ulm university universität ulm

Statistical Computing 2010

Abstracts der 42. Arbeitstagung

HA Kestler, H Binder, B Lausen

H-P Klenk, M Schmid, F Leisch (eds)

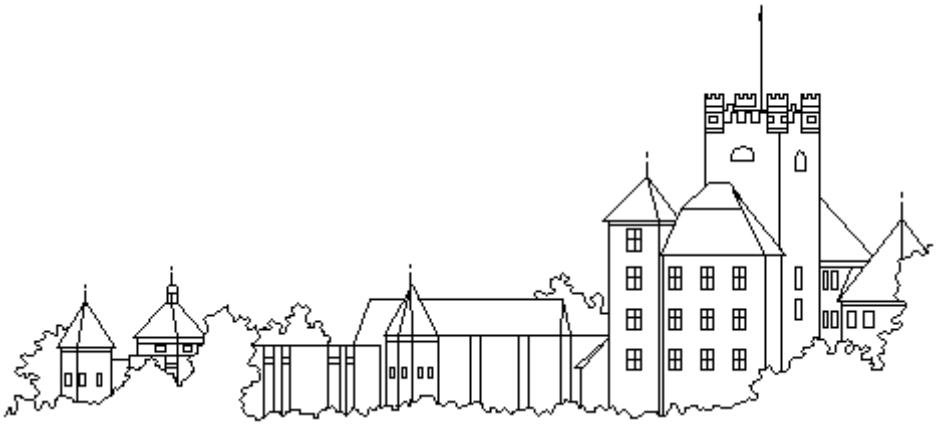
Ulmer Informatik-Berichte

Nr. 2010-05

Juni 2010

International Graduate School in Molecular Medicine Ulm

Statistical Computing 2010



42. Arbeitstagung

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

20.06.-23.06.2009, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, June 20, 2010

18:15-20:00		**Dinner**
20:00-21:00		**Chair: H.A. Kestler (Ulm)**
20:00-21:00	Michael Stadler (Basel)	Ultra-high Throughput Sequencing: Quantification by Counting

Monday, June 21, 2010

8:50		**Opening of the workshop: H.A. Kestler, H. Binder**
09:00-12:00		**Chair: A. Benner (Heidelberg)**
09:00-09:25	Veronika Reiser (Freiburg)	Matching, boosting or both? Identifying important genes in observational studies
09:25-09:50	Klaus Jung (Göttingen)	RepeatedHighDim: A new R-package for global tests of group effects in functional
09:50-10:15	Andreas Leha (Göttingen)	Practical Aspects of p-Value Adjustments at Interim Analyses in Microarray Studies
10:15-10:30		Discussion: Testing vs. risk prediction
10:30-11:00		**Coffee break**
11:00-11:30	Michael Glodek (Ulm)	Artificial neural networks trained by ensembles of Gaussian Mixture Models
12:15-14:00		**Lunch**
14:00-18:00		**Chair: F Leisch (München)**
14:00-14:30	Lea Vaas (Braunschweig)	Phenotype Microarray – New Technology, new realm of high dimensional data
14:30-15:00	Anna Telaar (Dummerstorf)	Biomarker Discovery: Classification using pooled samples
15:00-15:30	Sebastian Lück (Ulm)	Quantitative analysis of the intermediate filament network in scanning electron microscopy images by gray value-oriented segmentation methods
15:30-16:00	Sandra Andorf (Dummerstorf)	Analysis of heterosis in Arabidopsis thaliana: A molecular network structure based approach
16:00-16:30		**Coffee break**
16:30-18:00	Martin Hopfensitz, Christoph Müssel, Hans A. Kestler (Ulm), Marco Grzegorczyk (Dortmund), Tim Beißbarth (Göttingen), Holger Fröhlich (Bonn)	**Tutorial I**: “Network Modeling in Systems Biology with R”
18:15-20:00		**Dinner**
20:00-21:00		Tutorial II: “Network Modeling in Systems Biology with R

Tuesday, June 22, 2010

09:00-12:00		**Chair: B. Lausen (Essex)**
09:00-09:25	Ludwig Lausser (Ulm)	Assessing the robustness of feature selection techniques in resampling experiments
09:25-09:50	Werner Adler (Erlangen)	Evaluating the Predictive Power of Random Survival Forests for Breast Cancer
09:50-10:15	Matthias Schmid (Erlangen)	A Robust Alternative to the Schemper-Henderson Estimator of Prediction Error
10:15-10:30		Discussion: Model checking and prediction
10:30-11:00		**Coffee break**
11:00-11:30	Sebastian Krey (Dortmund)	Automatic sound recording segmentation for sound source and sound phase
11:30-12:00	Martin Schels (Ulm)	A Hybrid Information Fusion Approach to Discover Events in EEG Data
12:15-14:00		**Lunch**
14:00-18:00		**Chair: H. Binder (Freiburg)**
14:00-14:30	Nina Melzer (Dummerstorf)	A more realistic simulation approach for the prediction of genetic values from genome-wide Introduction marker data
14:30-15:00	Frank-Michael Schleif (Bielefeld)	Functional Vector Quantization based on Divergence Learning
15:00-15:30	Manuel Eugster (München)	Sequential Benchmarking
15:30-16:00	Bernd Bischl (Dortmund)	Benchmarking and Analysis of Local Classification Methods
16:00-16:30		**Coffee break**
16:30-18:00		Working groups meeting on **Statistical Computing 2011** and other topics (all welcome)
18:15-20:00		**Dinner**
20:00-21:00		**Software Demo** mlr and benchmark: Setup, Execution and Analysis of Benchmark Experiments in R

Wednesday, June 23, 2010

09:00-12:00		**Chair: M. Schmid (Erlangen)**
09:00-09:30	Nikolay Robinzonov (München)	Boosting for Estimating Spatially Structured Additive Models
09:30-10:00	Benjamin Hofner (Erlangen)	Constrained Regression Using mboost: An Application to Body Fat Prediction
10:00-10:30	Andreas Mayr (Erlangen)	Prediction Intervals and Quantile Boosting: A Simulation Study
10:30-11:00		**Coffee break**
11:00-11:30	Johann Kraus (Ulm)	Dimensionality reducing cluster analysis
11:30-12:00	Steffen Unkel (UK)	Zig-zag exploratory factor analysis with more variables than observations
12:15-14:00		**Lunch**

Ultra-high Throughput Sequencing: Quantification by Counting

Michael Stadler

Friedrich Miescher Institute,

Basel,

michael.stadler@fmi.ch

Ultra-high throughput sequencing has tremendously increased our ability to generate DNA sequence information over the last few years. Compared to classical capillary sequencers, the "next-generation sequencers" from companies such as 454-Life-Sciences and Solexa/Illumina produce data at a rate several orders of magnitude higher and at a drastically reduced cost. In the first part of the presentation, I introduce and compare the technologies behind next-generation sequencing.

Most next-generation sequencers produce millions of short sequence reads (around 50 nucleotides in length), which limits their value for sequencing and assembly of unknown repeat-rich genomes. However, they are ideal for quantifying DNA or RNA molecules by counting. In the second part, I present RNA-seq and ChIP-seq as two examples of counting applications, with an emphasis on the challenges in analyzing this new kind of data.

Matching, boosting or both? Identifying important genes in observational studies

Veronika Reiser, Christine Porzelius, Harald Binder
and Martin Schumacher

Freiburg Center for Data Analysis and Modeling,
University of Freiburg,
Institute of Medical Biometry and Medical Informatics,
University Medical Center Freiburg,
reiser@imbi.uni-freiburg.de

In observational studies with high-dimensional molecular data it is a typical aim to identify genes that provide information about the occurrence of a disease while adjusting for the effect of other factors. As one possible method for classification of subjects, we consider a matching approach (Heller et al., 2009), which is based on tests and selects important genes via p-values after constructing homogeneous groups for comparison. We contrast this approach with a regularized boosting technique (Tutz and Binder, 2007) that is directly based on prediction and produces sparse model fits, incorporating other factors as mandatory covariates. To give an insight into the characteristics of the two approaches, we analyze two gene expression microarray studies. The first study contains patients with B-cell and T-cell acute lymphoblastic leukemia and the second patients with acute megakaryoblastic leukemia. Additionally, we discuss potential combinations of both methods.

References

Heller, R., Manduchi, E. and Small, D. (2009). Matching methods for observational microarray studies. Bioinformatics, 25(7): 904-909.

Tutz, G. and Binder, H. (2007). Boosting ridge regression. Computational Statistics & Data Analysis, doi: 10.1016/j.csda.2006.11.041.

RepeatedHighDim: A new R-package for global tests of group effects in functional gene sets

Klaus Jung, Edgar Brunner and Tim Beißbarth

Department of Medical Statistics,
University Medicine Göttingen,
klaus.jung@ams.med.uni-goettingen.de

Sets of genes which are all associated with the same biological function (e.g. cell growth or cell death) are often in the focus of biomedical research. A particular question in such research is whether the genes of such a functional set are differentially expressed between the samples of different groups, for example between healthy and diseased individuals. Gene-specific expression levels are usually measured by DNA microarrays in such experiments. The typical way of comparing each gene individually between the groups has the disadvantage that a potential group effect might be ignored due to too small individual effects. If however many genes are altered just a little bit this might be detected as a group effect by performing one global test instead of many single tests. Several statistical approaches for global tests have been published, e.g. by Goeman (2004), by Hummel et al. (2008) and by Brunner (2009). In this talk, a new R-package – RepeatedHighDim – which implements the approach of Brunner (2009) is presented. This R-package provides functions for the case of two groups, where samples are either supposed to be all independent or are also allowed to be dependent between group. The latter case is for example of use when tumor and mucosa samples of the same individuals are compared. RepeatedHighDim has noticeable smaller calculation times than the respective R-packages of the two other approaches of global tests. In addition, the approach of Brunner exceeds the power of the two other approaches in certain situations.

References

Brunner, E. (2009) Repeated measures under non-sphericity. Proceedings of the 6th St. Petersburg Workshop on Simulation, 605-609.

Hummel, M., Meister, R. and Mansmann, U. (2008) GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics, 24, 78-85.

Goeman, J.J. et al. (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 20, 93-99.

Practical Aspects of p-Value Adjustments at Interim Analyses in Microarray Studies

Andreas Leha, Tim Beißbarth and Klaus Jung

Department of Medical Statistics,
University Medicine Göttingen,
andreas.leha@med.uni-goettingen.de

Microarrays and other high-throughput methods are widely used for simultaneously measuring expression levels of several hundreds or thousands of molecular features such as genes or proteins. Comparing each feature between the samples of different groups (for example healthy versus diseased, or therapy responder versus non-responder) by means of statistical testing makes adjustment of p-values necessary in order to avoid too many false positive results (Dudoit et al., 2003). In many microarray studies, samples become only available sequentially over longer time periods, particularly when survival times are studied or when low prevalent diseases are studied. In these cases, it is often desired to find interesting features which are correlated to the studied response at interim analyses prior to the end of the study. Interim analyses, however, make a second adjustment necessary due to the α-error increase caused by repeated testing. In a simulation study, we regard some practical aspects of the problem of adjusting twice, once for multiple testing and once for interim analyses using the methods proposed by Pocock (1977) or by O'Brien and Fleming (1979). In this simulation we study particularly the behavior of attained error rates (such as the false discovery rate) and attained average power during interim analyses. While Posch et al. (2009) found that adjustment for interim analyses is not necessary for microarray data that comprise a high number of features, our simulation yielded that adjustments for interim analyses are essential to maintain prespecified error rates when the number of features is rather low (< 500) such as in functional gene sets and when no gene is differentially expressed. We study further whether significant features detected in an interim analysis are to be tested again in subsequent interim analyses or whether they can be omitted. We verify our findings from the simulation study on a real data set of gene expression levels in tissue samples from leukemia patients (Chiaretti et al., 2004). We particularly compare sets of differentially expressed genes between two groups of patients detected by different ways of interim analysis. The example confirms the results of the simulation study.

References

Dudoit S., et al. (2003) Multiple hypothesis testing in microarray experiments. Statistical Science, 18, 71-103.

O'Brien P. and Fleming T.R (1979) A multiple testing procedure for clinical trials. Biometrics, 35, 549-556.

Posch M., et al. (2009) Hunting for significance with the false discovery rate. Journal of the American Statistical Association, 104, 832-840.

Pocock S.J. (1977) Group sequential methods in the design and analysis of clinical trials. Biometrika, 64, 191-199.

Chiaretti, S., X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival Blood, 103, 2771-2778 .

Artificial neural networks trained by ensembles of Gaussian Mixture Models

Michael Glodek and Friedhelm Schwenker

Institute of Neural Information Processing,
University of Ulm,
firstname.lastname@uni-ulm.de

Ensemble learning refers to machine learning algorithms employed to train multiple models and to combine their outputs by an appropriate fusion or combination method. The basic idea of ensemble learning is that decisions of the complete ensemble should have more accurate on average than any single ensemble member or at least be more robust against parameter settings. Outputs of ensemble members can be combined by averaging, voting, and many other statistical and machine learning methods. Ensembles have been used in many pattern recognition applications, such as regression or classification, but here we consider the problem of probability density function (PDF) estimation for a given data set $T = \{x^\mu : x^\mu \in R^n, \mu = 1, \ldots, M\}$ which is basically an unsupervised learning task. An algorithm is proposed that combines two ideas: (1) Model ensembling. and (2) supervised learning the PDF using an artificial neural network (ANN). In the first step of this algorithm an ensemble of Gaussian mixture models is computed for different parameter settings, e.g. number of components, subsets of T. For this the very well known expectation maximization algorithm has been applied. Then the fused ensemble output, e.g. through median or average fusion, is defined as a teaching signal for a supervised training procedure of an artificial neural network, in our study a multi-layer-perceptron (MLP) was used. One can presume that this approach will show its best performance for data of complex non-gaussian PDFs given only a few data samples. In order to enforce the ensemble of GMMs to have an increased diversity the prementioned subset of T and a random initialization of the GMMs was used. The testing of different types of ensemble combinations, revealed that the median offers the best performances due to its robustness against outliers in GMM overfitting. In order to train the MLP a set of trainings samples will be drawn within an interval determined by the mean and the variances of the ensemble members. The results of the numerical experiments show a superior approximation performance and robustness of the ensemble PDF and (the more compact) MLP approximator in comparison to the ensemble members.

Functional Vector Quantization based on Divergence Learning

T. Villmann, S. Haase, S. Simmuteit, M. Kästner
and F.-M. Schleif

University of Applied Sciences Mittweida,
Dep. of Mathematics, Natural and Computer Sciences,
Technical University Bielefeld,
CITEC,
thomas.villmann@hs-mittweida.de

Supervised and unsupervised vector quantization methods for prototype based classification and clustering traditionally use dissimilarities, frequently taken as Euclidean distances. Recent approaches in functional data processing investigates the applicability of divergences instead. In functional vector quantization the data as well as the prototypes represent functions. In this contribution we provide an extension of this focus. We suppose that the prototypes are convex combinations of independent basis functions. Thus learning of protypes is transferred to learning of the convex representation. We show, how this ideas can be incorporated in vector quantization learning based on divergences.

Model description Supervised and unsupervised vector quantization for prototype based classification and clustering supposes the evaluation of dissimilarities between data $\mathbf{v} \in \mathbb{R}^n$ and prototypes $W = \{\mathbf{w_r} \in \mathbb{R}^n\}_{\mathbf{r}\in A}$. Usually the dissimilarities are judged in terms of distances. In case of functional data, i.e. the several vector dimensions are not independent but can be taken as the domain of a function represented by the considered vector and the vector elements are the respective function values, advanced dissimilarity measures maybe more appropriate taking the functional structure of data into account[4]. Recent approaches focus on divergences instead of the usually applied Euclidean distance, if data are given as positive measures, i.e. non-negative functions or its discrete vectorial representation[1].

Online vector quantization frequently optimizes the reconstruction error as a cost

function, usually defined in terms of the used dissimilarity measure D:

$$E = \sum_{\mathbf{r} \in A} \int P(\mathbf{v}) \cdot h_{\mathbf{r},\mathbf{s}(\mathbf{v})} \cdot D(\mathbf{v}, \mathbf{w}_\mathbf{r}) \, d\mathbf{v} \tag{1}$$

with P being the data density, $h_{\mathbf{r},\mathbf{s}(\mathbf{v})} \leq 1$ is an weighting function describing the mapping and the 'spatial' relations between the prototypes and $\mathbf{s}(\mathbf{v}) = \arg\min_{\mathbf{r} \in A} (D(\mathbf{v}, \mathbf{w}_\mathbf{r}))$ is the best matching prototype according to the given dissimilarity measure. The online adaptation scheme is defined as a stochastic gradient descent on E: $\triangle \mathbf{w}_\mathbf{r} = -\varepsilon \frac{\partial E}{\partial \mathbf{w}_\mathbf{r}}$. In particular, the gradient $\frac{\partial E}{\partial \mathbf{w}_\mathbf{r}}$ explicitly depends from the derivative $\frac{\partial D}{\partial \mathbf{w}_\mathbf{r}}$ of the underlying dissimilarity measure, i.e. $\frac{\partial E}{\partial \mathbf{w}_\mathbf{r}} = F\left(\frac{\partial D}{\partial \mathbf{w}_\mathbf{r}}\right)$. In case of divergences, the derivative $\frac{\partial D}{\partial \mathbf{w}_\mathbf{r}}$ becomes the Fréchet-derivative $\frac{\delta D}{\delta w_\mathbf{r}}$ [6].

Powerful such vector quantization models are self-organizing map in the variant of T. Heskes (SOM) and neural gas (NG) for unsupervised learning or generalized learning vector quantization for supervised classification problems. One crucial property which all approaches have in common is the curse of dimensionality for high-dimensional data ($n \gg 0$) as occur for functional data. Hence, efficient and robust variants of these algorithms are needed for these data.

Here we propose the application of convex representations of the prototypes based on basis functions to avoid these problems. More specifically, we assume that the data prototypes are to be determined as $\mathbf{w}_\mathbf{r} = \sum_{k=1}^{m} \alpha_{\mathbf{r},k} \cdot \mathbf{b}_k$ with $\mathbf{b}_k$ represent basis functions and $m \ll n$. This representation can be written in functional form as $w_\mathbf{r}(t) = \sum_{k=1}^{m} \alpha_{\mathbf{r},k} \cdot b_k(t)$. The coefficients $\alpha_{\mathbf{r},k}$ are non-negative. Hence, learning reduces to learning of the coefficient vectors $\alpha_\mathbf{r}$ in that case. Yet, the gradient has to be adapted accordingly:

$$\frac{\partial E}{\partial \alpha_{\mathbf{r},k}} = \sum_{k=1}^{m} \frac{\partial E}{\partial \mathbf{w}_\mathbf{r}} \cdot \frac{\partial \mathbf{w}_\mathbf{r}}{\partial \alpha_{\mathbf{r},k}} = \sum_{k=1}^{m} \left\langle \frac{\partial E}{\partial \mathbf{w}_\mathbf{r}}, \mathbf{b}_k \right\rangle \tag{2}$$

with $\left\langle \frac{\partial E}{\partial \mathbf{w}_\mathbf{r}}, \mathbf{b}_k \right\rangle$ being the inner product of the respective vectors. Equivalently, we can write this in functional form as

$$\frac{\partial E}{\partial \alpha_{\mathbf{r},k}} = \sum_{k=1}^{m} \frac{\delta E}{\delta w_\mathbf{r}} \cdot \frac{\partial w_\mathbf{r}}{\partial \alpha_{\mathbf{r},k}} = \sum_{k=1}^{m} \left\langle \frac{\delta E}{\delta w_\mathbf{r}}, b_k \right\rangle \tag{3}$$

whereby $\left\langle \frac{\delta E}{\delta w_\mathbf{r}}, b_k \right\rangle$ is now the inner product in the Hilbert-space of quadratic integrable functions $\mathcal{L}_2$. The term $\frac{\delta E}{\delta w_\mathbf{r}}$ is the functional derivative also be known as *Fréchet-derivative* [2]. Hence, the derivative $\frac{\partial E}{\partial \alpha_{\mathbf{r},k}}$ again essentially depends on the underlying dissimilarity measure D

$$\frac{\partial E}{\partial \alpha_{\mathbf{r},k}} = \sum_{k=1}^{m} \left\langle \frac{\delta E}{\delta w_\mathbf{r}}, b_k \right\rangle = \sum_{k=1}^{m} \left\langle F\left(\frac{\delta D}{\delta w_\mathbf{r}}\right), b_k \right\rangle \ . \tag{4}$$

Emphasizing the aspect of functional data processing one possibility is to use divergences as dissimilarity measure between data and prototypes as mentioned above. In

fact, a large set of different divergences D exists. Thereby, the set of divergences can be partitioned into more or less disjunct subsets regarding their mathematical properties and (mathematical) topological categorization [1]. These include *Bregman*-divergences, *Czisars-f*-divergences as well as γ-divergences. All these classes share the *Kullback-Leibler*-divergence as the most prominent example. Recently, the Fréchet-derivatives $\frac{\delta D(v(t)||w_{\mathbf{r}}(t))}{\delta w_{\mathbf{r}}}$ for these divergence classes were proposed [5], the discretization of which immediately give the gradients needed for vector quantization algorithms. Thus, the divergence measures can plugged directly into the above outlined functional approach:

$$\frac{\partial E}{\partial \alpha_{\mathbf{r},k}} \sim -\left\langle \frac{\delta D\left(v\left(t\right)||w_{\mathbf{r}}\left(t\right)\right)}{\delta w_{\mathbf{r}}}, b_k \right\rangle \cdot h_{\mathbf{r},\mathbf{s}(\mathbf{v})}$$

For example, considering the Fréchet-derivative of the Kullback-Leibler-divergence $\frac{\delta D(v(t)||w_{\mathbf{r}}(t))}{\delta w_{\mathbf{r}}} = \frac{v(t)}{w(t)}$ gives for the SOM-update in case of functional representation $\frac{\partial E}{\partial \alpha_{\mathbf{r},k}} \sim -\left\langle \mathbf{d}, \mathbf{b}_k \right\rangle \cdot h_{\mathbf{r},\mathbf{s}(\mathbf{v})}$ with $\mathbf{d} = \left(\frac{v_1}{w_1}, \ldots, \frac{v_n}{wm}\right)$ for its discrete variant. In case of the very robust γ-divergence [3]

$$D_\gamma\left(v\left(t\right)||w_{\mathbf{r}}\left(t\right)\right) = \frac{1}{\gamma+1}\log\left[\left(\int v\left(t\right)^{\gamma+1}dt\right)^{\frac{1}{\gamma}} \cdot \left(\int w_{\mathbf{r}}\left(t\right)^{\gamma+1}dt\right)\right] - \log\left[\left(\int v\left(t\right)\cdot w_{\mathbf{r}}\left(t\right)^{\gamma}dt\right)^{\frac{1}{\gamma}}\right]$$

one obtains

$$\frac{\delta D_\gamma\left(v\left(t\right)||w_{\mathbf{r}}\left(t\right)\right)}{\delta w_{\mathbf{r}}} = \frac{w_{\mathbf{r}}\left(t\right)^{\gamma}}{\int w_{\mathbf{r}}\left(t\right)^{\gamma+1}d\mathbf{x}} - \frac{v\left(t\right)\cdot w_{\mathbf{r}}\left(t\right)^{\gamma-1}}{\int v\left(t\right)\cdot w_{\mathbf{r}}\left(t\right)^{\gamma}d\mathbf{x}}$$

as Fréchet-derivative.

Example application for clustering NMR data The data base consists of two different data sets both generated from wet lab metabolite mixtures measured by 1-H nuclear magnetic resonance (NMR) spectroscopy. The first set contains of 8 spectra with a metabolite mixture of iso-leucine (Ile), leucine (Leu), valine (Val), glutamine (Glu) and methionine (Met) at different concentrations. The second data set consists of 5 measurements with a mix of lactate (lac), threonine (Thr), proline (Pro) and alanine (Ala). All measurements consist of 2^{14} points in a chemical shift range of $\approx 0.5 - 5$ppm measured with a NMR system of 700.153MHz. The mixtures show in parts strong overlaps of peak structures, complicating the unmixing of the spectra. Reference spectra for the $m = 9$ pure metabolites serve as the set of basis functions and have been simulated using an NMR simulation software with the same settings as for the original measurement device. After preprocessing, the remaining dimensionality of the data as well as basis function vector are $n = 6200$. We show that it is possible to cluster the data base appropriately although the data spaces is very sparse using the above outlined functional clustering approach using a NG-model and divergence D_γ with $\gamma = 0.75$ to take care of the functional character of the data. In particular, the mixing coefficients $\alpha_{\mathbf{r},k}$ of the prototypes reflect the metabolite mixture in the data, see Fig.1.

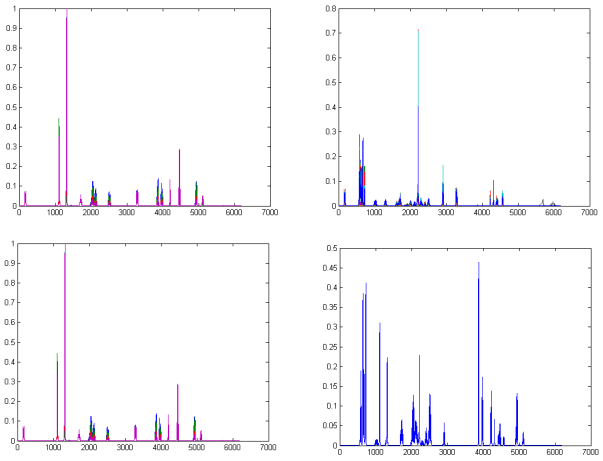


Figure 1: Subsets of the data base of NMR spectra (top, left and right) and the respective prototypes learned by the functional NG-learning based on the D_γ-divergence (bottom).

References

[1] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. Nonnegative Matrix and Tensor Factorizations. Wiley, Chichester, 2009.

[2] B. A. Frigyik, S. Srivastava, and M. Gupta. An introduction to functional derivatives. Technical Report UWEETR-2008-0001, Dept of Electrical Engineering, University of Washington, 2008.

[3] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy conta- mination. Journal of Multivariate Analysis, 99:2053–2081, 2008.

[4] J. Ramsay and B. Silverman. Functional Data Analysis. Springer Science+Media, New York, 2nd edition, 2006.

[5] T. Villmann and S. Haase. Mathematical aspects of divergence based vector quantization using fréchet-derivatives - extended and revised version -. Machine Learning Reports, 4(MLR-01-2010):1– 35, 2010. ISSN:1865-3960.

[6] T. Villmann, S. Haase, F.-M. Schleif, B. Hammer, and M. Biehl. The mathematics of divergence based online learning in vector quantization. In F. Schwenker and N. Gayar, editors, Artificial Neural Networks in Pattern Recognition – Proc. of 4th IAPR Workshop (ANNPR2010), volume 5998 of LNAI, pages 108–119, Berlin, 2010. Springer.

Phenotype Microarray – New Technology, new realm of high dimensional data

Lea Vaas and Markus Göker

DSMZ – German Collection of Microorganisms and Cell Cultures GmbH,
Braunschweig,
Lea.Vaas@dsmz.de

Recently, the set of techniques generating so called "-omics" data was augmented by yet another one, Phenotype Microarrays (PM). In contrast to the existing major technologies, i.e. DNA Microarrays, 2D-Proteomic and chromatographic applications, PM monitors cell respiration over time. Through a redox reaction that alters the colour of a tetrazolium dye in the presence of respiration, kinetic response curves are generated. This provides a high throughput means to characterize microbial metabolism. Yet, the system consists of about 2000 assays for monitoring the cells respiration in the presence of macro- and micronutrients or their reactions to osmotic stress factors, ion or pH effects. The application of a number of chemicals, such as antibiotics, antimetabolites, membrane-active agents, respiratory inhibitors and toxic metals to investigate the cells' sensitivity is also possible. While the PM technology is automated, simple and user-friendly data analysis methods lag behind. This talk will provide an overview of possible applications of this technique and the resulting data types. The two main kinds of applications are (i) identification and drug screening scenarios, where presence-absence calls from each assay can be used for analysis and (ii) comparisons of phenotypes where the evaluation of growth-curve data on a high dimensional level is in demand. Identification and screening analysis could be accomplished with existing clustering (or even phylogenetic inference) methods, whereas strategies for simultaneous analysis of a high number of growth-curves are more demanding. This talk will focus on the statistical challenges this new type of high dimensional data brings along.

Biomarker Discovery: Classification using pooled samples

Anna Telaar, Dirk Repsilber and Gerd Nürnberg

Bioinformatics and Biomathematics group,
Leibniz-Institute for Farm Animal Biology,
telaar@fbn-dummerstorf.de

It is known that a pooled sample design should be avoided for a biomarker search [1]. But pooling can be necessary, for example when not enough amount of RNA per sample exist for a microarray screening experiment. Therefore we consider the effects of sample pooling on classification performance in a simulation study. We simulate different scenarios of discriminating patterns as well as pooled designs with different pool size to find dependency on choice of method. For our simulation of the gene expression data, we simulate a matrix with 60 rows (30 individuals per class) and 1000 columns (corresponding to the features) partitioned in an informative part and non-informative part. We only consider biological variation and the following scenarios of univariate and multivariate patterns for the informative part. Scenario 1: Differentially expressed genes with a mean class difference between [0.1, 0.5]. Scenario 2: A pattern by threshold, one class with values in a special interval, the values of the other class lie outside the interval. But both classes have the same mean value. Scenario 3: A two-dimensional linear pattern, with two linear dependent features. Scenario 4: A two-dimensional circular pattern where the values of one class lie in a circle and the values of the other class outside the circle. The informative part is simulated in different relations to the non-informative part by choosing pattern proportions of 1%, 10% and 20%. We simulated 500 training sets and as test sets, for each repetition we simulated again a 60 x 1000 data matrix under the same conditions. We study four designs for a classification task, pooled sample designs with size of pool $m_p = 2, 3, 5$ and a single sample design ($m_p = 1$). In the single sample design a single sample is measured on an array. In the pooled design, first pools are built and then the pooled samples are analysed. In the pooled design the methods are trained with the pooled samples and tested with independent/new single samples. We use five statistical learning methods: support vector machines with a linear kernel (SVML), support vector machines with a radial basic kernel (SVMR), random forest (RF), powered partial least squares discriminant analysis (PPLS-DA), t-test/linear discriminant analysis (LDA). Comparing a single sample and a pooled design,

we report the prediction errors of these statistical learning methods. For scenario 1 for 1% informative features only the prediction error of PPLS-DA is similar for all poolsizes. Compared to all other learning methods RF shows low increasing prediction errors by comparing the single sample design and designs of poolsize $m_p = 2, 3, 1$ but the error is at least two times higher for a design with pools of size $m_p = 5$. Already for 10% informative features of scenario 1 the pooling effect (of higher prediction errors in pooled designs as for a single sample design) nearly disappears except for the methods SVMR and LDA. The scenario 2 with 1% informative features could only be used by RF for classification in the single sample design where the prediction error is less than 0.03. The prediction error of RF for the pooled designs lies between 0.45 and 0.5. For 10% and 20% informative features for scenario 2 SVMR could also classify the single samples. PPLS-DA shows no great differences between the prediction error in the different designs for the scenario 3 with 1% informative features. An increasing number of informative features (10%, 20%) in scenario 3 leads to a decreasing prediction error for all methods. For scenario 4 with 1% informative features RF has a prediction error less than 0.001 for the single sample design but a prediction error higher than 0.36 for the pooled designs. The other three methods show a prediction error over 0.46 for the scenario 4 in the case of 1% informative features. The prediction error of SVMR decreases while increasing the number of informative features in scenario 4 for the single sample design. We summarize that the prediction errors of pooling design depends strongly on choice of method, pattern scenario and the number of informative features. Most sensitive seems to be the number of informative features followed by pattern scenario. Further simulations are planned for different total numbers of features (5000,20000), to account for technical variance and including an additional error which Kendziorski et al. [2] call the pooling error (different proportions of single samples in a pool).

References

[1] Kerr MK (2003). Design considerations for efficient and effective microarray studies. Biometrics, 59(4): 822–8.

[2] Kendziorski CM, Zhang Y, Lan H, and Attie AD (2003). The efficiency of pooling mRNA in microarray experiments. Biostatistics, 4(3): 465–77.

Quantitative analysis of the intermediate filament network in scanning electron microscopy images by grayvalue-oriented segmentation methods

S. Lück, A. Fichtl, M. Sailer, H. Joos, R.E. Brenner,
P. Walther and V. Schmidt

Institute of Stochastics,
Ulm University,
Electron Microscopy Facility,
Ulm University,
Division for Biochemistry of Joint and Connective Tissue Diseases, Department of Orthopaedics,
Ulm University,
sebastian.lueck@uni-ulm.de

We combine methods from image analysis and spatial statistics to analyze the morphology of intermediate filament (IF) networks as visualized by scanning electron microscopy (SEM). As part of the cytoskeleton in biological cells the IF network determines the elastic properties of cells at large-scale deformations and is thus essential for processes linked to cell migration such as tissue regeneration or metastasis of cancer cells. The elastic response of the cells to large deformations is known to depend on the network architecture, which can be reorganized for tuning of mechanical properties [1]. Methods from spatial statistics can be used to assess physically relevant characteristics of the network graphs, which are automatically extracted from two-dimensional (2D) SEM image data. The estimation of the mesh-size distribution, connectivity or a point-process based analysis of mesh configurations can contribute to a deeper understanding of cell mechanics. Although in general IF networks represent three-dimensional (3D) structures, 2D morphological network characteristics still contain valuable information since the orientation of most filaments is close to parallel with respect to the imaging plane. Moreover, data acquisition for 3D-tomograms of IF networks as conducted in [7] is extremely time-consuming. Thus, 2D methods can be favorable for the efficient statis-

tical investigation of network morphology at large sample sizes, which capture the high variability of IF morphology that is frequently found within single biological scenarios. In our samples most of the cellular structures were removed by detergent extraction such that only the finely woven cytoskeleton remained with the filaments surrounded by vacuum. In combination with subsequent critical-point drying and carbon coating this preparation method yields a superior level of contrast in the SEM image data [8]. Control experiments with cryo-preparation techniques suggested a good preservation of network structure in our samples [8], which were imaged with a Hitachi S-5200 inlens SEM (Tokyo, Japan) at an accelerating voltage of 5 kilovolts and a magnification of 40,000. In previous studies 2D automatic image segmentation for IF networks has been conducted for samples from the cell periphery, where networks exhibited a single-layered planar structure of a rather homogeneous grayvalue distribution [1]. As a consequence, the images could be binarized by simple thresholding, and standard skeletonization techniques could be applied for the extraction of the graph structure. We are extending this approach to non-planar networks which exhibit a drastic variation of the grayvalues in the filamentous phase, where filament contrast decreases in lower network layers (Fig.1a). The resulting problem for binarization could not successfully be solved by the use of local thresholding techniques [2]. Instead we applied the concept of lower λ-leveling kernels to construct grayvalue-oriented crest-lines (Fig.1b) [4]. This algorithm modifies the grayvalue topology of an image in a controlled way by successively lowering pixel values with a specific topology in their neighborhood. The grayvalue-oriented crest-lines resulted in an oversegmentation in brightly imaged thicker network structures containing local minima. The final network graph was therefore computed from a combination of the crest-lines with standard skeletonization, where the latter was based on binarizations by high global thresholdding and dilation of the crest lines (Fig1c). A network graph extracted in this way (Fig.1d) provides only limited information on the network topology around vertices since cross-linked network segments are represented in the same way as filament trajectories which are crossing each other in the 2D images without actual contact in 3D. Correct assessment of the 3D connectivity is however essential for the computation of morphological characteristics such as segment length and mesh-size distribution. Therefore, our algorithmic approach exploits the strong gradient of the surface-dependent secondary electron signal along the filament boundaries to differentiate between merged and disconnected filament trajectories and thus identifies the 3D nature of vertices within the 2D graph. Based on these data a descriptive statistical analysis of network morphology will be presented for three related biological scenarios, namely undifferentiated mesenchymal stem cells, chondrocytes and osteoblasts. In cells from the mesenchymal lineage the cytoskeleton is centrally involved in transmitting signals from extracellular topographic cues and mechanical forces to the nucleus thereby influencing cellular differentiation and function. So far, most work has focused on actin microfilaments. In addition, the network of intermediate filaments (10-12 nm in thickness) is known to be involved in biologic signal transduction [5, 6]. In mesenchymal cells it is mainly composed of the protein vimentin and known to regulate chondrogenic differentiation [3]. The vimentin filaments are prone to active structural reorganization which cannot be comprehensively described by measuring its gross amount or cellular distri-

bution. Therefore, we performed a morphologic analysis based on electron microscopy and used subsequent statistical analysis to gain insight into the complex structural organization and crosslinking of intermediate filaments in undifferentiated mesenchymal stem cells and two well defined cellular phenotypes into which these adult stem cells can differentiate, namely chondrocytes and osteoblasts. A comprehensive profiling on this level could finally help to define novel cellular characteristics related to stemness and differentiation and useful screening methods for early lineage modulating effects of extracellular topographic, mechanical or soluble signals [9].

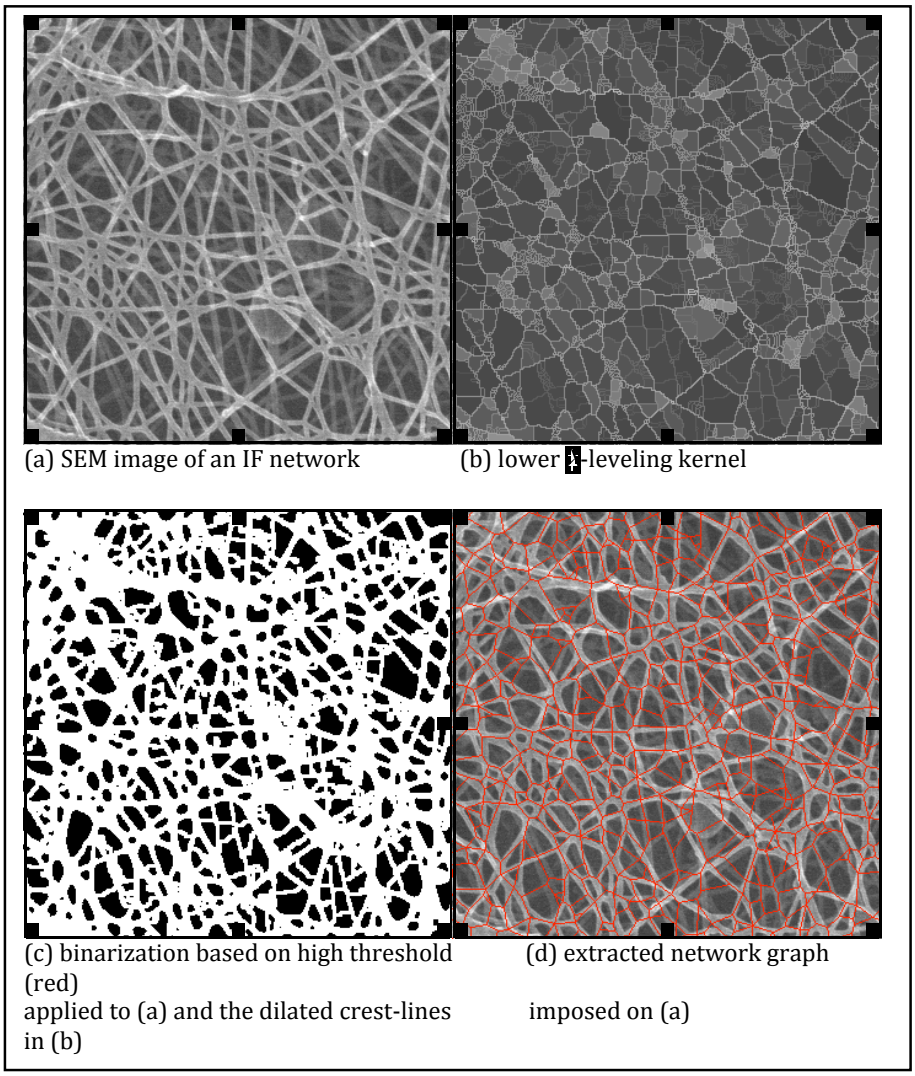


Figure 1

References

[1] M. Beil, H. Braxmeier, F. Fleischer, V. Schmidt, & P. Walther (2005) Quantitative analysis of keratin filament networks in scanning electron microscopy images of cancer cells. J Microsc 220, 84-95.

[2] I. Blayvas, A. Bruckstein & R. Kimmel (2006) Efficient computation of adaptive threshold surfaces for image binarization. Pattern Recogn 39, 89-101.

[3] B. Bobick, R.S. Tuan & F.H. Chen (2010) The intermediate filament vinculin regulates chondrogenesis of adult human bone marrow-derived multipotent progenitor cells. J Cell Biochem 109, 265-276.

[4] M. Couprie, F.N. Bezerra & G. Bertrand (2001) Topological operators for grayscale processing. J Electron Imaging 10, 1003-1015.

[5] H. Herrmann, H. Bär, L. Kreplak, S.V. Strelkov & U. Aebi (2007) Intermediate filaments: from cell architecture to nanomechanics. Nat Rev Mol Cell Biol 8, 562-573.

[6] J. Ivaska, H.M. Pallari, J. Nevo & J.E. Eriksson (2007) Novel functions of vimentin in cell adhesion, migration and signalling. Exp Cell Res 313, 2050-2062.

[7] S. Lück, M. Sailer, V. Schmidt & P. Walther (2010) Three-dimensional analysis of intermediate filament networks using SEM-tomography. J Microsc DOI: 10.1111/j.1365-2818.2009.03348.x.

[8] M. Sailer, K. Höhn, S. Lück, V. Schmidt, M. Beil & P. Walther (2010) Novel electron tomographic methods for three-dimensional analysis of keratin filament networks. Microsc Microanal (in print).

[9] M.D. Treiser, E.H. Yang, S. Gordonov, D.M. Cohen, I.P. Androulakis, J. Kohn, C.S. Chen & P.V. Moghe (2010) Cytoskeleton-based forecasting of stem cell lineage fates. PNAS 107, 610-615.

Analysis of heterosis in Arabidopsis thaliana: A molecular network structure based approach

Sandra Andorf, Joachim Selbig, Thomas Altmann,
Hanna Witucka-Wall and Dirk Repsilber

Leibniz Institute for Farm Animal Biology, Dummerstorf
University of Potsdam, Potsdam-Golm
Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben
andorf@fbn-dummerstorf.de

The heterosis phenomenon, also known as hybrid vigor, was defined already in the year 1908 by Shull [1] as the superiority in performance of heterozygous genotypes compared to their corresponding genetically different homozygous parents. Even though it is already known for a long time and it is widely used in plant breeding [2], the molecular mechanisms are not yet established. We propose a systems biological approach on the basis of molecular network structures to contribute to the understanding of heterosis. Our idea is based on Robertson and Reeve [3], who suggested already in 1952 that heterozygous individuals carry a greater diversity of alleles and are therefore likely to contain additional regulatory possibilities compared to their homozygous parents. These additional regulatory possibilities lead to a higher adaptability of the hybrids and, thus, the heterosis phenomenon. Werhli et al. [4] suggested the use of partial correlations of features of time series profiles to estimate regulatory interactions. So, we based our approach on the calculation of partial correlations of features in the regulatory network to differentiate between homozygous and heterozygous genotypes that show heterosis. We analyzed metabolite (GC-MS) as well as gene expression (microarray) time series data of seven time points of the model plant Arabidopsis thaliana. Under study were the two homozygous accessions C24 and Columbia (Col-0) and the reciprocal crosses for which it is known that they show a biomass heterosis effect [5]. We expect a higher number of regulatory possibilities in the hybrids compared to the homozygous parents. Therefore, regarding our network hypothesis for heterosis [6, 7] the presence of additional regulatory interactions in the regulatory networks of the heterozygous genotypes is expected. Following our hypothesis, these additional regulatory interactions can be identified as

increased significances of the partial correlations in the hybrids in comparison with the homozygous genotypes. This hypothesis of increased significance of the partial correlations of the hybrids which show heterosis could be confirmed for mid-parent and less strong for best-parent heterosis of either heterozygous genotype for our small experimental metabolite as well as gene expression data sets. A further result is that the outcome of the analysis depends on the filtering that is performed to exclude features from the heterosis analysis which do not show a significant time and/or genotype-time point interaction effect in the applied linear model. Too strong filtering leads to the exclusion of features that show a positive partial correlation heterosis effect and in this case no increase in the significance of partial correlations is detectable.

References

[1] G. H. Shull. The composition of a field of maize. American Breeding Association Report, 4:296–301, 1908.

[2] James A Birchler, Donald L Auger, and Nicole C Riddle. In search of the molecular basis of heterosis. Plant Cell, 15(10):2236–2239, 2003.

[3] Forbes W Robertson and E. C Reeve. Heterozygosity, environmental variation and heterosis. Nature, 170(4320):286, 1952.

[4] Adriano V Werhli, Marco Grzegorczyk, and Dirk Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics, 22(20):2523–2531, Oct 2006.

[5] Rhonda C Meyer, Ottó Törjék, Martina Becher, and Thomas Altmann. Heterosis of biomass production in arabidopsis. establishment during early development. Plant Physiol, 134(4):1813–1823, 2004.

[6] Sandra Andorf, Tanja Gärtner, Matthias Steinfath, Hanna Witucka-Wall, Thomas Altmann, and Dirk Repsilber. Towards systems biology of heterosis: A hypothesis about molecular network structure applied for the arabidopsis metabolome. EURASIP J Bioinform Syst Biol, articleID: 147157, 2009. Special Issue: Network structure and biological function: reconstruction, modelling, and statistical approaches.

[7] Sandra Andorf, Joachim Selbig, Thomas Altmann, Kathrin Poos, Hanna Witucka-Wall, and Dirk Repsilber. Enriched partial correlations in genome-wide gene expression profiles of hybrids (a. thaliana): a systems biological approach towards the molecular basis of heterosis. Theor Appl Genet, 120(2):249–259, 2010.

Modeling and simulation of gene-regulatory systems using Boolean networks – a step-by-step introduction

Martin Hopfensitz, Christoph Müssel and Hans A. Kestler

Institute of Neural Information Processing,
University of Ulm,
Department of Internal Medicine I,
University Hospital Ulm,
hans.kestler@uni-ulm.de

In recent years, life scientists have gained more and more understanding of the gene expression processes that control the behaviour of cells. Gene products can influence the expression of other genes, forming extensive gene-regulatory networks. Modeling and simulation of such networks in silico have become indispensible to gain insight into the functioning of cells and can replace costly biological experiments. Boolean networks provide a dynamical model of regulatory process that is able to capture the main properties of biological networks, while being of simple structure[1]. In these networks, a gene is modeled as active or inactive and represented by a Boolean variable. State transitions are calculated by applying Boolean functions associated with the genes synchronously or asynchronously. We recently proposed the R package BoolNet [2], a toolkit for Boolean networks. This package supports all major modeling and simulation steps for synchronous, asynchronous, and probabilistic Boolean networks. In this tutorial, we outline typical work flows in the context of gene-regulatory network simulations and show how they can be realized in BoolNet. Typically, the first step of in silico experiments is the inference of the network. Researchers may infer a network by collecting statements on gene dependencies from literature. Furthermore, they could measure the expression levels of interesting genes at different points of time and try to deduce a network structure from these time series. After constructing a network, researchers can simulate the network to study its dynamics. In this context, the identification of steady states and cycles (so-called attractors) is of particular interest. These represent the states in which the cell resides most of the time. Visualizations of the network structure and dynamics can help understanding the regulatory system. Virtual knock-out and overexpression

experiments can be conducted to assess the influence of certain genes on the regulatory process. The presented methods and tools can be helpful to obtain new hypotheses on a gene-regulatory system which can subsequently be verified in wet-lab experiments.

References

[1] S. A. Kauffman. Metabolic stability and epigensis in randomly constructed genetic nets. Journal of Theoretical Biology, 22(3):437467, 1969.

[2] C. Muüssel*, M. Hopfensitz*, and H. A. Kestler. BoolNet - an R package for generation, reconstruction, and analysis of Boolean networks. Bioinformatics [Epub ahead of print], 2010. * equal contribution.

A practical introduction to MCMC sampling of static Gaussian Bayesian networks

Miriam Lohr and Marco Grzegorczyk

Department of Statistics,

TU Dortmund,

lohr@statistik.tu-dortmund.de, grzegorczyk@statistik.tu-dortmund.de

The ultimate objective of systems biology is the elucidation of the regulatory networks and signalling pathways of the cell. The ideal approach would be the deduction of a detailed mathematical description of the entire system in terms of a set of coupled nonlinear differential equations. As high-throughput measurements at the cell level are inherently stochastic and most kinetic rate constants cannot be measured directly, the parameters of the system would have to be estimated from the data. Unfortunately, multiple parameter sets of nonlinear systems of differential equations can offer equally plausible solutions, and standard optimization techniques in high-dimensional multimodal parameter spaces are not robust and do not provide a reliable indication of the confidence intervals. Most importantly, model selection would be impeded by the fact that more complex pathway models would always provide a better explanation of the data than less complex ones, rendering this approach intrinsically susceptible to over-fitting. To assist the elucidation of regulatory network structures, probabilistic machine learning methods based on Bayesian networks can be employed, as proposed in the seminal paper by Friedman et al. (2000). In a nutshell, the idea is to simplify the mathematical description of the biological system by replacing coupled differential equations by simple conditional probability distributions of a standard form such that the unknown parameters can be integrated out analytically. This results in a scoring function (the marginal likelihood) of closed-form that depends only on the structure of the regulatory network and avoids the over-fitting problem referred to above. In this context it is worth mentioning that (static) Bayesian networks can also be employed for reverse-engineering the structures of regulatoy processes from static (steady-state) data, where approaches based on differential equation models are impossible. Consequently, Bayesian networks have been developed and established as a standard modelling tool in the computational systems biology literature. To obtain the closed-form expression of the marginal likelihood referred to above, two probabilistic models with their respective

conjugate prior distributions have been employed in the past: the multinomial distribution with the Dirichlet prior, leading to the so-called BDe score (Cooper and Herskovits, 1992), and the linear Gaussian distribution with the normal-Wishart prior, leading to the BGe score (Geiger and Heckerman, 1994). These approaches are restricted in that they either require the data to be discretized (BDe) or can only capture linear regulatory relationships (BGe). Practical Bayesian network inference usually follows the Bayesian paradigm and network structures are sampled from the posterior distribution with Markov chain Monte Carlo (MCMC) simulations. Novel fast MCMC algorithms, like Grzegorczyk and Husmeier (2008), can be applied to sys- tematically search the space of network structures for those that are most consistent with the data. In our tutorial we will focus on static Bayesian networks and the Gaussian BGe scoring metric for continuous data. Our MCMC inference will be based on the classical structure MCMC sampler for Bayesian networks (Madigan and York, 1995). Since there is no freely-available R package for this task, R functions, which were written by Miriam Lohr, will be provided and distributed at the tutorial.

References

Cooper, G. F. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9, 309–347.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7, 601–620.

Geiger, D. and Heckerman, D. (1994) Learning Gaussian networks. In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, pp. 235–243. Morgan Kaufmann, San Francisco, CA.

Grzegorczyk, M. and Husmeier, D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. Machine Learning, 71, 265–305.

Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. International Statistical Review, 63, 215–232.

Learning a System by Breaking it – Nested Effects Models at Work

Holger Fröhlich, Tim Beißbarth

Bonn-Aachen International Center for Information Technology,
University of Bonn,
Medical Statistics,
University Medicine Göttingen,
frohlich@bit.uni-bonn.de, tim.beissbarth@ams.med.uni-goettingen.de

The exact reconstruction of cellular pathways and protein interaction networks is a key for the understanding of biological systems. Its knowledge is a prerequisite for identifying target proteins for novel drugs for various diseases (e.g. cancer, diabetes, heart diseases, ...). Due the enormous complexity of cellular systems an exact picture with detailed knowledge on the role and interactions of proteins is still in far future. However, the development of novel perturbation techniques, like RNA interference [2], has added an important tool to the usage of target specific inhibitors and opened new perspectives. Specifically in combination with modern high-throughput methods such techniques offer a powerful tool to get new insights into protein interactions. Nested Effects Models (NEMs) have been established recently as a specific Bayesian network model [8, 7, 3, 9, 5, 4, 11, 6, 10, 1], which is especially designed to learn the signaling flow between perturbed genes from indirect, highdimensional effects. NEMs use a probabilistic framework to compare a given hypothetical network structure with the observed nested structure of downstream effects. The basic idea is that the perturbation of a gene G always leads to a number of observable downstream effects (on gene expression level). If now the perturbation of another gene G' leads to a (noisy) subset of these downstream effects, this is a hint that G indirectly influences G', e.g. is upstream of G'. NEMs have been applied successfully to a significant number of data in the past [8, 7, 9, 5, 11, 10, 1]. In this tutorial we first give a brief introduction into the NEM method itself and then show their application in our R package nem [4]. The R package nem not only contains a variety of different network inference methods, but also a rich number of functions for visualization and analysis of the obtained network. Different approaches for testing the statistical stability of a network via non-parametric bootstrap techniques as well as permutation tests to assess the statistical significance compared to a random network

have been implemented. We will demonstrate the usage of nem via an example workflow analyzing some experimental data. Participants are further encouraged to bring their own data.

References

[1] Benedict Anchang, Mohammad J Sadeh, Juby Jacob, Achim Tresch, Marcel O Vlad, Peter J Oefner, and Rainer Spang. Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. Proc Natl Acad Sci U S A, 106(16):6447–6452, Apr 2009. doi: 10.1073/pnas.0809822106.

[2] A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. Nature, 391:806 – 811, 1998.

[3] H. Fröhlich, M. Fellmann, H. Sültmann, A. Poustka, and T. Beißbarth. Large Scale Statistical Inference of Signaling Pathways from RNAi and Microarray Data. BMC Bioinformatics, 8:386, 2007.

[4] H. Fröhlich, T. Beißbarth, A. Tresch, D. Kostka, J. Jacob, R. Spang, and F. Markowetz. Analyzing gene perturbation screens with nested effects models in R and bioconductor. Bioinformatics, 24(21):2549–2550, Nov 2008. doi: 10.1093/bioinformatics/btn446.

[5] H. Fröhlich, M. Fellmann, H. Sültmann, A. Poustka, and T. Beibarth. Estimating Large Scale Signaling Networks through Nested Effect Models with Intervention Effects from Microarray Data. Bioinformatics, 24:2650– 2656, 2008.

[6] H. Fröhlich, A. Tresch, and T. Beissbarth. Nested effects models for learning signaling networks from perturbation data. Biometrical Journal, 2(51):304 – 323, 2009.

[7] F. Markowetz, D. Kostka, O. Troyanskaya, and R. Spang. Nested effects models for high-dimensional phenotyping screens. Bioinformatics, 23:i305 – i312, 2007.

[8] Florian Markowetz, Jacques Bloch, and Rainer Spang. Non-transcriptional pathway features reconstructed from secondary effects of rna interference. Bioinformatics, 21(21):4026–4032, Nov 2005. doi: 10.1093/bioinformatics/bti662.

[9] A. Tresch and F. Markowetz. Structure Learning in Nested Effects Models. Statistical Applications in Genetics and Molecular Biology, 7(1):Article 9, 2008.

[10] Charles J Vaske, Carrie House, Truong Luu, Bryan Frank, Chen-Hsiang Yeang, Norman H Lee, and Joshua M Stuart. A factor graph nested effects model to identify networks from genetic perturbations. PLoS Comput Biol, 5(1):e1000274, Jan 2009.

[11] C. Zeller, H. Fröhlich, and A. Tresch. A bayesian network view on nested effects models. EURASIP Journal on Bioinformatics and Systems Biology, 195272:8 pages, 2009.

Assessing the robustness of feature selection techniques in resampling experiments

Ludwig Lausser, Christoph Müssel and Hans A. Kestler

Institute of Neural Information Processing,
University of Ulm,
Department of Internal Medicine I,
University Hospital Ulm,
hans.kestler@uni-ulm.de

With the advent of high-throughput biomolecular technologies, highdimensional biological data is increasingly available for many clinical questions. Due to the large size of the data set, many challenges have to be faced when analyzing such data. A common hypothesis is that many biological processes only depend on a small number of genes. This motivates the development and use of feature selection techniques to identify the relevant genes. Recent research has proposed a broad variety of different feature selection methods, which makes it hard to keep track of the specific advantages and disadvantages of certain methods. In many settings, feature reduction is performed before validation steps that involve resampling strategies, such as cross-validation of classifiers. However, it is important to bear in mind that the feature selection methods themselves may yield different sets of features in different resampling settings. We expect a biologically meaningful feature selection to produce stable results in different resampling settings, as it should mainly return the same subset of genes for different probe sets. In this paper, we examine several popular methods with a special respect to stability and robustness. These include Prediction Analysis for Microarrays (PAM) [3], Threshold Number of Misclassifications (TNoM) [1], Relief [2], correlation-based approaches, the area under the ROC curve (AUC), and random selection of features. We analyzed the stability of several feature selection methods in jackknife experiments for different numbers of remaining features on real-world microarray data. Across several subsamples, we measured how often a certain gene was chosen by a specific method. In these experiments, PAM and ROC yield stable selections, whereas Relief is comparatively unstable. Furthermore, we determined groups of feature selection methods that behave similarly

using a hierarchical clustering approach. The clusterings confirm the above results by stating a similar behavior of Relief and random feature selection. Moreover, the results indicate that TNoM selects similar features as the correlation-based approaches.

References

[1] Amir Ben-Dor, L. Bruhn, N. Friedman, M. Schummer, I. Nachman, and Z. Yakhini. Tissue Classification with Gene Expression Profiles. Journal of Computational Biology, 7(3/4):559–584, 2000.

[2] K. Kira and L. A. Rendell. A practical approach to feature selection. In ML92: Proceedings of the ninth international workshop on Machine learning, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[3] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS, 99(10):6567–6572, 2002.

Evaluating the Predictive Power of Random Survival Forests for Breast Cancer Survival

Werner Adler, Sergej Potapov and Matthias Schmid

Department of Biometry and Epidemiology,
University of Erlangen-Nuremberg
werner.adler@imbe.med.uni-erlangen.de

Gene expression data have shown to be predictive for the survival of many diseases, including breast cancer (van't Veer et al., 2002). However, evaluating the performance of survival prediction rules is a complex issue and no standard procedure has been established so far. Heagerty & Zheng (2005) suggested to tackle this problem by using a semiparametric method based on time dependent ROC analysis, adopted from techniques for classification problems. In the classification domain, random forests (Breiman, 2001) are known for their accurate performance. Ishwaran et al. (2008) modified them for censored data, in- troduced as random survival forests (RSF). Based on microarray data, we apply RSF to generate a rule for the prediction of breast cancer survival and evaluate its predictive power by the area under the time dependent ROC curve (AUC). Efron & Tibshirani (1997) introduced the .632+ estimator to reduce the bias of bootstrap estimated classification errors. We present a modification for time dependent AUCs. With a simulation study based on the data of (van't Veer et al., 2002), we compare the modified .632+ method with cross-validation and the bootstrap and discuss the bias and variance of the different resampling techniques.

References

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

Efron, B and Tibshirani, R (1997): Improvements on Cross-Validation: The .632+ Bootstrap Method. JASA, 92(438), 548–560.

Heagerty, PJ and Zheng, Y (2005): Survival Model Predictive Accuracy and ROC Curves. Biometrics, 61, 92–105.

Ishwaran, H, Kogalur, UB, Blackstone, EH and Lauer, MS (2008): Random Survival Forests. Ann App Stat, 2(3), 841–860.

van't Veer, LJ, Dai, HY, van de Vijver, MJ, et al. (2002): Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871), 530–536.

A Robust Alternative to the Schemper-Henderson Estimator of Prediction Error

Matthias Schmid

joint work with T. Hielscher, T. Augustin and O. Gefeller,
Department of Medical Informatics, Biometry and Epidemiology,
Friedrich-Alexander-University Erlangen-Nuremberg,
matthias.schmid@imbe.med.uni-erlangen.de

In clinical applications, the prediction error of survival models has to be taken into consideration to assess the practical suitability of conclusions drawn from these models. Different approaches to evaluate the predictive performance of survival models have been suggested in the literature. In this talk, we analyze the properties of the estimator of prediction error developed by Schemper and Henderson (2000), which quantifies the absolute distance between predicted and observed survival functions. We show that the estimator proposed by Schemper and Henderson is not robust against misspecification of the survival model, i.e. the estimator will only be meaningful if the model family used for deriving predictions has been specified correctly. To overcome this problem, we propose an alternative estimator of the absolute distance between predicted and observed survival functions. We show that this modified Schemper-Henderson estimator is robust against model misspecification, allowing its practical application to a wide class of survival models. The properties of the Schemper-Henderson estimator and its new modification are illustrated by means of a simulation study and the analysis of real world data.

Automatic sound recording segmentation for sound source and sound phase separation

Sebastian Krey and Uwe Ligges

Fakultät Statistik,

Technische Universität Dortmund,

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

For quite a lot music information retrieval tasks a segmentation of the source sound recording in different phases is necessary (beside more complex sounds, even isolated instrumental sounds change over time). This segmentation can be used as a base for further data reduction steps, which are usually necessary for tasks using for example classification methods. For monophonic instrument recordings clustering methods based on classic soundfeatures like static Mel-frequency cepstral coefficients (MFCC) often result in good and interpretable results. For more difficult situations different segmentation approaches for multivariate time series, for example [1], can be used and their results are presented and compared to the clustering methods. Additionally the suitability of more modern sound features like cepstral modulation ratio regression (CMRARE)[2], which are also using a short-time fourier transformation for the intial spectral analysis, like static MFCCs, and wavelets for this task are investigated.

References

[1] Graves D., Pedrycz W. (2009). "Multivariate Segmentation of Time Series with Differential Evolution", Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, pp. 1108-1113.

[2] Martin R., Nagathil A. (2009). "Cepstral Modulation Ratio Regression (CMRARE) Parameters for Audio Signal Analysis and Classification", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 321-324.

A Hybrid Information Fusion Approach to Discover Events in EEG Data

Martin Schels, Stefan Scherer, Michael Glodek,
Hans A. Kestler, Friedhelm Schwenker and Günther Palm

Institute for Neural Information Processing,
University of Ulm,
Department of Internal Medicine I,
University Hospital Ulm,
firstname.lastname@uni-ulm.de

In the past few years mind reading and brain computer interfaces became a prominent application in neuroscience [1]. In the present study the goal is to investigate the possibility to automatically determine whether a human subject has just seen a target on a presented image by solely analyzing the individual event related potentials (ERP), recorded using an electroencephalograph (EEG). ERPs typically reflect cognitive processes in the brain that follow a more or less strict timely pattern that can be visualized by filtering and averaging of many ERP signal recordings [2]. In the present investigation visual stimuli were presented by following a typical oddball paradigm: the non target (background) type stimuli were presented very frequently, whereas the targets were displayed very rarely which leads to a prominent ERP, the P300 [3]. Such an oddball scenario can be found in the dataset provided by the Machine Learning for Signal Processing 2010 Competition: Mind Reading (http://www.bme.ogi.edu/∼hildk/mlsp2010Competition.html). The challenge of this dataset is the classification of stimuli by analyzing EEG recordings. For these recordings, satellite images were presented in a fast sequence to a test person, who was instructed to push a button when a surface to air missile (SAM) site was shown. The data consists of 64 EEG channels, that are recorded with a sampling rate of 256 Hz. A feature and decision fusion approach for this neuroscience application will be described subsequently. In order to prepare the data for classification five different features were extracted locally from every EEG channel. As it is possible to analyze ERP in both, the frequency and the time domain [4], we decided to follow both approaches to extract suitable features. Before passing the calculated features to our machine learning approach, the EEG channels were partitioned into nine overlapping

areas containing up to 18 channels each. Subsequently, the resulting 45 sets of channels – due to combining the five kinds of features with the nine partitions – were trained and classified separately using a fuzzy-input fuzzy-output support vector machine (F^2-SVM) with a RBF kernel as proposed in [5]. In order to mitigate the problems arising with a highly imbalanced distribution of the labels in the dataset due to the oddball property, the F^2-SVM was extended by loss parameters, punishing misses of the rare class stronger. The performance of the classifiers is determined by the area under the ROC curve. Results of this first classification step are ranging from a classification performance of 0.54 to 0.88, depending on the partition and feature. Finally, a decision fusion step was implemented to combine the outputs of the obtained F^2-SVM. We decided to use an averaging classifier fusion approach. The set of 45 classifiers offered $2^{45} - 1$ combinations for possible classifier fusions. In order to find a suitable combination, a very basic genetic search algorithm approach was implemented [6]. Thus, combinations of classifiers which further increased the area enclosed by the ROC curve up to 0.965, were found. It is shown, that difficulties of separating the categories using the proposed weak features for the particular EEG channel is mitigated by partitioning the 64 EEG channels. The obtained classifiers, based on the defined partitions, showed only moderate performances in terms of the area under the ROC curve. These results were further improved by applying a decision fusion approach.

Acknowledgment This paper is based on work done within the "Information Fusion" subproject of the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems", funded by the German Research Foundation (DFG). The work of Martin Schels is supported by a scholarship of the Carl-Zeiss Foundation.

References

1. Chumerin, N., Manyakov, N.V., Combaz, A., K., S.J.A., Yazicioglu, R.F., Torfs, T., Merken, P., Neves, H.P., Van Hoof, C., Van Hulle, M.M. In: P300 Detection Based on Feature Extraction in Online Brain-Computer Interface. Volume 5803/2009 of KI 2009: Advances in Artificial Intelligence. Springer Berlin / Heidelberg (2009) 339–346

2. Gray, H.M., Ambady, N., Lowenthal, W.T., Deldin, P.: P300 as an index of attention to self-relevant stimuli. Journal of Experimental Social Psychology 40(2) (2004) 216–224

3. Dujardin, K., Derambure, P., Bourriez, J.L., Jacquesson, J.M., Guieu, J.D.: P300 component of the event-related potentials (erp) during an attention task: effects of age, stimulus modality and event probability. International Journal of Psychophysiology 14(3) (May 1993) 255–267

4. Picton, T., Bentin, S., Berg, P., Donchin, E., Hillyard, S., Johnson, R., Miller, G., Ritter, W., Ruchkin, D., Rugg, M., Taylor, M.: Guidelines for using human event-related

potentials to study cognition: Recording standards and publication criteria. Psychophysiology 37(02) (2000) 127–152

5. Thiel, C., Scherer, S., Schwenker, F.: Fuzzy-input fuzzy-output one-against-all support vector machines. In: Knowledge-Based Intelligent Information and Engineering Systems 2007, Springer (2007) 156–165

6. Bäck, T.: Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford University Press, Oxford, UK (1996)

A more realistic simulation approach for the prediction of genetic values from genome-wide Introduction marker data

N. Melzer, D. Wittenburg and D. Repsilber

Leibniz Institute for Farm Animal Biology,
Dummerstorf,
n.melzer@fbn-dummerstorf.de

In our simulation study we want to validate and develop methods to predict milk phenotypes in cows. The phenotype prediction is based on genome-wide marker data. There exist approaches in the field of genomic selection, e.g. Meuwissen et al. (2001). They apply methods to simulated data. In this and similar simulation studies genetic variation is caused by additive and dominance effects at different loci. This method is labelled as conventional and it is compared to our approach. Our approach is to test existing methods on more realistic data. To simulate more realistic data we use the metabolome level as intermediate level of the the genotype-phenotype map. The metabolome level is integrated by using a curated SBML model (Holzhütter, 2004). Different scenarios are simulated for both approaches and resulting datasets are evaluated using the fast Bayes method (Meuwissen et al., 2009). The bovine genotype is simulated using annotated SNP positions from the Illumina SNP-Chip Bovine 50k, in total 52273 SNPs on a 30 Morgan genome. To simulate a realistic SNP distribution we adopt a population genetics simulation approach. We start with a homozygous population with an effective populations size of 100 and use a mutation-drift model. For the first to the 400th generation the following was done: random mating including recombination events and each SNP locus can mutate with a rate of $2.5 * 10^{-3}$. After the 400th generation linkage disequilibrium was obtained as $r^2 = 0.13$ (r^2 is the mean of pairwise correlation for all adjacent marker loci). The generations 401 and 402 are the training generations which consist of 1,000 individuals, in the design of 50 half-sib families with 20 offspring. Each individual was genotyped and phenotyped. The test generations 403 and 404 were built up as successive generations from the training set. For the conventional approach additive genetic effects were drawn from a gamma distribution (Hayes and Goddard, 2001) and dominance effects were drawn from a normal distribution

(Bennewitz and Meuwissen, 2009) to determine the genetic values. In contrast, in our approach similar to Mendes et al. (2003) the genotype is used to change enzyme kinetic parameters for the SBML model (in our case glycolysis, pentose phosphate pathway and glutathione metabolism). The SBML model was running till it reached the steady state. Further the parameter change leads to a varied equilibrium metabolite concentration. Afterwards the genetic value was determined as sum over the end concentrations of part of the metabolites. To receive the phenotype, a random error component, which depends on the chosen broad- sense heritability, is added to the simulated genetic value in both approaches. Datasets are simulated which include 23 QTL and 230 QTL with heritabilities $h^2 = \{0.1, 0.3, 0.5\}$. Further we used the fast Bayes method for the complete data set (52,273 SNPs) and for a reduced data set where each tenth SNP (in total 5,227 SNPs) was taken, including the true QTL positions. Both approaches are compared regarding predicting precision, which is determined as the correlation between estimated and simulated genetic values of the test generations. The following preliminary results are based on the analysis of 100 simulated datasets for each scenario. The results show that the use of different heritabilities and the quantity of QTLs has an influence on the accuracy of the fast Bayes method for both approaches. It was expected, and the results confirm that, the prediction precision for the conventional approach is higher than for the new approach. The reason for that lies in the approaches, because in the conventional approach only additive and dominance effects are simulated. These effects and further effects of interaction were contained in our approach. Additionally, we could observe that for using more QTLs the precision of prediction decreases. That means, if we use more QTLs the genetic variance is split over more positions. Using the complete dataset also led to de- creased prediction precision. Best results were achieved for the case of 23 QTLs and the reduce dataset. At the moment for the SBML model just the maximum velocity (Vmax) is changed according to the genotype. Analyses where also other parameters are affected are under way. It will also be of interest to compare the new approach with the conventional including effects of interaction.

References

Bennewitz, J. and Meuwissen, T. (2009). Book of Abstracts of the 60th Annual Meeting of the EAAP, page 320.

Hayes, B. and Goddard, M. E. (2001). Genet Sel Evol, 33(3):20929. Holzhütter, H. G. (2004). Eur J Biochem, 271(14):290522.

Mendes, P., Sha, W., and Ye, K. (2003). Bioinformatics, 19 Suppl 2:ii1229.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Genetics, 157(4):181929.

Meuwissen, T. H., Solberg, T. R., Shepherd, R., and Woolliams, J. A. (2009). Genet Sel Evol, 41:2.

Sequential Benchmarking

Manuel J. A. Eugster and Friedrich Leisch

Department of Statistics,
LMU München,
Firstname.Lastname@stat.uni-muenchen.de

Benchmark experiments draw B independent and identically distributed learning samples from a data generating process to estimate the (empirical) performance distribution of candidate algorithms. Formal inference procedures are used to compare these performance distributions and to investigate hypotheses of interests. In most benchmark experiments B is a "freely chosen" number, often specified depending on the algorithms' runtime to setup experiments which finish in reasonable time. In this presentation we provide first thoughts on how to control B and remove its "arbitrary" aftertaste. General sequential designs enable, amongst other things, to control the sample size, i.e., B. A benchmark experiment can be seen as a sequential experiment as each run, i.e., drawing a learning sample and estimating the candidates' performances, is done one by one. Currently, no benefit is taken from this sequential procedure: The experiment is considered as a fixed-sample experiment with B observations and the hypothesis of interest is tested using a test T at the end of all B runs. We propose to take the sequential nature of benchmark experiments into account and to execute a test T successively on the accumulating data. In a first step, this enables to monitor the benchmark experiment – to observe p-value, test statistic and power of the test during the execution of the benchmark experiment. In a second step, this information can be used to make a decision – to stop or to go on with the benchmark experiment. We present and discuss methods, group sequential and adaptive, suitable for benchmark experiments.

Benchmarking and Analysis of Local Classification Methods

Bernd Bischl and Julia Schiffner

Chair of Computational Statistics,
TU Dortmund,
{bischl, schiffner}@statistik.tu-dortmund.de

Local approaches to classification are widely used. Well-known examples are the k nearest neighbors method [3] and Cart [1]. In recent years many more local classification methods have been developed. Among these are, for example, localized versions of standard classification techniques like linear discriminant analysis [2] and Fisher discriminant analysis [5], logistic regression [4, 6] as well as boosting [7]. Here, two questions arise: How can local classification methods be characterized and when are they especially appropriate? In the relevant literature the term 'local' is often only vaguely defined as relating to the position in some space, to a part of a whole or to something that is not general or widespread. Often, it lacks an exact explanation of its particular meaning. We present steps towards a framework for a unified description of local methods and show that different types of local approaches can be distinguished. Moreover, it is not clear which properties local methods have and for which types of classification problems they may be beneficial. Generally, localized classification methods exhibit more flexibility than their global counterparts. Therefore they are expected to give good results in case of irregular class boundaries. A special situation, which is addressed by many local methods, is multimodality of class distributions. To our knowledge there are very few extensive studies in literature that compare several types of local methods across many data sets. In order to assess their performance we conduct a benchmark study on real-world and synthetic data. Different types of local methods are compared to global methods. We try to identify subgroups of similar algorithms and set these subgroups into relation to our theoretical intuitions about them. Also we will try to give general guidelines in what situation which method will probably lead to good results.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. CART: Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.

[2] I. Czogiel, K. Luebke, M. Zentgraf, and C. Weihs. Localized linear discriminant analysis. In R. Decker and H.-J. Lenz, editors, Advances in Data Analysis, volume 33 of Studies in Classification, Data Analysis, and Knowledge Organization, pages 133–140, Berlin Heidelberg, 2007. Springer.

[3] E. Fix and J. L. Hodges. Discriminatory analysis – nonparametric discrimination: Consistency properties. Report 4, U.S. Airforce School of Aviation Medicine, Randolph Field, Texas, 1951.

[4] D. J. Hand and V. Vinciotti. Local versus global models for classification problems: Fitting models where it matters. The American Statistician, 57(2):124–131, May 2003.

[5] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. Journal of Machine Learning Research, 8:1027–1061, May 2007.

[6] G. Tutz and H. Binder. Localized classification. Statistics and Computing, 15:155–166, 2005.

[7] C.-X. Zhang and J.-S. Zhang. A local boosting algorithm for solving classification problems. Computational Statistics & Data Analysis, 52:1928–1941, 2008.

mlr and benchmark: Setup, Execution and Analysis of Benchmark Experiments in R

Bernd Bischl and Manuel J. A. Eugster

Chair of Computational Statistics,
TU Dortmund,
Workgroup Computational Statistics,
LMU Munich,
bischl@statistik.tu-dortmund.de, manuel.eugster@stat.uni-muenchen.de

The benchmarking process in machine learning experiments abstractly consists of three levels: (1) The Setup defines the design of a benchmark experiment; data sets, candidate algorithms, performance measures and a suitable resampling strategy are declared. (2) In the Execution level the defined setup is carried out. Here, computational aspects play a major role; an important example is the parallelization of the experiment on different computers. (3) And finally the Analysis, where the calculated raw performance measures are analyzed using exploratory and inferential methods. Main objective is a statistically correct order of the candidate algorithms according to one ore more relevant measures. The Setup and Execution layer of benchmark experiments are realized using the mlr package [1]. It provides a generic, object-oriented interface to many machine learning methods in R for classification or regression. It enables the researcher to succinctly define the complete Setup stage, allows for various meta-optimization algorithms, e.g. to tune the learning machines or to perform variable selection. Regarding the Execution stage, it supports the relevant packages in R to easily parallelize code with different job-size levels. The Analysis layer is realized using the benchmark package [2]. It provides exploratory and inferential methods for different units of observations of a benchmark experiment; two examples are algorithms' (empirical) performance distributions and dataset characteristics. In case of the algorithms' performance distributions (the most common unit of observation) the package provides a specialized benchmark experiment plot, formal methods to investigate hypotheses of interest and to finally infer statistical correct order relations of the algorithms. The demonstration will feature different use cases, mainly with toy examples to explain the functionality. We will also add comments from our ongoing research in this area regarding real-world examples and our experiences with modeling and software. R examples will be provided on-line and users may execute

them on their laptops during the talk. If possible, we will also present an example of a parallelized experiment on an R cluster.

References

[1] B. Bischl. mlr: Machine learning in R. http://mlr.r-forge.r-project.org, 2010.

[2] M. J. A. Eugster, T. Hothorn, and F. Leisch. Exploratory and inferential analysis of benchmark experiments. Technical Report 30, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, 2008.

Boosting for Estimating Spatially Structured Additive Models

Nikolay Robinzonov and Torsten Hothorn

Institut für Statistik,

Ludwig-Maximilians-Universität München,

nikolay.robinzonov@stat.uni-muenchen.de, torsten.hothorn@stat.uni-muenchen.de

Spatially structured additive models offer the flexibility to estimate regression relationships for spatially and temporally correlated data. Here, we focus on the estimation of conditional deer browsing probabilities in the National Park "Bayerischer Wald". The models are fitted using a componentwise boosting algorithm. Smooth and non-smooth base learners for the spatial component of the models are compared. A benchmark comparison indicates that browsing intensities may be best described by non-smooth base learners allowing for abrupt changes in the regression relationship.

Constrained Regression Using mboost: An Application to Body Fat Prediction

Benjamin Hofner

Institut für Medizininformatik, Biometrie und Epidemiologie,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
benjamin.hofner@imbe.med.uni-erlangen.de

Today, overweight and obesity are widespread and have a huge impact on public health as they strongly increases risks for cardiovascular diseases, diabetes and other popular diseases. Thus, the assessment of body fat is important to properly diagnose the nutritional status in individuals. DXA (dual energy X-ray absorptiometry) is considered to be a highly reliable but complex and expansive method to asses the body composition in epidemiological studies. Garcia et al. (2005) proposed a linear model to replace the DXA measurements by easily obtained measures such as skinfold thickness and body circumferences. We improve this model by using a generalized additive model (GAM) with smooth effects, which offers a very flexible scheme for model estimation. By defining appropriate degrees of freedom, the flexibility of the smooth function estimates can be controlled. However, the actual shape remains unspecified.In many applications this is not desirable as researchers have a priori assumptions on the shape of the estimated effects such as monotonicity. Skin-fold thickness and body circumference are, for example, considered to have a monotone (increasing) influence on body fat. We will present a method that allows to incorporate such monotonicity constraints in GAMs: We base estimation on B-splines with difference penalty (i.e., P-splines) and use an additional asymmetric L2 penalty to enforce monotonicity (Eilers, 2005). Estimation and variable selection is conducted using the very flexible boosting framework as implemented in the R package mboost (Hothorn et al., 2010). Practical considerations and results regarding monotonic P-splines will be illustrated in the context of body fat prediction. An outlook to further constrained regression problems which can be tackled using mboost will be given.

References

Eilers, P. H. C. (2005). Unimodal smoothing, Journal of Chemometrics 19: 317–328.

Garcia, A. L., Wagner, K., Hothorn, T., Koebnick, C., Zunft, H.-J. F. and Trippo, U. (2005). Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths, Obesity Research 13: 626–634.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010). mboost: Model-Based Boosting. R package version 2.0-4.
URL: http://cran.R-project.org/package=mboost

Prediction Intervals and Quantile Boosting: A Simulation Study

Andreas Mayr, Nora Fenske and Torsten Hothorn

Institut für Medizininformatik, Biometrie und Epidemiologie,
Friedrich-Alexander-Universitaät Erlangen-Nürnberg,
Institut für Statistik,
Ludwig-Maximilians-Universitaät München,
andreas.mayr@imbe.med.uni-erlangen.de

Prediction intervals are a useful tool to express uncertainty in the prediction of future or unobserved realizations of the response variable in a regression setting. Standard approaches typically assume an underlying distribution function and use the variance of the estimation method to compute boundaries around the expected mean. Meinshausen (2006) suggested to use quantile regression forests to construct nonparametric prediction intervals for new observations. He adopted this generalization of random forests to estimate not only the conditional mean but the full conditional distribution of a response variable and, therefore, also conditional quantiles. Compared with classical methods, the resulting intervals have the advantage that they do not depend on distributional assumptions and are computable for high-dimensional data sets. In this talk, we present an adaptation of gradient boosting algorithms to compute intervals based on additive quantile regression (Fenske et al., 2009), as available in the R package mboost (Hothorn et al., 2010). The boundaries of prediction intervals are modeled by applying nonparametric quantile regression with linear as well as smooth effects, fitted by component-wise boosting providing intrinsic variable selection and model choice. The main advantage of this highly flexible approach is that it allows to quantify and to interpret the influence of single covariates on the response and on the prediction accuracy. We found that the correct interpretation of prediction intervals involves the risk of running into a severe pitfall in practice since only the conditional view based on fixed predictor variables is adequate to prove the correct coverage of the proposed intervals. Hence, we analyze simulated data sets to evaluate the accuracy of our methods and show a real-life example to emphasize their practical relevance.

References

Fenske, N., Kneib, T. and Hothorn, T. (2009). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression, Technical Report, Department of Statistics, University of Munich 52.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010). Model-Based Boosting. R package version 2.0-4.

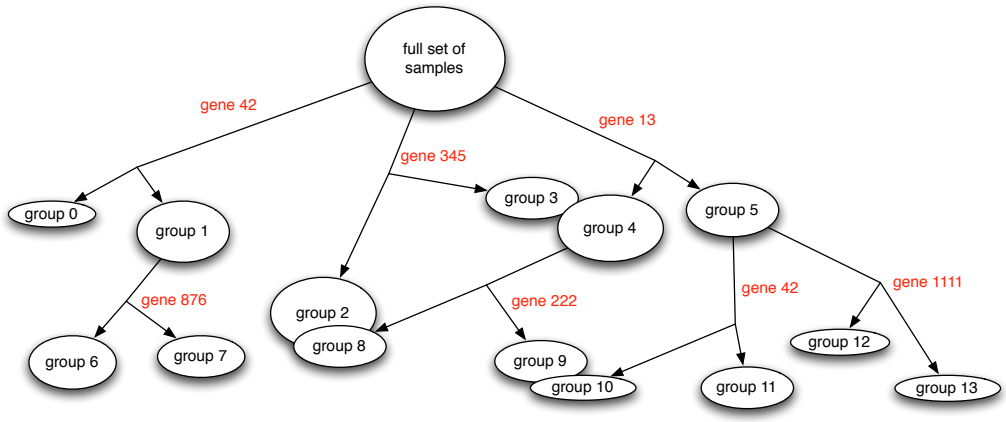
Meinshausen, N. (2006). Quantile regression forests, Journal of Machine Learning Research 7: 983–999.

Dimensionality reducing cluster analysis

Johann M. Kraus, Stefan Schultz and Hans A. Kestler

Department of Internal Medicine I,
University Hospital Ulm,
Institute of Neural Information Processing,
University of Ulm,
hans.kestler@uni-ulm.de

Unsupervised machine learning operates on data without access to labels. The aim of cluster analysis is to infer relevant characteristics of the data and partition the data into groups of similar appearance. In the context of mircroarray data analysis we deal with high-dimensional and noisy data. It is of essential importance that the cluster methods are not too much influenced by that noise. Our goal is the incorporation of dimensionality reducing techniques into clustering giving special attention to noise robustness. In this regard we developed a method that exhibits noise robustness as a means of finding relevant one-dimensional subspaces of the data. This includes the significant quantization of values in order to find patterns of similarly regulated genes. We propose a cluster procedure based on scan statistics[1] that unfolds a directed acyclic graph by repeatedly splitting groups along the most important genes.



In the Figure shown a significant grouping of samples is found when scanning gene 42 for a region with high density (group 1). Scanning gene 876 on this group identifies two more subgroups (group 6, group 7). To summarize, our method does not partition the data set into a fixed number of groups. Every data point will be associated with several groups, as well as different sequences of single-axis splits may lead to similar groupings.

References

[1] J. Glaz, J. Naus, S. Wallenstein: Scan statistics. Springer, Heidelberg (2001)

Zig-zag exploratory factor analysis with more variables than observations

Steffen Unkel and Nickolay T. Trendafilov

Department of Mathematics and Statistics,
The Open University Milton Keynes,
s.unkel@open.ac.uk

In this paper, the problem of fitting the exploratory factor analysis (EFA) model is reconsidered. A new approach for EFA of data matrices with more variables than observations is presented. The EFA model is viewed as a specific data matrix decomposition with fixed unknown matrix parameters. A new algorithm named zig-zag factor analysis is introduced for the least squares estimation of all EFA model unknowns. As in principal component analysis, zig-zag factor analysis is based on the singular value decomposition of data matrices. Another advantage of the proposed computational routine is that it facilitates the estimation of both common and unique factor scores. Numerical examples with simulated data and a high-dimensional data set from genome research illustrate the algorithm and the EFA solutions.

Liste der bisher erschienenen Ulmer Informatik-Berichte

Einige davon sind per FTP von `ftp.informatik.uni-ulm.de` erhältlich
Die mit * markierten Berichte sind vergriffen

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

95-12 *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^{cc} 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 Beth, 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

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm