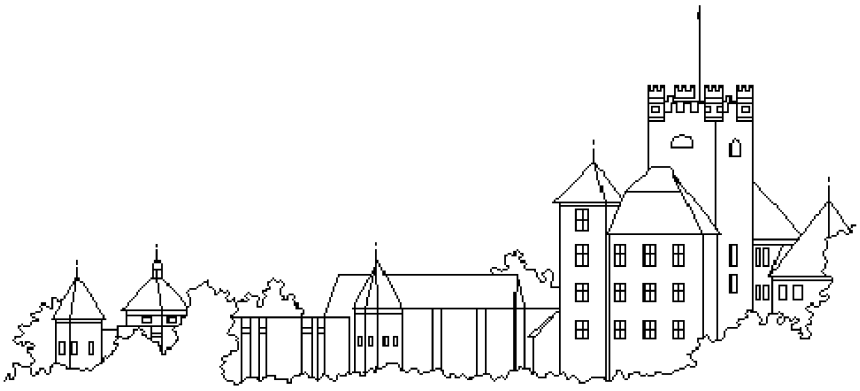
International Graduate School
in Molecular Medicine Ulm

SYSTAR

Leibniz Institute for Age Research
Fritz Lipmann Institute (FLI)

Friedrich-Schiller-Universität Jena
seit 1558

Statistical Computing 2015



47. Arbeitstagung

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

19.07. - 22.07.2015, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, July 19, 2015

18:00-20:00		**Dinner**
20:00-21:00		**Chair: H. A. Kestler**
20:00-21:00	Andreas Ziegler (Lübeck)	Random forests: Current concepts and implementations

Monday, July 20, 2015

09:15-09:30		**Opening of the workshop: H. A. Kestler, M. Schmid**
09:30-14:30		**Chair: M. Schmid**
09:30-09:50	Anne-Sophie Stöhlker (Freiburg)	Integrative Analysis of Case-Control Data on Multiple Cancer Subtypes
09:50-10:10	Elisabeth Waldmann (Erlangen)	Joint Modelling of Quantile Regression and Survival Time of Lung Function Decline in Cystic Fibrosis Patients
10:10-10:30	Silke Janitza (München)	Variable Importance Measures in Random Forests
10:30-11:00		**Coffee Break**
11:00-12:00	Holger Fröhlich (Bonn)	Probabilistic Modeling of biological high throughput data
12:00-13:30		**Lunch**
13:30-14:30	Tutorial: Marvin Wright (Lübeck)	Implementation of random forests in the R package Ranger
14:30-14:50		**Coffee Break**
14:50-21:00	Social Program (Ulm)	Nabada (Schwörmontag)

Tuesday, July 21, 2015

09:30-12:00		**Chair: L. Lausser**
09:30-09:50	Max Schneider (Potsdam)	Spatial Variation of Agricultural Abandonment with Spatically Boosted Models
09:50-10:10	Tobias Hepp (Erlangen)	Regularization methods in statistical modelling - A comparison between gradient boosting and lasso
10:10-10:30	Andreas Mayr (Erlangen)	Analysing measurement errors via boosting location and scale models
10:30-11:00		**Coffee Break**
11:00-12:00	Joaquin Vanschoren (Eindhoven)	OpenML: Networked science in machine Learning
12:00-13:30		**Lunch**
13:30-14:50		**Chair: E. Sträng**
13:30-13:50	Giuseppe Casalicchio (München)	The Residual-based Predictiveness Curve - A Visual Tool to Assess the Performance of Prediction Models
13:50-14:10	Alfred Ultsch (Marburg)	3-D printing as a tool for knowledge discovery in high dimensional data spaces
14:10-14:30	Rolf Hühne (Jena)	Lifespan Data Analysis - 3D Gene Network Visualization
14:30-14:50	Uwe Ligges (Dortmund)	Model based optimization of a statistical simulation model for single diamond grinding
14:50-15:30		**Poster-Teaser, Postersession + Coffee Break**
15:30-16:50		**Chair: H. Binder**
15:30-15:50	Andre Burkovski (Ulm)	Handling Missing Values in Rank Aggregation
15:50-16:10	Werner Adler (Erlangen)	Ensemble pruning to optimize the performance of ensembles of decision trees with unbalanced data sets
16:10-16:30	Dirk Surmann (Dortmund)	Predicting Measurements at Unobserved Locations in an Electrical Transmission System
16:30-16:50	Gunnar Völkel (Ulm)	Parallel Algorithms with Sputnik
16:50-17:50		Working group meeting on **Statistical Computing 2016** and other topics (all welcome)
18:00-20:00		**Dinner**
20:00-21:00	Tutorial: Bernd Bischl, Giuseppe Casalicchio (München)	OpenML with R and mlr

Wednesday, July 22, 2015

09:30-10:30		**Chair: U. Ligges**
09:30-09:50	Marius Felder (Jena)	Long-read sequencing for repetitive structure resolution
09:50-10:10	Philipp Koch (Jena)	Using a novel read-based method for the discovery and annotation of repetitive elements in the genome of the short-lived killifish Nothobranchius furzeri
10:10-10:30	Florian Schmid (Ulm)	Data transformations for invariant classifiers
10:30-11:00		**Coffee Break**
11:00-12:00		**Chair: A. Groß**
11:00-11:20	Lyn-Rouven Schirra (Ulm)	Evaluating feature selection methods for binarized Data
11:20-11:40	Alexander Engelhardt (München)	A statistical model for signal detection and bias correction in chip-seq data
11:40-12:00	Jakob Richter (Dortmund)	Multi Fidelity Model-Based Hyper-Parameter tuning for Large Scale Learning Problems
12:00-13:30		**Lunch**

Contents

Random forests: Current concepts and implementations

Andreas Ziegler[1,2,3], *Marvin N. Wright*[1]

Random Forests are fast, flexible and represent a robust approach to mining high-dimensional data. They are an extension of classification and regression trees (CART) and perform well even in the presence of a large number of features and a small number of observations. Random forests can deal with continuous outcome, categorical outcome and time-to-event outcome with censoring. The tree-building process of random forests implicitly allows for interaction between features and high correlation between features. Approaches are available to measuring variable importance and variable selection. Although random forests perform well in many applications, their theoretical properties are not fully understood. Recently, several articles have provided a better understanding of random forests, and we summarize these findings. We survey different versions of random forests, including random forests for classification, probability estimation and regression. Emphasis will be put on random forests for survival data. We discuss various approaches for generating forests. We briefly review backward variable elimination and forward variable selection, the determination of trees representing a forest and the identification of important variables in a random forest. We also provide a brief overview of different areas of application of random forests.
In a separate tutorial, we introduce the C++ application and R package Ranger. The software is a fast implementation of random forests and particularly suited for high dimensional data. Ensembles of classication, regression and survival trees are supported. We describe the handling of the software and compare runtime and memory usage with other implementations. The handling of several datasets is demonstrated, including a genome-wide association study.

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany
[2] Zentrum für Klinische Studien, Universität zu Lübeck, Germany
[3] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

ziegler@imbs.uni-luebeck.de, wright@imbs.uni-luebeck.de

Integrative Analysis of Case-Control Data on Multiple Cancer Subtypes

Anne-Sophie Stöhlker[1], Alexandra Nieters[2], Harald Binder[3] and Martin Schumacher[1]

In general, one cancer entity does not have a simple structure but usually breaks down into several (heterogeneous) subtypes. When investigating the associations between genes or, respectively, SNPs and one cancer, this diversity turns into a challenge for the analysis: Some genes might be associated with the respective cancer in general, while some other genes could be related to specific subtypes. However, subgroup analysis might overlook shared genes which determine central characteristics of the cancer whereas an overall analysis could miss type-specific genes representing the singularity of subtypes. Thus, an analysis combining those aspects would be beneficial to understand relations and differences of the subtypes. Data on several cancer subtypes is mostly investigated by comparing analysis results of single subtypes and therefore might suffer from an insufficient amount of data per subtype.
We consider an approach for integrative analysis that analyzes the data on all subtypes simultaneously [1]. While this approach was developed for prognosis data, we modify it for case-control settings. It is based on the heterogeneity model allowing a gene to be associated with all, only some, or none of the subtypes, respectively. Building upon this, a tailored compound penalization method is applied to actually find out whether a gene is associated with any of the present subtypes and if so, with which of them. In this context, genes are selected if they contain important SNPs associated with any subtype. The proposed method uses an iterative algorithm to identify interesting genes. To contrast the above-mentioned approach, we also investigate a componentwise boosting approach. Both procedures are applied to real genotyping data derived from a case-control study on the etiology of various subtypes of lymphoma.

References

1 Liu, J., Huang, J. et al. (2014), Integrative analysis of prognosis data on multiple cancer subtypes. Biometrics, 70:480–488.

[1] Center for Medical Biometry and Medical Informatics; University Medical Center Freiburg
[2] Center for Chronic Immunodeficiency; University Medical Center Freiburg
[3] Institute of Medical Biostatistics, Epidemiology and Informatics; University Medical Center Johannes Gutenberg University Mainz

stoehlker@imbi.uni-freiburg.de

Joint Modelling of Quantile Regression and Survival Time of Lung Function Decline in Cystic Fibrosis Patients

Elisabeth Waldmann[1] *, David Taylor-Robinson*[2]

Lung function decline in cystic fibrosis (CF) patients is faster when the patient is infected by pulmonary diseases [1]. The direction of causality is unclear: does fast lung function decline increase susceptibility towards infections or does the onset of infection accelerate the decline? When modelling repeated measurement and time to event data simultaneously, analysis is often based on combining mixed models with survival analysis. Some data sets however have a more complex structure than the one underlying normality assumption. In the case of modelling lung function of cystic fibrosis patients over a long period of time the impact of the covariates differs between quantiles of the dependent variable. This work aims at explaining the setup of a Bayesian joint quantile model and illustrates it using data from the United Kingdom CF registry.
We propose a Markov Chain Monte Carlo (MCMC) method to estimate parameters of survival models and quantile regression as well as an association parameter between the two structures simultaneously. This is rendered possible by using the asymmetric Laplace distribution (ALD) as an auxiliary distribution that helps modeling quantile regression in a Bayesian setup[2]. When using a location scale presentation of the ALD inference on quantile regression can be done similar as in MCMC approaches for Gaussian mean regression. The combination with the survival model is hence straight forward and only differs slightly from Bayesian joint modelling in mean regression as suggested by [3].

References

1 Qvist, T., Taylor-Robinson, D., Waldmann, E., Olesen, H., Rønne Hansen, C., Mathiesen, M.I., Høiby, N., Katzenstein, T. L. , Smyth, R. L. , Diggle P., Pressler T. (2015). Infection and lung function decline; comparing the harmful effects of cystic fibrosis pathogen. in preparation.

2 Waldmann, E., Kneib, T., Yue, Y., Lang, S. and Flexeder, C. (2013). Bayesian Semiparametric Additive Quantile Regression. Statistical Modelling, 13, 223–252.

3 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.

[1] Institut für Medizininformatik, Biometrie und Epidemiologie,Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Department of Public Health and Policy, University of Liverpool

elisabeth.waldmann@fau.de

Variable Importance Measures in Random Forests

Silke Janitza, Anne-Laure Boulesteix

Random forests are a commonly used tool for classification with high-dimensional data as well as for ranking candidate predictors based on the so-called variable importance measures. There are different importance measures for ranking predictors, the two most common measures are the Gini importance and the permutation importance. The latter has been found to be more reliable than the Gini importance. It is computed from the change in prediction accuracy if removing any association between the response and a predictor variable, with large changes indicating that the predictor variable is important. A drawback of those variable importance measures is that there is no natural cutoff which can be used to discriminate between important and non-important variables. Several approaches, for example approaches based on hypothesis testing, have been developed for addressing this problem. The existing testing approaches are permutation-based and require the repeated computation of forests. While for low-dimensional settings those permutation-based approaches might be computationally tractable, for high-dimensional settings typically including thousands of genes, computing time is enormous. In this article we propose a computationally fast heuristic procedure of a variable importance test which is appropriate for high-dimensional molecular data where many variables do not carry any information. The testing approach is based on a modified version of the permutation variable importance measure, which is inspired by cross-validation procedures. In our studies on complex high-dimensional binary classification settings this new approach controlled the type I error and had higher power at a substantially smaller computation time than the permutation-based approach of Altmann and colleagues.

Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, D-81377 Munich, Germany.

janitza@ibe.med.uni-muenchen.de

Probabilistic Modeling of biological high throughput data

Holger Fröhlich

Modern biological high-throughput techniques allow for measuring large sets of molecular features in a massively parallel manner. Such omics data can cover different biological aspects (e.g. gene expression, DNA methylation) of a biological system and are of high relevance in modern biomedicine. Probabilistic models can help to integrate heterogeneous omics data, to decipher parts of the underlying complex biological system and to derive experimentally verifiable predictions.
As a first example I will consider the question, how different omics and clinical data can be combined effectively in order to obtain a consensus clustering of patients. I propose a Bayesian approach using a Dirichlet Process Mixture (DPM) coupled with an Accelerated Failure Time (AFT) model for this purpose. The model specifically includes an automated feature selection for each omics data entity.
Molecular features represent parts of a complex biological system. Classically, ordinary differential equations (ODEs) have been used to describe such systems mechanistically. However, a major difficulty lies in the fact that any biological sub-system is embedded into and influenced by the surrounding system. I will present a probabilistic approach, which aims to detect these latent influences based on time series data. Moreover, I will demonstrate that the method can be also used to estimate missing and wrong reactions in ODE systems. My last example focuses on reverse engineering of biological networks from perturbation data. Besides molecular data there is an increasing interest in imaging based techniques for this purpose. I will present a probabilistic graphical model, which can be used to learn the structure of a biological network from such data.

Institute for Computer Science, University of Bonn

frohlich@bit.uni-bonn.de

Implementation of random forests in the R package Ranger

Marvin Wright

Random forests are widely used in applications, such as gene expression analysis, credit scoring, image processing or genome-wide association studies. With currently available software, the analysis of high dimensional data is time-consuming or even impossible for very large datasets. We therefore introduce ranger, a fast implementation of random forests, which is particularly suited for high dimensional data. Ranger is available as standalone C++ application and R package. It is platform independent and designed in a modular fashion. Due to efficient memory management, datasets on genome-wide scale can be handled on a standard personal computer. We briefly describe the implementation and compare runtime and memory usage with other implementations.
In the hands-on tutorial, we illustrate the usage of the R version of ranger. Based on examples, we use random forests for classification, survival analysis and probability prediction. Furthermore, we investigate different methods to handle missing data and analyze a small genome wide association study.

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany

wright@imbs.uni-luebeck.de

Spatial Variation of Drivers of Agricultural Abandonment with Spatially Boosted Models

Max Schneider[1,2], Gilles Blanchard[1], Christian Levers[2], Tobias Kümmerle[2]

Agricultural abandonment (AA) is a significant land use process in the European Union (EU) and modeling its driving factors has great scientific and policy interest. Past studies of drivers of AA in Europe have been limited by their restricted geographic regions and their use of traditional statistical methods, failing to consider the spatial variation in both predictors and AA itself. In this study, we broaden our experimental region to cover the EU, which is the level of many agricultural policies governing AA, thus covering far more data as well as diverse landscapes. We implement a modeling framework based on boosted regression, choosing the squared loss function and P-splines as base learners, as well as cross validation for the early stopping criterion, for their superior statistical properties. By building models containing both constant and spatially-varying coefficients, as well as modelizing the spatial variation of AA, we assess the importance of the spatial relationship between the drivers and AA. Models are built to classify agricultural land as abandoned or not. The potential drivers come from a set of a dozen landscape, environment and socio-economic variables hypothesized to influence AA, which is derived using fallow-active classifications of agricultural land from satellite images over the last 12 years. Models are then evaluated by sensitivity, specificity and empirical risk calculated over training data; better performance of the models containing spatially-varying regressors over spatially-constant models indicates that the spatial variation of top predictors must be considered in process understanding of AA. Our analysis uncovers marked spatial variability in the behavior of top drivers of AA over the spatial domain; in fact, models with spatially-varying predictors outperform their spatially-constant counterparts. Our models thus help capture the nuanced spatial relationships leading to observed rates of EU farmland abandonment and offer insights for better agricultural policy.

[1] Institut für Mathematik, Universität Potsdam
[2] Institut für Geographie, Humboldt Universität zu Berlin

maxsch03@uni-potsdam.de

Regularization methods in statistical modelling - A comparison between gradient boosting and the lasso

Tobias Hepp, Andreas Mayr

In many fields of study, classical statistical models quickly reach their limits facing the constantly increasing data volume. Especially in high-dimensional data settings with more predictors than observations, regularization methods for variable selection and penalization thus enjoy great popularity. The "lasso" (least absolute shrinkage and selection operator) by R. Tibshirani probably is the most famous application for regularized regression, outperforming many other approaches with regard to prediction accuracy and interpretability.
Although originally motivated from a different perspective, the stagewise model building of gradient boosting exhibits a striking similarity to the coefficient paths of the lasso. While in the linear setting Efron et al. defined the condition for both profiles to be identical, simulation in Hastie et al. suggests that in some situations forward stagewise algorithms like gradient boosting might be preferable.
In this talk, we compare the performance of both approaches in simulation studies regarding Gaussian linear regression, logistic regression and Cox proportional hazard models for time-to-event data applying the R add-on packages mboost, glmnet and penalized.

References

1 Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. Statistical Science. 22(4), 477-505.

2 Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. The Annals of Statistics. 32(2), 407-499.

3 Hastie, T., Taylor, J., Tibshirani, R. and Walther, G. (2007). Forward stagewise regression and the monotone lasso. Electric Journal of Statistics. 1, 1-29.

4 Tibshirani, R. and Walther, G. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 58, 267-288.

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

tobias.hepp@uk-erlangen.de

Analysing measurement errors via boosting location and scale models

Andreas Mayr[1], Matthias Schmid[2], Annette Pfahlberg[1], Wolfgang Uter[1] and Olaf Gefeller[1]

When medical data or some characteristics are obtained using a medico-technical device, users expectation is typically that the measurements are precise and reflect the true status of what should be measured correctly. Measurement errors are, however, inevitable in practical situations. The analysis of measurements errors focuses on two different aspects, systematic bias and random error.
We propose a new method to address both simultaneously via generalized additive models for location, scale and shape (GAMLSS, [1]) in combination with permutation tests. More precisely, we extend a recently proposed boosting algorithm for GAMLSS [2] to provide a test procedure to analyse potential device effects on the measurements. We carried out a large-scale simulation study to provide empirical evidence that our method is able to identify possible sources of systematic bias as well as random error under different conditions. Finally, we apply our approach to compare measurements of skin pigmentation from two different devices in an epidemiological study [3].

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 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 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. Statistical Methods in Medical Research. doi: 10.1177/0962280215581855.

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

andreas.mayr@fau.de

OpenML: Networked science in machine learning

Joaquin Vanschoren

Today, the ubiquity of the internet is allowing new, more scalable forms of scientific collaboration. Networked science uses online tools to share and organize data on a global scale so that scientists are able to build directly on each other's data and techniques, reuse them in unforeseen ways, and mine all data to search for patterns.
OpenML.org is a place where researchers can easily share and reuse machine learning data sets, tools and experiments. It helps researchers win time by automating machine learning experiments as much as possible, and gain more credit for their work by making it more visible and easily reusable.
Moreover, OpenML helps scientists and students to explore different machine learning techniques, find out which are most useful in their work, and collaborate with others to analyze scientific data online.

Department of Mathematics and Computer Science, Eindhoven University of Technology

j.vanschoren@tue.nl

The Residual-based Predictiveness Curve - A Visual Tool to Assess the Performance of Prediction Models

Giuseppe Casalicchio

It is agreed among biostatisticians that prediction models for binary outcomes should satisfy two essential criteria: First, a prediction model should have a high discriminatory power, implying that it is able to clearly separate cases from controls. Second, the model should be well calibrated, meaning that the predicted risks should closely agree with the relative frequencies observed in the data.
The focus of this work is on the predictiveness curve, which has been proposed by Huang *et al.* (Biometrics 63, 2007) as a graphical tool to assess the aforementioned criteria. By conducting a detailed analysis of its properties, we review the role of the predictiveness curve in the performance assessment of biomedical prediction models. In particular, we demonstrate that marker comparisons should not be based solely on the predictiveness curve, as it is not possible to consistently visualize the added predictive value of a new marker by comparing the predictiveness curves obtained from competing models. Based on our analysis, we propose the "residual-based predictiveness curve" (RBP curve), which addresses the aforementionened issue and which extends the original method to settings where the evaluation of a prediction model on independent test data is of particular interest. Similar to the predictiveness curve, the RBP curve reflects both the calibration and the discriminatory power of a prediction model. In addition, the curve can be conveniently used to conduct valid performance checks and marker comparisons.

Institut für Statistik, Ludwig-Maximilians-Universität München

giuseppe.casalicchio@stat.uni-muenchen.de

3-D printing as a tool for knowledge discovery in high dimensional data spaces

Alfred Ultsch[1], Michael Weingart[2], and Jörn Lötsch[3,4]

Three-dimensional printing is a currently quickly evolving technique, apparently consisting of a technical change from spraying toner on paper to adding up layers of materials to a 3-D object, however, by enabling a machine to produce objects of any shape it has the potential to impact on virtually most areas [1]. This includes biomedical applications where 3-D printing receives increasing scientific interest reflected in a quickly raising number of publications. A PubMed search for "(("3D printing" OR "3-D printing" OR "Three dimensional printing" OR "Three-dimensional printing")) NOT review[Publication Type]" on June 14, 2015, produced 659 hits, of which 656 originated from the year 2000 and later. Main biomedical applications were so far 3-D printing in vascular implants, aerosol delivery technologies, cellular transplantation, endoprosthetics, tissue engineering, biomedical device development and pharmacology including techniques such as individualized drug delivery formulations [2].

3-D printing is also employed for the visualization of biomedical data, for example to produce graspable three-dimensional objects for surgical planning [3]. The present work proposes the application of 3-D printing to the enhancement of knowledge discovery in high-dimensional data transferring them in to 3-D haptic physical models. This addresses a most important step in the mining of high dimensional data consisting of the visualization of structures, which serves as a basis for the identification of cluster structures. The addition of the third dimension ameliorates the principal impossibility to represent a high dimensional space in a lower space without losing distance relations. For example, popular projection methods into a 2 dimensional plane, such as PCA, ICA, or MDS, occasionally produce severe distortions of the input space, for example distant points in the input space can be projected close together in in the output space, or vice versa.

An inherently 3-D approach to high-dimensional data is pursued in the U-matrix [4] that has been originally developed for visualization of data structure following projection on emergent self-organizing maps (ESOM). This technique is able to represent the original distances of the input space on top of the projected points by generating a 3-D landscape analogue. This has been applied to visualize the cluster structure of high-dimensional complex pain measures in human volunteers that allowed successful association of the genetic [5] or the psychological [6] background causatively underlying these clusters. When colored according to the convention of physical maps this landscape gains the haptic

[1]DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany
[2]Weingart Ingenieur-Büro + CNC Fräsen, Kirchheim-Teck, Germany
[3]Goethe - University, Institute of Clinical Pharmacology, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany
[4]Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Project Group Translational Medicine and Pharmacology TMP, Theodor-Stern-Kai 7, 60596 Frankfurt am Main

ultsch@informatik.uni-marburg.de

properties that allow researchers to literally grasp the structure of the data. This 3-D print of an U-matrix, produced as an output of a public domain ESOM program [7], is to our knowledge the first to apply 3-D printing techniques directly for data mining and knowledge discovery in high-dimensional data in a haptic form. The recently devolved abstract U-matrix [8] extends the application of 3-D haptic representation of high-dimensional data to virtually all projection methods not limited to ESOM.

Funding
The work has been supported by the Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (LOEWE, JL), Schwerpunkt: Anwendungsorientierte Arzneimittelforschung. The funders had no role in method design, data selection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist.

References

1 D'Aveni, R.A.: 3-D Printing Will Change the World. Harward Buisness Review, (2013)

2 Pillay, V., Choonara, Y.E.: 3d printing in drug delivery formulation: you can dream it, design it and print it. How about patent it? Recent patents on drug delivery & formulation (2015)

3 Rengier, F., Mehndiratta, A., von Tengg-Kobligk, H., Zechmann, C.M., Unterhinninghofen, R., Kauczor, H.-U., Giesel, F.L.: 3D printing based on imaging data: review of medical applications. Int J Comput Assist Radiol Surg 5, 335-341 (2010)

4 Ultsch, A.: The U-Matrix as Visualization for Projections of high-dimensional data. In: Proc. 11th IFCS Biennial Conference. (2003)

5 Lötsch, J., Ultsch, A.: A machine-learned knowledge discovery method for associating complex phenotypes with complex genotypes. Application to pain. Journal of biomedical informatics 46, 921-928 (2013)

6 Dimova, V., Oertel, B.G., Kabakci, G., Zimmermann, M., Hermens, H., Lautenbacher, S., Ultsch, A., Lötsch, J.: A more pessimistic life-orientation is associated with experimental inducibility of neuropathy-like pain pattern in healthy subjects. J Pain (2015)

7 Ultsch, A., Moerchen, F.: ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46, (2005)

8 Lötsch, J., Ultsch, A.: Exploiting the structures of the U-matrix. In: Villmann, T., Schleif, F.-M., Kaden, M., Lange, M. (eds.) Advances in Intelligent Systems and Computing, vol. 295, pp. 248-257. Springer, Heidelberg (2014)

Lifespan Data Analysis – 3D Gene Network Visualization

Rolf Hühne, Hans A. Kestler

The JenAge ageing factor database AgeFactDB [1] contains a large amount of lifespan data in form of lifespan observations. They describe the effect of modifications like gene deletions or dietary restriction on the lifespan of a model organism. For a single gene there can be hundreds of observations made under different conditions, involving many other genes or other ageing factors. Interactive 3-dimensional visualization can assist in analyzing complex data like this in many ways, e.g.: compact presentation, activation of human visual pattern recognition, exploration by zooming in/out, rotating and filtering. Our goal is to build a visualization tool specialized for ageing-related data and to integrate it into AgeFactDB as an analysis tool. But it should also be customizable for any other kind of network data.

Networks can be build from AgeFactDB data in many different ways. For each ageing factor their could be build for example a network containing all lifespan observations where it is directly involved, including all directly involved other ageing factors. For a more comprehensive look this could be expanded recursively to include these networks for all other ageing factors within the smaller network.

Any of the networks could also be expanded by adding annotations like Gene Ontology (GO) [2,3] terms as nodes. It is planned to integrate an increasing number of such annotations into AgeFactDB. Existing network visualization tools like Biolayout Express3D [4,5] and Cytoscape [6,7] do not really fit our needs. Instead we decided to expand the open source chemical and biomolecular 3D structure viewer Jmol [8]. Its focus is on the integration into a web browser as Javascript application or Java applet. But it is also available as standalone Java program for batch processing. Due to a very powerful Javascript-like scripting language it can be expanded by user-defined functions.

In a first step its scripting language was used to add network-specific capabilities to Jmol. The next step will be to build a browser-based network-specific graphical user interface that can easily be adapted to different kinds of networks.

Bioinformatics and Systems Biology of Aging, Leibniz Institute for Age Research – Fritz Lipmann Institute

rhuehne@fli-leibniz.de

References

1 Hühne R, Thalheim T, Sühnel J. AgeFactDB – The JenAge Ageing Factor Database – Towards data integration in ageing research. Nucleic Acids Res. 2014; 42(Database issue): D892-6. Epub 2013 Nov 11.

2 Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res. 2001; 11(8):1425-33.

3 http://geneontology.org/

4 Goldovsky L., Cases I., Enright A.J., Ouzounis C.A. BioLayout(Java): versatile network visualisation of structural and functional relationships. Appl Bioinformatics. 2005;4(1):71-4.

5 http://www.biolayout.org

6 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 2003 Nov; 13(11):2498-504

7 http://www.cytoscape.org

8 Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

Model based optimization of a statistical simulation model for single diamond grinding

Swetlana Herbrandt[1], Uwe Ligges[1], Manuel Pinho Ferreira[2], Michael Kansteiner[3], Dirk Biermann[3], Wolfgang Tillmann[2], Claus Weihs[1]

With the presented model for single diamond grinding we simulate normal forces arising during a grinding process in cement. Assuming the diamond to have the shape of a pyramid, a very fast calculation of force and removed volume can be achieved. The basic approach is the simulation of the scratch track. Its triangle profile is determined by the shape of the diamond. The approximation of the scratch track is realized by stringing together polyhedral. Their sizes depend on the actual cutting depth and an error describing the material brittleness. Each scratch part can be subdivided into three simplices for a straightforward calculation of the removed volume. Since the scratched mineral subsoil is generally inhomogeneous, the forces at different positions of the workpiece are expected to be different. This heterogeneous nature is considered by sampling from a Gaussian random field.
To achieve realistic outcome the model parameters are adjusted applying model based optimization methods. A noisy Kriging model is chosen as surrogate to approximate the deviation between modelled and observed forces. This deviation is minimized and the results of the modelled forces and the actual forces from conducted experiments are rather similar.

[1] Fakultät Statistik, Technische Universität Dortmund
[2] Fakultät Maschinenbau, Lehrstuhl für Werkstofftechnologie, Technische Universität Dortmund
[3] Institut für Spanende Fertigung, Technische Universität Dortmund

herbrandt@statistik.tu-dortmund.de, ligges@statistik.tu-dortmund.de

Handling Missing Values in Rank Aggregation

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

Rank aggregation is the process of computing a consensus ranking with a minimal number of disagreements to a set of input rankings, partial rankings or pairwise comparisons. While common algorithms were developed for the aggregation of full rankings, an aggregation of partial rankings is an active research area and different problem-specific methods exist that are able to cope with missing values. Common aggregation methods rely on independent preprocessing strategies in the presence of missing information.
In this work, we study the interaction of rank aggregation methods and different preprocessing strategies for missing values. We utilize common generative ranking models, such as the simplified Thurstone, the Plackett-Luce, or the Mallows model in order to generate sets of rankings with a controlled variability and a controlled number of missing values. A suitable combination of a preprocessing strategy and a rank aggregation method should be able to reconstruct a consensus ranking that it close to the generated ground truth ranking of a chosen generative ranking model.
We compare our results and findings in simulation studies and data from the field of microarray analyses. In this context partial rankings or missing values occur in the comparison of gene expression signatures.

[1] Core Unit Medical Systems Biology and Institute of Neural Information Processing, Ulm University, D-89069 Ulm, Germany
[2] Leibniz Institute for Age Research - Fritz Lipmann Institute and FSU Jena, D-07745 Jena

andre.burkovski@uni-ulm.de, llausser@fli-leibniz.de, hkestler@fli-leibniz.de

Ensemble pruning to optimize the performance of ensembles of decision trees with unbalanced data sets.

Werner Adler[1], Asma Gul[2], Zardad Khan[2,3], Berthold Lausen[2]

A natural way of optimizing classification performance is to minimize the classification error. However, in many situations, classes in the data set are represented in an unbalanced way, i.e. there exist many observations of one class and few observations of the second class. This problem is common e.g. in medical screening situations when the prevalence of the disease is very small. We compare several strategies to optimize the performance of random forests [1] in this special situation, i.e. when a high specificity is the goal. The approaches we examine are based on ensemble pruning based on the ensemble performance, and ensemble pruning based on single classification tree performance. These pruning strategies are compared to the performance when the ensembles are constructed alternatively [2]: either by changing the criteria to construct a single classification tree, or by applying a different sampling strategy to construct the ensemble. The analyses are based on a simulation study and partial AUC [3] is calculated to report the performance of the ensemble.

References

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

2 Chen, C., Liaw, A., Breiman, L. (2004): Using random forest to learn imbalanced data. Technical report, Department of Statistics, University of Berkeley.

3 Ricamato, M.T., Molinara, M., Tortorella, F. (2011): Selection strategies for pAUC-based combination of dichotomizers. MCS11 Proceedings of the 10th international conference on multiple classifier systems, 177–186.

[1] Department of Biometry and Epidemiology, University of Erlangen-Nuremberg, Germany
[2] Department of Mathematical Sciences, University of Essex, Colchester, UK
[3] Department of Statistics, Abdul Wali Khan University, Mardan, Pakistan

werner.adler@fau.de

Predicting Measurements at Unobserved Locations in an Electrical Transmission System

Dirk Surmann, Uwe Ligges, Claus Weihs

The European electrical transmission system is working closely to its operational limits due to market integration, energy trading and the increasing feed-in by renewable energies. There- fore the system has become more vulnerable for disturbances in different areas, for example energy permanently oscillating with a low frequency. Analysing this Low Frequency Oscillation (LFO) requires measurements of voltage angle and magnitude at different locations in the transmission system. Due to the fact that the considered system consists of a large number of locations, our aim is to derive a subset of locations which contains sufficient information about the LFO. This subset is easier manageable than interrogating all locations. We derive a parameter set for the Low Frequency Oscillation based on differential equations which characterises every single measuring locations. We construct a feasible Kriging meta-model over the whole transmission system or, respectively, for a subset of locations with the thought in mind to predict the remaining locations.
To obtain a smooth spatial effect in which the coefficients of neighbouring regions are similar, we utilise a penalised least square estimator. The Kriging model works in a discrete spatial domain with the assumption of a known covariance matrix. However, we do not know the covariance matrix between the locations neither are we able to measure the electrical distance between all locations easily. Leveraging a reproducing kernel Hilbert space on the energy network, we devise a kernel function on the basis of the adjacency matrix. This matrix reflects the (weighted) neighbourhood structure of the transmission system. With the known covariance structure, we are able to predict the remaining locations which are not in the training subset. The talk will describe the methods and compare prediction errors for two different kernel functions on subsets that predict one or two locations in an energy transmission system.

Technische Universität Dortmund, Fakultät Stochastik, Vogelpothsweg 87, 44221 Dortmund, Germany

surmann@statistik.tu-dortmund.de, ligges@statistik.tu-dortmund.de,
weihs@statistik.tu-dortmund.de

Parallel Algorithms with Sputnik

Gunnar Völkel[1], Ludwig Lausser[2], Hans A. Kestler[1,2]

Sputnik is a framework for distributed parallel computations on the Java Virtual Machine. Its leightweight ad hoc setup and its tool support allow an easy and flexible usage. The distributed system of Sputnik offers the possibility to add or remove worker nodes dynamically. A built-in web user interface provides an overview of the worker node performance and the completion of computation jobs. Neither a permanent setup nor administrator privileges are required to use Sputnik.
We demonstrate the capabilities of Sputnik in the context of biomarker selection. We show an implementation of a parallel evolutionary algorithm utilizing a correlation-based heuristic. It will be applied for extracting marker signatures out of gene expression profiles.

References

Hall, M. A. (2000): Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In Proceedings ICML, pages 359–366. Morgan Kaufmann.

Gunnar Völkel, Ludwig Lausser, Florian Schmid, Johann M. Kraus, Hans A. Kestler (2015): Sputnik: ad hoc distributed computation. Bioinformatics, 31 (8).

[1]Core Unit Medical Systems Biology, Ulm University
[2]Leibniz Institute for Age Research-Fritz Lipmann Institute and FSU Jena

gunnar.voelkel@uni-ulm.de, llausser@fli-leibniz.de, hkestler@fli-leibniz.de

OpenML with R and mlr

Giuseppe Casalicchio

OpenML is an online machine learning platform where researchers can automatically log and share data, code, and experiments, and organize them online to work and collaborate more effectively. In this tutorial, we present an R package to interface with the OpenML platform, and illustrate its use in combination with the mlr machine learning package. We show how the OpenML package allows R users to easily search, download and/or upload machine learning datasets, as well as R scripts solving machine learning tasks. It also allows them to automatically log their results online, share them with others, and download results from other researchers to build on them. Beyond ensuring reproducibility of results, it automates much of the drudge work, speeds up research, facilitates collaboration, and increases user's visibility online.

Institut für Statistik, Ludwig-Maximilians-Universität München

giuseppe.casalicchio@stat.uni-muenchen.de

Long-read sequencing for repetitive structure resolution

Marius Felder, Marco Groth, Philipp Koch, Matthias Platzer, Hans A. Kestler

Obtaining high-quality sequences of complex regions remains one of the major challenges of finishing genome assemblies. Typically this required time-consuming and expensive Sanger-sequencing of large-insert clones. The use of 2nd generation sequencing has complicated the assembly of repetitive sequences. Although much more sequences can be generated, the short sequence read data leads to more gaps, missing data, and thus more incomplete reference assemblies.
We explored the possibility to resolve repetitive regions using long-read Single-Molecule Real-Time (SMRT) sequencing technology from Pacific Biosciences.
This "third-generation" sequencing technology provides kilobase-sized reads using zero-mode waveguides to observe the base incorporation of an anchored polymerase (Eid et al., Science 2009). As testset for the PacBio approach BAC clones of the short-lived fish Nothobranchius furzeri were chosen.
Our results suggest that the PacBio technology allows resolving complex repetitive regions and could be used to significantly improve sequence assembly quality.

Leibniz Institute for Age Research - Fritz Lipmann Institute and FSU Jena, D-07745 Jena

mfelder@fli-leibniz.de, mgroth@fli-leibniz.de, philippk@fli-leibniz.de,
mplatzer@fli-leibniz.de, hkestler@fli-leibniz.de

Using a novel NGS read-based method for the discovery and annotation of repetitive elements in the genome of the short-lived killifish Nothobranchius furzeri

Philipp Koch[1], Andreas Petzold[1], Bryan R Downie[1], Domitille Chalopin[2], Jean-Nicolas Volff[2] and Matthias Platzer[1]

Many denovo assemblies of complex eukaryotic genomes sequenced by next generation sequencing (NGS) technologies face the challenge of a high repeat content. These NGS reads are too short to span many types of repeats which may comprise several kilobases. In order to properly analyze repeats (e.g. 50% of the human, 55% of the zebrafish and 80% of the barley genome, respectively), species-specific repeat libraries are required. We have recently reported a k-mer based method ("RepARK") [1] for the discovery and annotation of the fraction of genomic NGS data sets representing repetitive elements.
We analyzed the repeat content of the genome of the short-lived killifish Nothobranchius furzeri, a new model organism for aging research. Without an available genome assembly we found 64% to be repetitive including 21% of tandem repeats [2]. We then built a long-range, chromosomal scale genome assembly (N50: 57Mb, 1.24Gb) which was conducted in five tiers employing not only NGS technologies but also optical and genetic mapping as well as synteny comparison to other fish genomes [3]. Combining repeats identified by RepARK with those identified by reference based programmes produced a comprehensive library of repetitive elements (25,000 elements, 5.6Mb). Using this library, we identified 62.4% of the NGS reads and 35% of the genome assembly as repetitive. We also calculated the evolutionary history of individual repeat families which revealed a possible recent transposon activity in the N. furzeri genome. These potentially active DNA transposons (hAT & TcMar) and LINE retrotransposons (L2, REX-Babar & RTE) have to be further analyzed with regard to representation in transcriptome data and their potential role in development and aging.

References

1 Koch P, et al. RepARK - de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 2014, 42, e80

2 Reichwald K, et al. High tandem repeat content in the genome of the short-lived annual fish Nothobranchius furzeri: a new vertebrate model for aging research. Genome Biol 2009, 10:R16.

3 Reichwald K, Petzold A, Koch P, Downie BR, Hartmann N, et al. The genome of a short-lived vertebrate provides insights into the early stages of XY sex chromosome evolution and the genetic control of life-history traits. (submitted to Nat Genet)

[1] Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute, Jena, Germany
[2] Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, Lyon, France

philippk@fli-leibniz.de, mplatzer@fli-leibniz.de

Data transformations for invariant classifiers

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

Molecular diagnostic is based on the analysis of high-dimensional profiles of simultaneously measured biomarkers (e.g. gene expression profiles). All markers of a single profile can be assumed to be measured under nearly identical conditions. The variation among samples is not directly controlled. Glitches in the preparation of a sample are likely to globally affect a profile. Such samples might be misleading for the training of a classification model or might not be classified correctly afterwards.
In this work we analyze two data representations that neglect the effects of certain misleading data transformations. Both representations extract the ordinal structure of gene expression profiles and incorporate invariances in the training of decision rules. Any classifier trained on the corresponding profiles is guaranteed to be invariant against the function class of featurewise strictly monotonically increasing data transformations such as global scaling or global transition. The first transformation replaces the absolute values of a gene expression profile by a relative gene ranking [1]. The second one is based on pairwise feature comparisons, which leads to a mapping of the dataset to a binary feature space.

References

1 Lausser, L., Schmid, F., Schirra, L.R., Kestler, H.A.: Rank-based classifiers for extremely high-dimensional data. Advances in Data Analysis and Classification (2015). Under review

[1] Core Unit Medical Systems Biology and Institute of Neural Information Processing, Ulm University, D-89069 Ulm, Germany
[2] Leibniz Institute for Age Research - Fritz Lipmann Institute and FSU Jena, D-07745 Jena
[3] Institute of Number Theory and Probability Theory, Ulm University, D-89069 Ulm, Germany

llausser@fli-leibniz.de, florian-1.schmid@uni-ulm.de, lyn-rouven.schirra@uni-ulm.de, hkestler@fli-leibniz.de

Evaluating feature selection methods for binarized Data

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

Feature Selection has proven to be a powerful tool for handling classification tasks on high-dimensional biological or medical data such as gene expression profiles. It allows for constructing interpretable models of cellular mechanics by reducing the initial set of features to an observable signature of genes.
Another approach for obtaining interpretable models is to discretize continuous measurements to a simple two state switch. In the context of gene expression levels, these binarized states can be taken as the presence or absence of a mRNA-molecule.
Combining both approaches seems to be a promising for reducing the calculation complexity for the classifiers and for simultaneously creating easy-to-interpret models of the underlying diseases. However, not all feature selection methods, sufficent for real-valued data, are suitable for the application on binarized data.
In this work, we present an empirical study of the effects of combined application of binarization and suitable feature selection methods on the achieved performances of common classifiers. We investigate the selection and binarization stability of varying methods and compared the results to the perfomance on unaltered data and solely binarized data.

[1] Core Unit Medical Systems Biology and Institute of Neural Information Processing, Ulm University, D-89069 Ulm, Germany
[2] Leibniz Institute for Age Research - Fritz Lipmann Institute and FSU Jena, D-07745 Jena
[3] Institute of Number Theory and Probability Theory, Ulm University, D-89069 Ulm, Germany

lyn-rouven.schirra@uni-ulm.de,florian-1.schmid@uni-ulm.de, llausser@fli-leibniz.de, hkestler@fli-leibniz.de

A statistical model for signal detection and bias correction in chip-seq data

Alexander Engelhardt[1], *Georg Stricker*[2]

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) is a widely used method for studying interactions between proteins and DNA to better understand processes such as gene expression. ChIP-Seq data, however, has certain biases: Due to differences in factors like chromatin accessibility, mappability, and GC content, control experiments are performed that are used to correct for these biases. We developed a method based on Generalized Additive Models (GAMs), that smoothly models the coverage tracks as piecewise polynomials and is able to extract the signal component from the data. The framework allows for flexible inclusion of additional covariates such as GC content, and elegantly handles multiple proteins and replicates. Furthermore, using piecewise polynomials for the smoothed coverage tracks leads to a simple way to find peaks based on derivatives. Our method also yields corrected coverage tracks with confidence bands for further analyses. Benchmarks against existing methods on yeast data look promising. Furthermore the higher fraction of correct peaks out of all peaks (precision) suggests that correctly fit smooth functions are able to detect the summit of a peak more accurately than moving-average-based methods.

[1] Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany
[2] Gene Center Munich, Ludwig Maximilians University, Munich, Germany

alexander.engelhardt@ibe.med.uni-muenchen.de

Multi Fidelity Model-Based Hyper-Parameter Tuning for Large Scale Learning Problems

Jakob Richter[1] , Bernd Bischl[2], Michel Lang[1], Heike Trautmann[3] and Holger Hoos[4]

Introduction

State-of-the-art machine learning algorithms make us of so-called hyperparameters, whose settings are known to affect performance. Hence finding good values for these hyperparameters is an important problem that received much attention. Hyperparameter optimization becomes more challenging with bigger hyperparameter spaces and higher evaluation costs; the latter happens as training sets get larger and training a machine learning algorithm, such as a SVM, becomes more expensive. Sequential model-based optimization (SMBO) makes use of a regression model as a surrogate to predict the outcome for unknown hyperparameter settings and is an increasingly popular approach for hyperparameter tuning. The key idea behind our multi-fidelity model-based optimization (mfMBO) algorithm is to make SMBO faster on larger data sets, by sampling the training data and performing SMBO combined on the different sample sizes of the training data and thus decreasing average evaluation costs. Findings on so called lower fidelity levels will be propagated to improve the predictive accuracy of higher fidelity surrogate models. This approach is inspired by earlier work of Huang et al. [3] and generalizes their multi-fidelity Kriging approach; different from their algorithm, our mfMBO approach is not restricted to Gaussian process (aka Kriging) models.

Related Work

The superiority of automated and systematic optimization approaches for tuning hyperparameters over a manual approach is evident. Genetic algorithms such as CMA-ES [2] are common candidates but lack of good exploration qualities, making restarts necessary and leading to a fairly high amount of evaluations of the target function. More recent works on hyperparameter optimization and automatic algorithm configuration incorporate the idea to use a regression model for the response in dependency of the hyperparameter. Primarily featured in [5] the method of Model Based Optimization gained momentum and became popular with the SMBO Framework [4] and the SMAC Algorithm. Application of MBO to tune machine learning methods is found e.g. in Auto-WEKA [6] and in the R-Package mlrMBO [1].

[1] Faculty of Statistics, TU Dortmund, Germany
[2] Institute of Statistics, LMU Munich, Germany
[3] Information Systems and Statistics Group, University of Münster, Germany
[4] Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

richter@statistik.tu-dortmund.de, bernd.bischl@stat.uni-muenchen.de,
lang@statistik.tu-dortmund.de, trautmann@wi.uni-muenster.de, hoos@cs.ubc.ca

Conclusion
The presented mfMBO Algorithm is implemented in mlrMBO and can be used easily. It is not specifically designed for machine learning problems but can be used for deterministic and noisy optimization problems which can be adapted to a multi fidelity scenario. It has been benchmarked on empirical functions and in an SVM tuning scenario and was able to compete with the standard MBO approach. Especially for short run times and on large data sets mfMBO was able to outperform MBO. The algorithm worked reliably even if lower fidelity target functions were distorted and would have led to wrong conclusions if used solely.

References

1 Bernd Bischl, Jakob Bossek, Daniel Horn, and Michel Lang. mlrMBO: Model-Based Optimization for mlr. R package version 1.0. url: https://github.com/berndbischl/mlrMBO.

2 Nikolaus Hansen and Andreas Ostermeier. "Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation". In: Evolutionary Computation, 1996., Proceedings of IEEE International Conference on. IEEE. 1996, pp. 312–317.

3 D. Huang, T. Allen, W. Notz, and R. Miller. "Sequential kriging optimization using multiple-fidelity evaluations". In: Structural and Multidisciplinary Optimization 32.5 (2006), pp. 369–382.

4 Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. "Sequential model-based optimization for general algorithm configuration". In: Learning and Intelligent Optimization. Springer, 2011, pp. 507–523.

5 Donald R Jones, Matthias Schonlau, and William J Welch. "Efficient global optimization of expensive black-box functions". In: Journal of Global optimization 13.4 (1998), pp. 455–492.

6 Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms". In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2013, pp. 847–855.

Adding unlabeled samples to treatment sensitivity prediction in a breast cancer data set

Anton Moll

Introduction
In classification tasks of biological data, there are usually fewer labeled than unlabeled samples because labeling samples is costly or time-consuming. In addition, labeled data sets can be re-used in different contexts as additional unlabeled data sets. For example, when searching the Gene Expression Omnibus (GEO) repository for microarray data of drug sensitivity and resistance experiments, the largest one has 2,522 samples, but the median has only 12 samples. In machine learning in general, utilizing unlabeled data in classification tasks is called semi-supervised learning. One such algorithm uses Restricted Boltzmann Machines or Autoencoders (both are a type of artificial neuronal network), to pre-train on unlabeled data before fine-tuning via back-propagation with labeled data. It has gained attention since around 2000, since when it has been among the best-performing algorithms in visual object recognition. Here we compared support vector machines and artificial neuronal networks, both in supervised and semi-supervised mode. We measured accuracies in the task of classifying tissue taken from breast cancer patients at reductive surgery as chemotherapy-resistant or -sensitive.
Results
We constructed 6 different data sets by subsampling from GEO data set GSE25055 and GSE25065 with 40 labeled training samples from GSE25055, between 0 and 200 unlabeled training samples from GSE25055 and GSE25065, and 84 testing samples from GSE25065. This subsampling step was repeated 20 times. We always sampled the same number of sensitive and resistant cases. Using these data sets, we compared the learning methods support vector machine (SVM), Transductive SVM (TSVM), the neuronal network Deep Belief Network (DBN) with pre-training and 3 hidden layers, and neuronal network with one hidden layer without pre-training. Overall, TSVMs performed significantly better than the DBNs on all 6 data sets. TSVMs trained with 0 and 40 unlabeled samples performed significantly better than the TSVMs trained on 160 and 200 samples. There were no significant differences between DBNs utilizing different numbers of unlabeled samples. The best average classification accuracy was 67.8%, reached by TSVM without unlabelled cases. We also investigated the effect of different normalization procedures on the classification accuracy. Before classifying, the data were normalized with either RMA or MAS5, followed by either no batch-effect correction or Combat batch-effect correction. Only MAS5 profited from added Combat batch-effect correction, but normalization with RMA alone yielded the best classification accuracy.
Conclusion
Our results that adding more unlabeled samples to learning does not improve accuracy could be explained due to learning being hindered by the different sample sources coming from different distributions. Usually this is addressed by normalization and batch-effect

University Regensburg, Institute of Functional Genomics, Dep. for Statistical Bioinformatics

anton.moll@klinik.uni-regensburg.de

correction. However, in our data set, batch effect correction was only beneficial together with MAS5 as normalization procedure, and RMA alone was best. In our breast cancer data set we observed that adding unlabeled samples to learning by TSVM or DBN does not improve prediction of treatment sensitivity. Moreover, TSVM was the better classification method in both computing efficiency and classification accuracy.

Development of a Risk Score for n-smaller-p problems in Stratified Samples

Norbert Krautenbacher

The Objective of the project is the development of a risk score for childhood asthma based on genetic predictors (SNPs) as well as environmental predictors on 1708 children. Besides the n-smaller-p problem (the number of SNPs is 2.5 Million) we have to face the following issue: the sample is stratified which results from a two-phase sampling procedure. In a first phase samples were systematically taken from strata for "farm-exposure" categories, in a second phase each of these strata was divided again in strata for the outcome "asthma". In order to incorporate the sample bias an appropriate approach seemed to be the application of feature selection based on univariate survey regression models with inverse-probability weighting and adjustment of standard errors in order to perform LASSO regression with observation weights on the selected features. Further investigations contain purely multivariate approaches (like e.g. LASSO or Random Forest on all predictors). These approaches, however, have to be capable of handling the big amount of predictors and especially should incorporate the complex sampling design in an appropriate way.

Helmholtz Zentrum München, Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH)

krautenbacher@helmholtz-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

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm