ulm university universität ulm

Statistical Computing 2014

Abstracts der 46. Arbeitstagung

HA Kestler, M Schmid,

L Lausser, JM Kraus (eds)

Ulmer Informatik-Berichte

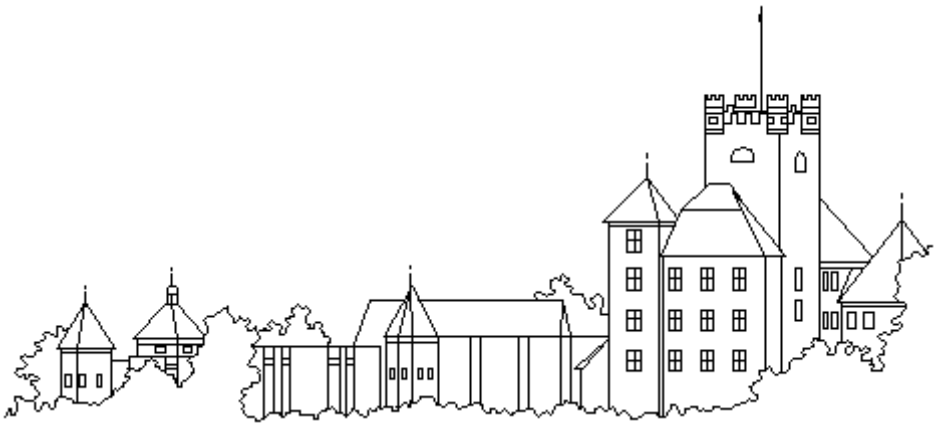
Nr. 2014-04

July 2014

International Graduate School in Molecular Medicine Ulm

SYSTAR

Statistical Computing 2014



46. Arbeitstagung

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

20.07.-23.07.2014, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, July 20, 2014

18:00-20:00		**Dinner**
20:00-21:00		**Chair: Hans A. Kestler**
20:00-21:00	Ulrich Mansmann (München)	Genes and Functions: Prediction, Regression and systems' modelling

Monday, July 21, 2014

08:50-09:00		**Opening of the workshop: H.A. Kestler, H. Binder, M. Schmid**
09:00-10:30		**Chair: Berthold Lausen**
09:00-09:20	Thomas Villmann (Mittweida)	Statistical Quality Measures and ROC-Optimization by Learning Vector Quantization Classifiers
09:20-9:40	Giuseppe Casalicchio (München)	Beyond discrimination and calibration: why the predictiveness curve is not sufficient for assessing the performance of prediction models
9:40-10:00	Marianna Grinberg (Dortmund)	Statistical methods for large scale gene expression data sets in toxioogenomics
10:00-10:20	Matthias Kuhn (Dresden)	Specific identification of small genomic structural variations using next generation sequencing data
10:20-10:50		**Coffee break**
10:50- 14:40		**Chair: Harald Binder**
10:50-11:10	Andre Burkovski (Ulm)	Rank aggregation of heterogeneous data for identification of common genesets
11:10-11:30	Werner Adler (Erlangen)	Comparing Classification for Optical Tissue Differentiation
11:30-11:50	Jörn Lötsch, Alfred Ultsch (Marburg)	What do all those MIRna do?
12:00-13:30		**Lunch**
13:30-14:30	Eyke Hüllermeier (Paderborn)	Learning from imprecise and fuzzy data
14:30-14:40	Johanna Mazur, Aslihan Gerhold-Ay (Mainz)	Teasers: Tutorials
14:40-15:00		**Coffee break**
20:30-21:30	Johanna Mazur, Aslihan Gerhold-Ay (Mainz)	Tutorial: Statistical Workflows for Sequencing Data

Tuesday, July 22, 2014

09:00-10:20		**Chair: Axel Benner**
09:00-09:20	Eric Sträng (Ulm)	Di-methylation is necessary for a sharp Notch response
09:20-09:40	Melanie Grieb (Ulm)	Predicting new Phenotypes with a Boolean Network incorporating uncertainty
09:40-10:00	Alexander Groß (Ulm)	Predicting the dynamic behavior of Wnt/ß -catenin and Wnt/JNK signaling by a rule based probabilistic modeling approach
10:00-10:20	Markus Maucher (Ulm)	A critical noise level for the reconstruction of Boolean functions from time series data
10:20-10:50		**Coffee break**
10:50-11:50		**Chair: Bernd Bischl**
10:50-11:10	Julia Schiffner (Düsseldorf)	A Mixture of Experts Approach for the Analysis of SNP data
11:10-11:30	Katrin Madjar (Dortmund)	Subgroup-specific survival analysis in high-dimensional datasets
11:30-11:50	Andreas Mayr (Erlangen)	Boosting the concordance index for survival data
12:00-13:30		**Lunch**
13:30-18:00		**Chair: Matthias Schmid**
13:30-14:30	Axel Benner (Heidelberg)	Genomic Biomarkers for Personalised Medicine: Identification and Validation
14:30-15:00	Anthony Rossini (Basel)	Putting Statistics back into Statistical Computing
15:00-15:30		**Coffee break**
15:30-15:50	Manuela Zucknick (Heidelberg)	Non-identical Twins: Comparison of Frequentist and Bayesian Lasso for Cox Models
15:50-16:10	Michael Glodek (Ulm)	Inequality-constraint Multi-class Fuzzy-in Fuzzy-out Support vector machines
16:10-16:30	Markus Kächele (Ulm)	Importance based hierarchical Langrange multiplier filtering for parallel training of Support Vector machines
16:30-16:50	Christoph Müssel, Ludwig Lausser (Ulm) Bernd Bischl (München), Florian Schmid (Ulm)	Teasers: Tutorials

16:50-17:50		Working groups meeting on **Statistical Computing 2015** and other topics (all welcome)
18:00-20:00		**Dinner**
20:00-21:00	Christoph Müssel, Ludwig Lausser (Ulm)	Tutorial: Boolean networks
	Bernd Bischl (München), Florian Schmid (Ulm)	Tutorial: Algorithm Configuration / Tuning with R

Wednesday, July 23, 2014

09:00-10:20		**Chair: Markus Maucher**
09:00-09:20	Rainer Dangl (Wien)	A Web Application for Generating Benchmarking Data
09:20-09:40	Riccardo De Bin (München)	Subsampling versus bootstrap in resampling-based model selection for multivariable regression
09:40-10:00	Johann Kraus (Ulm)	Semantic clustering
10:00-10:20	Florian Schmid (Ulm)	Semantic multi-classifier systems
10:20-10:50		**Coffee break**
10:50-11:50		**Chair: Johann Kraus**
10:50-11:10	Ludwig Lausser (Ulm)	Fold change classifiers
11:10-11:30	Sebastian Krey (Dortmund)	An Statistical approach for Modeling of Low Frequency Oscillations in Electricity Networks
11:30-11:50	Axel Fürstberger (Ulm)	Interactive zoomable alignment graphs for pairwise wild base nucleotide protein alignment
12:00-13:30		**Lunch**

Genes and Function: Prediction, Regression and systems' modelling

Ulrich Mansmann

On the one hand, prediction and regression techniques applied to molecular data represent the so called top-down approach to molecular complex data: knowledge free approaches which identify central players in a complex system. On the other hand, systems modelling combines existing knowledge on the interaction of molecular components in a bottom-up process to a complex system. My talk discusses both approaches in the light of the data made available by The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). It is of interest how both approaches are complementary and how both combined can help to discover new functionalities in the complex system.

IBE, LMU, München

ulrich.mansmann@lmu.de

Statistical Quality Measures and ROC-Optimization by Learning Vector Quantization Classifiers

Michael Biehl[1], Marika Kaden[2], and Thomas Villmann[2]

Introduction - Classification by Learning Vector Quantization

Learning vector quantization (LVQ) models are prototype-based adaptive classifiers for processing vectorial data (Kohonen, 95). Training samples are assumed to be of the form $\mathbf{v} \in V \subseteq \mathbb{R}^n$ with class labels $x_{\mathbf{v}} = x(\mathbf{v}) \in \mathcal{C} = \{1, \ldots, C\}$. The set of prototypes $W = \{\mathbf{w}_j \in \mathbb{R}^n, j = 1 \ldots M\}$ contains representatives of the classes carrying prototype labels $y_j \in \mathcal{C}$. Classification decisions for unknown data samples $\tilde{\mathbf{v}}$ are usually made according to a winner take all rule, i.e.

$$x_{\tilde{\mathbf{v}}} := y_{s(\tilde{\mathbf{v}})} \text{ with } s(\tilde{\mathbf{v}}) = argmin_j \left(d(\tilde{\mathbf{v}}, \mathbf{w}_j)\right)$$

where $d(\tilde{\mathbf{v}}, \mathbf{w}_j)$ is a dissimilarity measure in the data space, frequently chosen as the Euclidean distance. LVQ training amounts to distributing the prototypes in the data space such that the classification error is minimized. Stochastic gradient descent learning have been introduced which is based on objective function

$$E(W, f) = \frac{1}{2} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})) \tag{1}$$

approximating the classification error (Sato et al., 96). Here, the function

$$\mu(\mathbf{v}) = \frac{d^+(\mathbf{v}) - d^-(\mathbf{v})}{d^+(\mathbf{v}) + d^-(\mathbf{v})} \tag{2}$$

is the so-called classifier function. This approach is known as Generalized LVQ (GLVQ) (Sato et al. 96). Here $d^+(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^+)$ denotes the dissimilarity between the data vector $\mathbf{v}$ and the closest prototype $\mathbf{w}^+$ with the same class label $y_{s^+} = x_{\mathbf{v}}$, while $d^-(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^-)$ is the distance from the best matching prototype $\mathbf{w}^-$ with a class label y_{s^-} different from $x_{\mathbf{v}}$. The *modu-*

[1]Johann-Bernoulli-Institute for Mathematics and Computer Sciences, University Groningen, The Netherlands
[2]Computational Intelligence Group, University of Applied Sciences Mittweida, Germany

thomas.villmann@hs-mittweida.de

lation function f in (1) is a monotonically increasing function usually chosen as a sigmoid or the identity function. A typical choice is the Fermi function

$$f_\theta(x) = \frac{1}{1 + a \cdot \exp\left(-\frac{(x-x_0)}{2\theta^2}\right)} \tag{3}$$

with $x_0 = 0$ and $a = 1$ as standard parameter values. The parameter θ determines the slope of f_θ but is frequently fixed as $\theta = 1$.

Stochastic gradient learning performs update steps of the from

$$\triangle \mathbf{w}^{\pm} \propto -\frac{\partial f_\theta(\mu(\mathbf{v}))}{\partial \mu(\mathbf{v})} \cdot \frac{\partial \mu(\mathbf{v})}{\partial d^{\pm}(\mathbf{v})} \cdot \frac{\partial d^{\pm}(\mathbf{v})}{\partial \mathbf{w}^{\pm}} \tag{4}$$

for a randomly chosen data sample $\mathbf{v}$.

Classification Accuracy and Statistical Measures in GLVQ

As described above, the standard GLVQ optimizes the approximated classification error E from (1). We observe that the classifier function $\mu(\mathbf{v})$ from (2) becomes negative if the data point $\mathbf{v}$ is correctly classified, i.e. if $x_{\mathbf{v}} = y_{s(\mathbf{v})}$ is valid. Further, in the limit $\theta \searrow 0$ the sigmoid f_θ (3) becomes the Heaviside function

$$H(x) = \begin{cases} 0 \;\; if \;\; x \leq 0 \\ 1 \; else \end{cases}, \tag{5}$$

such that *border sensitive classification learning* takes place (Villmann et al., 2013). In this case, $E(W, H)$ counts the *misclassifications*. Considering a two-class problem with a positive class C_+ labeled by '$\oplus$' and a negative class C_- with class label '$\ominus$', the misclassifications are the false positives (FP) and false negatives (FN) according to the contingency table Tab. .

labels		true		
		C_+	C_-	
predicted	C_+	TP	FP	$\widehat{N}_+$
	C_-	FN	TN	$\widehat{N}_-$
		N_+	N_-	N

Table 1 Contingency / Confusion matrix: TP - true positives, FP - false positives, TN - true negatives, FN - false negatives, $N_\pm$- number of positive/negative data, $\widehat{N}_+$ - number of predicted positive/negative data.

Yet, counting of misclassifications is not always an appropriate evaluation of classifier, in particular, if the data are imbalanced (Sachs, 1992). In statistical analysis contingency table evaluations are well-known to deal with this problem more properly. Several measures were developed to judge the classification quality based on the confusion matrix emphasizing different aspects.

For example, *precision* π and *recall* ρ, defined by

$$\pi = \frac{TP}{TP + FP} = \frac{TP}{\widehat{N}_+} \text{ and } \rho = \frac{TP}{TP + FN} = \frac{TP}{N_+}, \tag{6}$$

respectively, are used in the widely applied F_β-measure

$$F_\beta = \frac{\left(1 + \beta^2\right) \cdot \pi \cdot \rho}{\beta^2 \cdot \pi + \rho} \tag{7}$$

developed by Rijsbergen (1979).

To integrate these contingency quantities into a GLVQ-like cost function, we have to approximate them properly while ensuring their dependence on the prototypes is differentiable. For this purpose we introduce the quantity $\hat{\mu}(\mathbf{v}) = f_\theta(-\mu(\mathbf{v}))$ with $\hat{\mu}(\mathbf{v}) \approx 1$ iff the data point $\mathbf{v}$ is correctly classified and $\hat{\mu}(\mathbf{v}) \approx 0$ otherwise for small values θ, with the derivative

$$\frac{\partial \hat{\mu}(\mathbf{v})}{\partial \mathbf{w}^\pm} = -\frac{\partial \hat{\mu}(\mathbf{v})}{\partial f_\theta} \cdot \frac{\partial f_\theta}{\partial \mu} \cdot \frac{\partial \mu}{\partial d^\pm(\mathbf{v})} \cdot \frac{\partial d^\pm(\mathbf{v})}{\partial \mathbf{w}^\pm}.$$

Thus we can express all quantities of the confusion matrix in terms of the new classifier function $\hat{\mu}(\mathbf{v})$:

$$TP = \sum_{\mathbf{v}} \delta_{\oplus, x_\mathbf{v}} \cdot \hat{\mu}(\mathbf{v}), \quad FP = \sum_{\mathbf{v}} \delta_{\ominus, x_\mathbf{v}} \cdot (1 - \hat{\mu}(\mathbf{v}))$$

$$FN = \sum_{\mathbf{v}} \delta_{\oplus, x_\mathbf{v}} \cdot (1 - \hat{\mu}(\mathbf{v})) \text{ and } TN = \sum_{\mathbf{v}} \delta_{\ominus, x_\mathbf{v}} \cdot \hat{\mu}(\mathbf{v})$$

with $\delta_{\oplus, x_\mathbf{v}}$ is the Kronecker symbol and $\delta_{\ominus, x_\mathbf{v}} = 1 - \delta_{\oplus, x_\mathbf{v}}$. Obviously, all these quantities are also differentiable with respect to $\hat{\mu}(\mathbf{v})$ and, hence, also with respect to the prototypes $\mathbf{w}_k$. In consequence, an arbitrary general statistical measure S can be optimized by a GLVQ-like stochastic gradient learning of the prototypes, if it is *continuous and differentiable* with respect to TP, FP, FN, and TN. Clearly, the above mentioned quantities precision π and recall ρ as well as the F_β-measure belong to this function class and, therefore, can be plugged into the GLVQ learning scheme.

Receiver Operation Characteristic Optimization and GLVQ

The Receiver Operation Characteristic (ROC) is an important tool for performance comparison of classifiers. A classifier is considered superior if it delivers a higher value of the *area under the ROC-curve* (AUC). Suppose a two-class problem of classes A and B according to datasets V_A and V_B with cardinalities $\#V_A$, $\#V_B$, respectively. Further assume that a classifier delivers a continuous output (discriminant function) used for the classification decision. Then the AUC can be interpreted as the probability P_{AB} that

a classifier will rank a randomly chosen A-instance $\mathbf{v}_A \in V_A$ higher than a randomly chosen B-instance $\mathbf{v}_B \in V_B$ (Fawcett, 2006).

This interpretation of the AUC can be facilitated in the GLVQ-framework (Villmann, 2014): To this end, the discriminant function

$$\mu_{AB}(\mathbf{v}) = \frac{d^B(\mathbf{v}) - d^A(\mathbf{v})}{d^A(\mathbf{v}) + d^B(\mathbf{v})} \tag{8}$$

is defined with $d^A(\mathbf{v}) = d^A(\mathbf{v}, \mathbf{w}_A^*(\mathbf{v}))$ where $\mathbf{w}_A^*(\mathbf{v})$ is the closest prototype to $\mathbf{v}$ responsible for class A. Analogously, $\mathbf{w}_B^*$ and $d^B(\mathbf{v})$ are defined in the same manner. We consider the (local) ordering function

$$O_\theta(\mathbf{v}_A, \mathbf{v}_B) = f_\theta(\mu_{AB}(\mathbf{v}_A) - \mu_{AB}(\mathbf{v}_B)) \tag{9}$$

for an ordered pair $(\mathbf{v}_A, \mathbf{v}_B)$ of vectors. Then, the ROC cost function can be calculated as

$$E_{ROC}(\theta, V_A, V_B) = \frac{1}{\#V_{AB}} \sum_{(\mathbf{v}_A, \mathbf{v}_B)} O_\theta(\mathbf{v}_A, \mathbf{v}_B) \tag{10}$$

depending on the slope parameter θ of the sigmoid function $f_\theta(x)$ from (3). Border sensitive learning, i.e. forcing $\theta \searrow 0$ in (9) , leads to the limit

$$E_{ROC}(\theta, V_A, V_B, W) \xrightarrow{\theta \searrow 0} P_{AB} \; . \tag{11}$$

Further, using the derivatives

$$\frac{\partial \mu_{AB}(\mathbf{v})}{\partial \mathbf{w}_A^*(\mathbf{v})} = \frac{d^B(\mathbf{v})}{d^{AB}(\mathbf{v})} \cdot \frac{\partial d^A(\mathbf{v})}{\partial \mathbf{w}_A^*(\mathbf{v})} \text{ and } \frac{\partial \mu_{AB}(\mathbf{v})}{\partial \mathbf{w}_B^*(\mathbf{v})} = -\frac{d^A(\mathbf{v})}{d^{AB}(\mathbf{v})} \cdot \frac{\partial d^B(\mathbf{v})}{\partial \mathbf{w}_B^*(\mathbf{v})}$$

we can calculate the gradients of the ordering function $O_\theta(\mathbf{v}_A, \mathbf{v}_B)$ regarding both $\mathbf{v}_A$ and $\mathbf{v}_B$, respectively:

$$\frac{\partial O_\theta(\mathbf{v}_A, \mathbf{v}_B)}{\partial \mathbf{w}_A^*(\mathbf{v}_A)} = \left.\frac{\partial f_\theta}{\partial z}\right|_z \cdot \left(\frac{\partial \mu_{AB}(\mathbf{v}_A)}{\partial \mathbf{w}_A^*(\mathbf{v}_A)} - \frac{\partial \mu_{AB}(\mathbf{v}_B)}{\partial \mathbf{w}_A^*(\mathbf{v}_A)}\right) \tag{12}$$

$$\frac{\partial O_\theta(\mathbf{v}_A, \mathbf{v}_B)}{\partial \mathbf{w}_A^*(\mathbf{v}_B)} = \left.\frac{\partial f_\theta}{\partial z}\right|_z \cdot \left(\frac{\partial \mu_{AB}(\mathbf{v}_A)}{\partial \mathbf{w}_A^*(\mathbf{v}_B)} - \frac{\partial \mu_{AB}(\mathbf{v}_B)}{\partial \mathbf{w}_A^*(\mathbf{v}_B)}\right) \tag{13}$$

$$\frac{\partial O_\theta(\mathbf{v}_A, \mathbf{v}_B)}{\partial \mathbf{w}_B^*(\mathbf{v}_A)} = \left.\frac{\partial f_\theta}{\partial z}\right|_z \cdot \left(\frac{\partial \mu_{AB}(\mathbf{v}_A)}{\partial \mathbf{w}_B^*(\mathbf{v}_A)} - \frac{\partial \mu_{AB}(\mathbf{v}_B)}{\partial \mathbf{w}_B^*(\mathbf{v}_A)}\right) \tag{14}$$

$$\frac{\partial O_\theta(\mathbf{v}_A, \mathbf{v}_B)}{\partial \mathbf{w}_B^*(\mathbf{v}_B)} = \left.\frac{\partial f_\theta}{\partial z}\right|_z \cdot \left(\frac{\partial \mu_{AB}(\mathbf{v}_A)}{\partial \mathbf{w}_B^*(\mathbf{v}_B)} - \frac{\partial \mu_{AB}(\mathbf{v}_B)}{\partial \mathbf{w}_B^*(\mathbf{v}_B)}\right) \tag{15}$$

with $z = \mu_{AB}(\mathbf{v}_A) - \mu_{AB}(\mathbf{v}_B)$.

In consequence, GLVQ-like stochastic gradient learning is possible also for the ROC cost function E_{ROC} from (10), which delivers an AUC-optimizing scheme in the limit $\theta \searrow 0$ of border sensitive learning.

Conclusion

We present in this extended abstract the mathematical framework for learning of prototype-based LVQ-classifiers to optimize statistical quality measures based on the confusion matrix or receiver operating characteristic by stochastic gradient learning. Obviously, this approach can be easily combined with other advanced GLVQ-techniques like relevance and matrix learning or kernelized variants (Villmann et al., 2002, Schneider et al.,2009, Villmann et al. 2014).

References

1. T. Kohonen. Self-Organizing Maps. Springer, 1995 (Second Extended Edition 1997).
2. A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference (Eds), pp. 423–9. Cambridge, MA, USA: MIT Press, 1996.
3. M. Kästner, M. Riedel, M. Strickert, W. Hermann, T. Villmann. Border-sensitive learning in kernelized learning vector quantization. In Rojas, G. Joya, and J. Cabestany (Eds.), Proc. of the 12th International Workshop on Artificial Neural Networks (IWANN), ser. LNCS, I. , vol. 7902, no. Part I., pp. 357–366. Springer, Berlin, 2013.
4. L. Sachs, Angewandte Statistik, 7th ed. Springer Verlag, 1992.
5. C. Rijsbergen, Information Retrieval, 2nd ed. London: Butterworths, 1979.
6. T. Fawcett. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874, 2006.
7. M. Biehl, M. Kaden, P. Stürmer, and T. Villmann. ROC-optimization and statistical quality measures in learning vector quantization classifiers. Machine Learning Reports 8, no. MLR-01-2014, 23–34, 2014, ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_01_2014.pdf.
8. B. Hammer and T. Villmann. Generalized relevance learning vector quantization. Neural Networks 15(8-9), 1059–1068, 2002.
9. P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. Neural Computation 21, 3532–3561, 2009.
10. T. Villmann, S. Haase, M. Kaden. Kernelized vector quantization in gradient-descent learning. Neurocomputing, 2014 (in press)

Beyond discrimination and calibration: why the predictiveness curve is not sufficient for assessing the performance of prediction models

Giuseppe Casalicchio[1], Bernd Bischl[1], Matthias Schmid[2]

To assess the performance of prediction models, it is commonly agreed that the prediction model should satisfy two major criteria. First, they should have a high discriminative power, meaning that they are able to well separate cases from controls. Second, they should be well calibrated, meaning that the expected number of events should closely agree with the observed number of events. The predictiveness curve is often used as a graphical tool to visualize and evaluate calibration and discrimination (see e.g. Huang et al., 2007; Pepe et al., 2008; Moons et al., 2012).

In this talk, we will analyze the properties of the predictiveness curve and will review its role in the assessment of the performance of prediction models. Based on these considerations, we will illustrate that the concept has several major shortcomings and should therefore not be solely used for the evaluation of prediction accuracy.

References

1. Y. Huang et al. Evaluating the predictiveness of a continuous marker. Biometrics 63(4), 1181–8, 2007.
2. K.G.M. Moons et al. Quantifying the added value of a diagnostic test or marker. Clinical chemistry 58(10), 1408–17, 2012.
3. M.S. Pepe et al. (2008). Integrating the predictiveness of a marker with its performance as a classifier. American journal of epidemiology 167(3), 362–8.

[1]Statistical Consulting Unit, Department of Statistics, Ludwig-Maximilians-University, 80539 Munich, Germany
[2]Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany

matthias.schmid@ukb.uni-bonn.de

Statistical methods for large-scale gene expression data sets in toxicogenomics

Marianna Grinberg[1], Eugen Rempel[1], Jan G. Hengstler[2], Jörg Rahnenführer[1]

Understanding chemically-induced toxicity is important for the development of drugs and for the identification of biomarkers. Recently, large-scale gene expression data sets have been generated to understand molecular changes on a genome-wide scale. The Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system (TG-GATEs) is an open-source project in Japan (http://toxico.nibio.go.jp). The database contains data for more than 150 compounds applied to cells from rats (liver, kidney) and to human hepatocytes. Our goal is the characterization of the compounds with statistical methods. For many compounds, incubations with different concentrations and for different time periods are available. Besides the curse of dimensionality (many more variables than observations) the statistical analysis is faced with additional complexity including batch effects and implausible concentration progression.

Main initial analysis goals are discriminant analysis and cluster analysis for compounds. Prior to this statistical analysis, a concentration progression analysis is performed to identify compounds with implausible measurements. Principal component analysis provides an overview of the associations between the compounds and also pinpoints artifacts like batch effects in the data sets. We present techniques to control for batch effects and to correct for unreasonable results. The curated database serves as basis for the extraction of relevant biomarkers. The most reliable ones are assigned to the corresponding mechanisms of toxicity and are compared to gene sets related to different diseases, such as fatty liver, cirrhosis and hepatocellular cancer. Finally, we aim at additional insight into biological relationships between compounds by analyzing differential effects on the level of transcription fac-

[1]Department of Statistics, TU Dortmund, Dortmund, Germany
[2]Leibniz Research Centre for Working Environment and Human Factors (IfADo), Dortmund, Germany

rahnenfuehrer@statistik.tu-dortmund.de

tors. Therefore, we apply enrichment type tests for discovering transcription factors that target more differentially expressed genes than expected under a random model.

Specific identification of small genomic structural variations using next generation sequencing data

Matthias Kuhn

Introduction

Next generation sequencing (NGS) is a technique that promises to unbundle genetic variability with low bias and hence to advance our understanding of, e.g. tumorigenesis. Changes of single nucleotides (SNV) and insertions/deletions of up to 50 nucleotides (indels) form the best known source of genetic variation. Another broad class of genetic changes are structural variations (SV) that involve bigger chunks of DNA. SVs can be further classified as insertions, duplications, deletions, inversions, translocations or mixtures thereof.

Currently, there exist well established methods which allow to call SNVs and short indels by basically mapping and comparing NGS reads to a reference genome. Also concerning SVs, methods have been developed for their identification. In contrast to SNV calling, the methods for SV detection mainly utilize reads that could not be properly mapped to the reference genome.

However, most SV calling methods are tuned to find rather big SVs that go beyond kilobases. Hence, there is a grey zone of genetic variation in between SNVs and SVs that may not be detected by any of these methods. Therefore, we aimed at developing a method to specifically find genomic variation in the range of 50 up to 350 nucleotides with NGS data. This focus on small SVs gives hope that the method is applicable also to exome sequencing where only short target regions (for instance exome regions) are covered.

Department of Medical Informatics and Biometry, University Hospital Carl Gustav Carus, Technical University Dresden, Dresden, Germany

matthias.kuhn@statistik.tu-dresden.de

Material and Methods

As input the method expects NGS reads of an exome-enriched tumor-normal pair of samples from one patient. The sequencing reads are mapped to the human reference genome. In particular, the method relies on a NGS read mapper that is able to map reads partially, hence generating clipped reads or even multi-part alignments. The recently released BWA MEM mapping software is used here. Differential small SVs are called in a two-step procedure.

First, we identify candidate positions for a small SV by filtering at loci that are sufficiently well covered in the normal sample and that show an increase in the rate of clipped reads in the tumor sample. We rely on clipped reads as it is a universal marker for any type of SVs.

At those candidate positions, we extract further features that characterize the differences of the local mapping patterns between tumor and normal sample. A support vector machine (SVM) is employed as pattern learning algorithm in order to identify outlier regions that are different from most other regions with regard to the extracted mapping features. We prefer a one-class SVM over a classification SVM because there are plenty of different types of SVs (like insertion, deletion or inversion and mixtures thereof) that do not form a homogeneous class in itself. For the fitting of one-class SVMs we relied on the libsvm library.

To train the SVM on the individual patient data the mapping pattern of non-candidate positions are used as examples of loci with no SV-style genetic differences between tumor and normal sample. Adding in-silico SVs at some of these regions in the tumor sample simulates loci with differential SVs. This allows to train and evaluate a personalised SVM for each patient. Hyper-parameters of the SVM (like the kernel parameter) are tuned with cross-validation (CV) on the training data. Finally the selected one-class SVM model is applied to the initially screened candidate regions as external test data. Candidate loci that are marked as outliers by the SVM represent the predicted small differential SVs of the method.

Results

To start evaluating the described method in a controlled setting we applied it to fully simulated tumor-normal samples. We randomly selected a number of enrichment targets, added in-silico SVs and simulated Illumina NGS reads with an average read depth of 60, a coverage depth that is often achieved in exome sequencing.

The usage of clipped reads as screening method for regions with small differential SVs was assessed in the simulation run. The proportion of clipped reads at a region achieved high sensitivity (say, 95%) only at the cost of a too high false positive rate (70%). Instead, we found that the average number

of clipped bases per read in a locus is a much more sensitive and specific screening method.

We further found that for the problem at hand SVMs with a radial kernel performed better than linear and polynomial kernels. In a cross-validation round we tuned the kernel parameter of the radial kernel to yield best f1-score. On a separate validation data we reached an overall accuracy of 90.4%. Specificity was 91.9% while sensitivity was only 61.5%. We implemented the described method as an R package. It works currently only for simulated tumor-normal data.

Discussion

We present a method to detect genetic variation in the grey zone between SNVs and SV, ranging from 50 to 350 nucleotides. Because it targets such small SVs there is reason to believe the method actually also works with exome sequencing. Another novelty is that the method tries to incorporate the typical tumor-normal setting in tumorigenesis research by directly considering the differences between the local mapping patterns at a locus of tumor and normal samples of an individual patient.

The method screens for small differential SVs by looking for an increased proportion of clipped bases at a locus in the tumor probe. Subsequently, a one-class SVM pattern learning algorithm is trained on inconspicuous genetic regions of the individual patient together with regions with in-silico generated differential SVs. These regions of known differential SVs allow to tune the SVM to end up with a personalised SVM per patient. The features are based on the differences of the local mapping pattern between tumor and normal sample at sufficiently covered loci.

For now, the results are restricted because we have tested the method only to fully simulated data. There it looks promising although the achieved sensitivity is still poor. We think that it will be fruitful to further explore new features based on differences in mapping patterns. Also, a sensitivity analysis on the effect of key parameters from sequencing (enrichment, coverage), mapping (penalty scores) and the learning machine remains to be done. Once the method works reliably with high sensitivity on simulated data it still needs to stand the test of real exome sequencing data.

Rank aggregation of heterogeneous data for identification of common genesets

Andre Burkovski, Florian Schmid, Ludwig Lausser, Hans A. Kestler

Gene expression studies are based on a large variety of model organisms and different experimental conditions. Single studies include geneset analyses which help to identify commonly regulated pathways. The integration of the individual results can help to identify the common processes through data interaction.

We present a method for cross-platform and inter-species integration of gene expression data. The method combines ranked gene lists though rank aggregation procedures that produces a consensus ranking - a ranking with which all studies least disagree. Differentially expressed genes that are common across the individual studies tend to have high ranks in the resulting consensus ranking. This consensus rank information is further utilized in a geneset analysis. Each geneset is rated according to the consensus rank information of the member genes. The geneset's rating is computed by an adaption of an AUC measure. The adapted AUC-statistic can be used to assess the significance of gene sets. Furthermore, our method can be used to identify common processes (e.g. KEGG pathways) in gene expression data that are only discovered by the combination of different experiments.

We apply the proposed method to gene expression data from different experiments for characterization of young and old. Our goal is to identify common processes in aging that may be revealed by combining different dataset that use different aging models and gene expression arrays.

Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

{andre.burkovski, florian-1.schmid, ludwig.lausser, hans.kestler}@uni-ulm.de

Comparing Classifiers for Optical Tissue Differentiation

Alexander Engelhardt[1], Rajesh Kanawade[2], Christian Knipfer[3], Matthias Schmid[4], Florian Stelzle[3], Werner Adler[1]

In the field of oral and maxillofacial surgery, laser surgery provides a number of advantages over traditional surgery, e.g. a lower risk of infections or increased precision. For example to prevent damage to nerve tissue, it is of high importance to correctly recognize the tissue type that is cut by the laser. Diffuse reflectance spectroscopy is an optical method that provides information on the tissue type. First results on using diffuse reflectance spectra for the discrimination of tissue types in oral laser surgery are available (e.g. Stelzle et al., 2010). However, due to the small sample sizes of currently available data a sound comparison of several classification methods is not possible with real data.

Based on a multivariate Gaussian model, we perform the simulation of a large number of diffuse reflectance spectra of different tissue types and compare the ability of several classification methods like LDA, classification trees, random forest, or penalized discriminant analysis (PDA) to correctly recognize these tissue types. We report and discuss classification results and compare performances based on the simulated data with results obtained using a small real world data set.

[1]Department of Biometry and Epidemiology, University of Erlangen-Nuremberg, Germany
[2]SAOT - Graduate School in Advanced Optical Technologies, Erlangen, Germany
[3]Department of Oral and Maxillofacial Surgery, Erlangen University Hospital, Erlangen, Germany
[4]Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany

Werner.Adler@fau.de

References

1. F. Stelzle, K. Tangermann-Gerk, W. Adler, A. Zam, M. Schmidt, A. Douplik, and E. Nkenke. Diffuse reflectance spectroscopy for optical soft tissue differentiation as remote feedback control for tissue-specific laser surgery. Lasers in surgery and medicine, 42(4), 319–325, 2010.

What do all those MIRnas do?

Alfred Ultsch[1], Christian Pallasch[2], Sabine Herda[1] and Jörn Lötsch[3]

Micro-RNAs (miRNA) are attributed to the systems biological role of a regulatory mechanism of the expression of protein coding genes. The present work uses the set of all genes currently known to be directly regulated by miRNAs to describe the systems biological role of all miRNAs. More than thousand empirically verified miRNA-gene interactions yield a set of n = 798 miRNA regulated genes. Knowledge Discovery on a gene ontology over-representation analysis (ORA) identified that a particular function of a large set of miRNA regulations is to control the expression of those genes that in turn regulate the expression of genes. The ORA on a comparative set of genes, where numerical methods predict a miRNA-gene interaction, independently revealed regulation of genes involved in control of gene expression. As such, we have identified a novel type of miRNA regulation principle that affects not only the known translational level in the cytoplasm, but in particular the control of gene expression by mRNA transcription and processing in the nucleus. In conclusion, we propose that a fundamental function of miRNAs is exerted on a superior regulatory level by the control of the expression of genes that control the expression of genes, i.e., hyper-regulation of gene expression.

[1]Department of Computer Science and Mathematics, Data Bionics Research Group, University of Marburg, Hans-Meerwein-Strae 22, D-35032 Marburg, Germany
[2]Laboratory of Cellular Immunotherapy, Clinic I, University of Cologne, Cologne
[3]Institute of Clinical Pharmacology, Goethe-University Hospital, Theodor Stern Kai 7, D-60590 Frankfurt am Main, Germany

ultsch@nformatik.uni-marburg.de

Learning from imprecise and fuzzy data

Eyke Hüllermeier

Methods for analyzing and learning from imprecise or "fuzzy" data have attracted considerable attention in recent years. In many cases, however, existing methods (for precise, non-fuzzy data) are extended to the imprecise/fuzzy case in an ad-hoc manner, and without carefully considering the interpretation of modeling concepts such as intervals or fuzzy sets when being used for representing data. Distinguishing between an ontic and an epistemic interpretation of (fuzzy) set-valued data, and focusing on the latter, it will be argued that a generalization of learning algorithms based on an application of the generic extension principle is not appropriate. In fact, the extension principle fails to properly exploit the inductive bias underlying typical machine learning methods, although this bias, at least in principle, offers a means for "disambiguating" the imprecise/fuzzy data. Alternatively, a novel method is proposed which is based on the generalization of loss functions in empirical risk minimization, and which performs model identification and data disambiguation simultaneously.

Department of Computer Science, University of Paderborn, Pohlweg 47-49, 33098 Paderborn, Germany

eyke@upd.de

Statistical Workflows for Sequencing Data

Johanna Mazur, Aslihan Gerhold-Ay

Next-generation sequencing (NGS) data, like RNA-Seq or ChIP-Seq, are becoming more and more important for medical research. They enable us to obtain information for the development of gene signatures for prediction of clinical endpoints like death, combine information from different molecular genetic levels and select differentially expressed genes between different groups.

The tutorial "Statistical Workflows for Sequencing Data" will combine theoretical information about RNA-Seq and ChIP-Seq data, issues concerning experiment design from the statistical perspective and practical information for the analysis of RNA-Seq data. With a hands-on example the participants will be enabled to perform the statistical analysis of RNA-Seq to find differentially expressed genes.

Core Facility Bioinformatics, Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg-Universitt Mainz, Mainz, Germany

mazur@uni-mainz.de, aslihan.gerhold-ay@unimedizin-mainz.de

Di-methylation is necessary for a sharp Notch response

Eric Sträng[1], Franz Oswald[2], Hans A. Kestler[1]

The Notch signaling pathway plays a crucial role in development, maintenance and differentiation of cells. The Notch receptor is membrane bound which upon ligand activation releases the Notch intracellular domain (NICD) which localizes in the nucleus. NICD then binds to RBPJ to create the core activation complex which binds to DNA binding sites to enable transcription of the target gene. New experimental data suggests that the the di-methylation status of NICD transactivation domain plays an important role on stability. This is however not reflected by the transcriptional output of the cells when measured by Luciferase gene reporter assay. Furthermore embryonic phenotypes (Xenopus laevis, Dario rerio) show responses which are inconsistent with the prior findings.

In order to shed some light on this apparent contradiction, we model the Notch pathway in order to quantify the transcriptional efficiency of the pathway. An ODE model of the core component is set up using data from litterature. Simulation suggest that di-methylation of the transactivation domain plays a crucial role in the Notch response. In particular proper di-methylation is necessary for a sharp response. Methylation mutant show a lesser peak but longer response upon NICD induction.

[1]Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany
[2]Internal Medicine I, University Hospital of Ulm, 89069 Ulm, Germany

{eric.straeng, franz.oswald, hans.kestler}@uni-ulm.de

Predicting new Phenotypes with a Boolean Network incorporating uncertainty

Melanie B. Grieb[1], Andre Burkovski[1], J. Eric Sträng[1], Johann M. Kraus[1], Alexander Groß[1], Susanne Kühl[2], Günther Palm[3], Michael Kühl[2], Hans A. Kestler[1]

Background

Boolean networks (Kauffman, 1993) are models for Gene Regulatory Networks that allow a gene to be on (1) or off (0). However, the state of a gene can have multiple values, e.g. due to polyploidy, gene expression measurements from multiple cells or noise. We created a Boolean Network Extension (BNE) that allows values in the interval $[0, 1]$. As the Boolean transition functions are only defined for Boolean values, we use product-sum fuzzy logic transition functions for the BNE. The transformation of Boolean network functions to BNE functions is created in two steps: In the first step, the rules are transformed to canonical DNF. In the second step, the canonical DNF rules are directly translated with product-sum fuzzy logic.

The BNE only uses the structure of the Boolean network, without any additional parameters. We apply the BNE to a Boolean network model of cardiac development (Herrmann et al., 2012). We numerically find the fixed points of the BNE and map them to the closest Boolean phenotype.

Results

The fixed points obtained by the simulation of the network with the BNE lie on curves that depend on an external parameter of the network. When the

[1]Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany
[2]Institute for Biochemistry and Molecular Biology, Ulm University, Ulm, Germany
[3]Institute of Neural Information Processing, Ulm University, Ulm, Germany

hans.kestler@uni-ulm.de

fixed points are mapped to their closest Boolean phenotype, we obtain the fixed points found by the original Boolean network model. More importantly we find additional fixed points that were not found in the Boolean network. The values of the additional fixed points found by our model are consistent with biological experiments in *Xenopus* (Gessert et al., 2009).

References

1. S.A. Kauffman. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, USA, 1993.
2. F. Herrmann, A. Groß, D. Zhou, H.A. Kestler, M. Kühl. A Boolean Model of the Cardiac Gene Regulatory Network Determining First and Second Heart Field Identity. PLoS ONE 7: e46798, 2012.
3. S. Gessert, M. Kühl. Comparative gene expression analysis and fate mapping studies suggest an early segregation of cardiogenic lineages in Xenopus laevis. Developmental Biology 334: 395–408, 2009.

Predicting the dynamic behavior of Wnt/β-catenin and Wnt/JNK signaling by a rule based probabilistic modeling approach

Alexander Groß[1], Barbara Kracher[2], Johann M. Kraus[1], Katrin Luckert[3], Oliver Pötz[3], Thomas Joos[3], Luc de Raedt[4], Michael Kühl[2], Hans A. Kestler[1]

Recent data indicates that a large number of proteins participate and interact in intracellular signal transduction forming large signaling networks. Due to the inherent complexity of such networks prediction of their behavior requires mathematical models and computational simulations. Here, we show that a static interaction network can be transformed into a semi-quantitative simulation model, which is able to reproduce the global behavior of the modeled signaling network. The model is based on the specification of probabilities for the actual occurence of so-called protein interactions including binding, enzymatic activation or phosphorylation. A set of local rules derived from experimental data and literature modifies these interaction probabilities according to the interdependencies between the different interactions. This enables the model to respond to external stimuli. Moreover, the model behavior can be observed under different conditions like knockout of network components or inhibition of specific interactions. The new rule-based probabilistic approach is able to represent dynamics of common network motifs found in signal transduction networks. We applied the presented computational method to Wnt/β-catenin and Wnt/JNK signaling. These signal transduction networks are involved in various biological processes ranging from development to aging. Our in-silico observations are in agreement with previously published findings as well as our own experimental data. These results suggest that protein-protein interaction maps augmented by local interaction rules can be a suitable means of predicting the global behavior of complex intracellular signaling networks under physiological and pathological conditions.

[1]Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany
[2]Institute for Biochemistry and Molecular Biology, Ulm University
[3]NMI Natural and Medical Sciences Institut at the University of Tübingen
[4]Department of Computer Science, Katholieke Universiteit Leuven

hans.kestler@uni-ulm.de

A critical noise level for the reconstruction of Boolean functions from time series data

Markus Maucher, Hans A. Kestler

Boolean networks are well-suited for the modeling and simulation of regulatory systems, such as gene regulatory systems. Reconstructing such networks from time series measurements can reveal the functionality of regulatory systems based on observations. While general methods for the reconstruction of Boolean networks have been devised, the modeling of biological systems brings two major challenges. First, time-resolved gene expression measurements at different stages of a cell are difficult and expensive. Therefore all reconstruction methods are faced with a relatively small number of time points compared to the number of genes. Second, measurements in biological systems are subject to noise, which can lead to bit flips after the necessary binarization.

In this work, we present an analysis of Boolean functions and the possibility to reconstruct them in the case of noisy data. We introduce the notion of the critical noise level, a function characteristic which measures the complexity of the reconstruction of a function from noisy time series data. This measure constitutes a natural upper bound for the noise probability under which a function can still be reconstructed, but can also be incorporated into the reconstruction process to improve reconstruction results. We show how to efficiently compute the critical noise level of any given Boolean function and present experimental data that shows how it can be used to improve the best-fit extension algorithm for the reconstruction of a Boolean network from noisy time series data.

References

1. S.A. Kauffman. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, 1993.
2. H. Lähdesmäki, I. Shmulevich, O. Yli-Harja. On learning gene regulatory networks under the boolean network model. Machine Learning, 52(1-2):147–167, 2003.

Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

{markus.maucher,hans.kestler}@uni-ulm.de

A Mixture of Experts Approach for the Analysis of SNP Data

Julia Schiffner, Holger Schwender

Single nucleotide polymorphism (SNP) data allow to gain insight into the molecular background of diseases. In order to find sets of SNPs that are associated with a disease normally a case-control setting is considered. A common approach is to regard the case-control status as binary response and apply a classification method combined with some dimensionality reduction technique, e.g., partial least squares (PLS). Originally, PLS has been developed for quantitative responses but can be applied to classification problems by embedding it into the logistic regression framework (see e.g. Chung and Keles (2010)). The resulting model is linear in the SNP variables, but often the relationship between SNPs and disease status is assumed to be more complex, with interactions playing an important role. Moreover, the class of diseased individuals may be heterogeneous in the sense that quite different genetic profiles can lead to an increased disease risk. For these reasons we propose a mixture of experts approach where the gating as well as all local expert models are PLS logistic regression models. This mixture model is nonlinear and takes potential heterogeneity and interactions into account. The performance of the proposed approach is assessed on simulated and real-world data and compared to standard methods.

References

1. D. Chung and S. Keles. Sparse partial least squares classification for high dimensional data. Statistical Applications in Genetics and Molecular Biology 9(1):Article 17, 2010.

Heinrich-Heine-Universität Düsseldorf

schiffner@math.uni-duesseldorf.de

Subgroup-specific survival analysis in high-dimensional datasets

Katrin Madjar, Christian Netzer, Jörg Rahnenführer

Survival analysis is a central aspect in cancer research with the aim of predicting the survival time of a patient on the basis of his features as precisely as possible. Often it can be assumed that specific links between covariables and the survival time not equally exist in the entire cohort but only in a subset of patients. The aim is to detect subgroup-specific effects by an appropriate model design. A survival model based on all patients might not find a variable that has a predictive value within one subgroup. On the other hand fitting separate models for each subgroup without considering the remaining patient data reduces the sample size. Especially in small subsets it can lead to instable results with high variance. As an alternative we propose a model that uses all patients but assigns them individual weights. The weights correspond to the probability of belonging to a certain subgroup. Patients whose features fit well to one subgroup are given higher weights in the subgroup-specific model. We define six independent non-small cell lung cancer cohorts resulting from publicly available datasets as different subgroups. We apply and evaluate our model approach using these datasets with some clinical variables and high-dimensional Affymetrix gene expression data and compare it to a survival model based on all patients as well as separate models for each subgroup.

Another aspect we investigate is whether the prediction accuracy of a survival model can be improved by combining clinical and genetic variables. It is known that some clinical variables like the tumor histology or stage are important prognostic factors and correlated with the survival time. Therefore it can be reasonable to take them into account when fitting a survival model with gene expression data.

Department of Statistics, TU Dortmund, Dortmund, Germany

rahnenfuehrer@statistik.tu-dortmund.de

Boosting the concordance index for survival data

Andreas Mayr[1], Matthias Schmid[2]

Although there exist numerous approaches for the derivation and evaluation of biomarker combinations for time-to-event outcomes, the underlying methodology often suffers from the problem that different optimization criteria are mixed during the feature selection, estimation and evaluation steps. For example, the estimation of biomarker combinations is usually based on Cox regression and is hence carried out via the optimization of a partial likelihood criterion. On the other hand, the resulting combinations are often evaluated by using the concordance index (C-index) which is a non-parametric measure to quantify the discriminatory power of a prediction rule and has its roots in the receiver operating characteristics (ROC) methodology. This methodological inconsistency is problematic from a practical point of view, as the marker combination that optimizes the partial log likelihood criterion is not necessarily the one that optimizes the C-index.
To address this issue, we propose a unified framework to derive and evaluate biomarker combinations based on the C-index. Specifically, we propose a componentwise gradient boosting algorithm that results in linear biomarker combinations that are optimal with respect to a smoothed version of the C-index. We investigate the performance of our algorithm in a large-scale simulation study and in molecular data for the prediction of survival in breast cancer patients. Our numerical results show that the new approach is not only methodologically sound but can also lead to a higher discriminatory power than traditional approaches for the derivation of gene signatures.

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

matthias.schmid@ukb.uni-bonn.de

References

1. A. Mayr, M. Schmid. Boosting the concordance index for survival data - a unified framework to derive and evaluate biomarker combinations. PLoS ONE 9(1); e84483, 2014.
2. M. Schmid, S. Potapov. A comparison of estimators to evaluate the discriminatory power of time-to-event models. Statistics in Medicine 31(23): 2588–2609, 2012.

Genomic Biomarkers for Personalised Medicine Identification and Validation

Axel Benner

A century ago Paul Ehrlich produced the first "rationally designed" drug. Since then progress of drug development evolved following his ideas, e.g. screening chemotherapeutic drugs for efficacy by assaying cytotoxicity against cancer cells first in culture, then in animals, and finally in clinical studies. During the last twenty years the targets for drug development more and more became products of aberrantly functioning genes that cause cancer. This was made possible by better knowledge about oncogenes and suppressor genes, and by the development of new tools for more accurately detecting genomic abnormalities.

A recent change in therapeutic research is denoted as "precision medicine" for better diagnosis and treatment of patients based on genomic data. It is anticipated that information from more than one molecular source will contribute to individualize treatment, including genetic aberrations, gene expression, epigenetic changes, immunological interventions and protein expression.

Computational statistics has invaded in a very short time many cutting edge research areas of molecular biomedicine. New statistical models achieved improvements with respect to statistical properties and computational efficiency, but unfortunately had more or less no impact on clinical practice. We expect that the full power of predictive and prognostic profiling will come only from integrating data from different levels into prediction models. Thus, a substantial part of our work need to be on the development of complex predictors based on data from multiple cellular levels. The availability of molecular and clinical data would then allow for developing rules for future treatment recommendations. To combine integrative statistical analysis with new concepts for marker-guided clinical trial design is strongly required.

Division of Biostatistics, German Cancer Research Center Heidelberg, Germany

benner@dkfz-heidelberg.de

Putting Statistics back into Statistical Computing

Anthony Rossini

Much recent work in statistical computing has focused on computational, mathematical, and informatics/data issues. I discuss a few limitations that exist in the common way that we use tools, and describe some features which would enable a better understanding for improving our currently "silo'd" statistical research. While all of these concepts can be built into R, some of these features are a fundamental part of the design of a new statistical system, Common Lisp Statistics, which is currently under development.

QSE / DSE / Global Development, WSJ-27.3.106, Novartis Pharma AG, CH-4002 Basel, SWITZERLAND

anthony.rossini@novartis.com

Non-identical Twins: Comparison of Frequentist and Bayesian Lasso for Cox Models

Manuela Zucknick, Maral Saadati, Axel Benner

When using high-dimensional genomic data in cancer research, the identification of prognostic factors, which can influence clinical parameters such as therapy response or survival outcome, and the evaluation of their prediction performance are some of the main issues. In these applications, the number of genome features p is usually much larger than the number of observations n ($p \gg n$ problem). Penalized likelihood methods, for example lasso regression, are often applied in this context. Frequentist lasso estimates correspond to Bayesian posterior mode estimates, when the regression parameters have independent double-exponential priors (Park and Casella, 2008), but while much attention has been paid to the frequentist lasso, less attention has been given to the Bayesian alternative. In this talk, we will investigate the lasso method in the frequentist and Bayesian frameworks in the context of Cox models through simulation studies and in an application to chronic lymphocytic leukemia. For the Bayesian lasso we extend the approach by Lee et al. (2011); in particular, we impose the lasso penalty only on the genome features, but not on relevant clinical covariates, to allow the mandatory inclusion of important established clinical factors. I will also spend some time to illustrate some of the computational aspects involved in the implementation of the Markov chain Monte Carlo sampling algorithm for posterior inference from this model.

References

1. K.H. Lee, S. Chakraborty, J. Sun. Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. The International Journal of Biostatistics, 7(1), 1–32, 2011.
2. T. Park, G. Casella. The Bayesian lasso. Journal of the American Statistical Association, 103(482), 681–686, 2008.

Division of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

m.zucknick@dkfz-heidelberg.de

Inequality-constraint Multi-class Fuzzy-in Fuzzy-out Support vector machines

Michael Glodek, Markus Kächele, Friedhelm Schwenker

In machine learning data is often afflicted with uncertainty arising from an insufficient representation of the classes in the data, but also from ambiguous class definitions which leads to deficiently annotated samples. Especially in real-world problems, it is likely to encounter noise, pattern variety and a large number of classes. Many approaches have been proposed to address these challenges. Support vector machines (SVM) embodies one of the most popular statistical learning algorithms which aims at finding the maximal margin between the separating hyperplane and the data (Schölkopf & Smola, 2002). Thiel et al. (2007) proposed a fuzzy-in fuzzy-out SVM (F^2SVM) for binary problems, in which fuzzy labels can be utilized in the training process. The output of the F^2SVM is based on the probabilistic outputs as proposed by Platt (1999). Weston and Watkins (1998) proposed extensions for multi-class SVM (MCSVM). Recently, Schwenker et al. combined these two approaches and introduced a multi-class fuzzy-in fuzzy-out SVM (MC-F^2SVM) (Schwenker et al., 2014).

The present work builds on the ideas of Schwenker et al. and proposes a novel MC-F^2SVM which incorporates the fuzzy labels using the inequality constraints instead of the objective function. The objective function of the inequality constraints based MC-F^2SVM (IC-MC-F^2SVM) is given by

$$\text{minimize} \quad \frac{1}{2}\sum_k \mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k + C\sum_n\sum_k \xi_k^{(n)} y_k^{(n)} \tag{1}$$

where y_k denotes the fuzzy label of class k, the data points are indexed by $n = 1, \ldots, N$ and $\xi_k^{(n)}$ is the slack variable. The higher the membership of a sample with respect to k, the higher the penalty of the slack variable. As a result, certain samples are more likely to

Institute of Neural Information Processing, University of Ulm, Germany

{firstname.lastname}@uni-ulm.de

be on the right side of the hyperplane. The inequality constraints capture the relations of the fuzzy labels

$$1 - \xi_k^{(n)} \leq (y_k^{(n)} - y_l^{(n)})(\mathbf{w}_k^{\mathrm{T}}\mathbf{x}^{(n)} + b_k) - (y_k^{(n)} - y_l^{(n)})(\mathbf{w}_l^{\mathrm{T}}\mathbf{x}^{(n)} + b_l) \quad (2)$$

$$0 \leq \xi_k^{(n)}. \quad (3)$$

The data points are given by $\mathbf{x}^{(n)}$ and the hyperplane is described by a weight vector $\mathbf{w}_k$ and a bias b_k. Due to the difference $(y_k^{(n)} - y_l^{(n)})$ in Equation 2, no differentiations of cases are needed. The corresponding dual form is derived by taking the derivatives of the parameter and setting the outcome equal to zero. With further substituting and rearranging the dual form is then given by

$$L(\mathbf{w}_k, b_k, \xi_k^{(n)}) = \sum_n \sum_l \sum_k \alpha_k^{(n)} \quad (4)$$

$$- \frac{1}{2} \sum_k \sum_{n_1} \sum_{n_2} \big(u_k^{(n_1)} u_k^{(n_2)} + u_k^{(n_1)} v_k^{(n_2)} \quad (5)$$

$$+ v_k^{(n_1)} u_k^{(n_2)} + v_k^{(n_1)} v_k^{(n_2)}\big)\mathbf{x}^{(n_1)\mathrm{T}}\mathbf{x}^{(n_2)} \quad (6)$$

with additional inequality constraints $0 \leq \alpha_k^{(n)} \leq C y_k^{(n)}$ and equality constraints $\sum_n \sum_l (y_k^{(n)} - y_l^{(n)})\big(\alpha_k^{(n)} + \alpha_l^{(n)}\big) = 0$.

The proposed algorithm has been compared to MC-F^2-SVM and standard approaches, i.e. one-vs-one and one-vs-rest classifiers using SVM and F^2-SVM, and outperforms them in terms of similarity while accuracy remains stable.

Acknowledgments

This paper is based on work done within the the Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems funded by the German Research Foundation (DFG).

References

1. B. Schölkopf, A.J. Smola. Learning with kernels. MIT Press, 2002.
2. C. Thiel, S. Scherer, F. Schwenker. Fuzzy-Input Fuzzy-Output One-Against-All Support Vector Machines. In: Proceedings of the International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES), Part III. pp. 156-165, Springer, 2007.
3. J.C. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schölkopf, C. Burges, A.J. Smola: Advances in Kernel Methods. pp 185-208, MIT Press, 1999.
4. J. Weston, C. Watkin. Multi-class Support Vector Machines.1998 (Tech-Report)
5. F. Schwenker, M. Frey, M. Glodek, M. Kächele, S. Meudt, M. Schels, M. Schmidt. A new multi-class fuzzy support vector machine algorithm. In: Proceedings of the International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR) (accepted).

Importance based hierarchical Lagrange multiplier filtering for the parallel training of Support Vector Machines

Markus Kächele, Friedhelm Schwenker

Support Vector Machines (SVM) are one of the most widely used classification algorithms due to their beneficial characteristics such as the maximum margin property. They have been applied to various machine learning tasks and usually rank among the top performing classifiers for a specific task. Training an SVM involves solving the convex optimization problem:

$$\max_{\alpha} \quad \sum_{i} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{1}$$

with Lagrange multipliers α_i under the constraints $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0$. $x_{i,j}$ and $y_{i,j}$ denote data points and labels, respectively. Standard quadratic programming (QP) methods can be utilized for solving Equation 1, however their complexity is too high for large datasets. Therefore, other algorithms have been developed that solve this task iteratively, such as Platt's Sequential Minimal Optimization (Platt, 1999) (SMO). Based on the decomposition scheme by Osuna et al., 1997, the algorithm selects two Lagrange multipliers, solves the QP subproblem analytically and repeats the process until convergence is achieved. The algorithm usually converges much faster than the standard method but has the drawback that there are only heuristics for the choice of the next pair of Lagrange multipliers to optimize. This leads to expensive scans through the whole dataset to find a suitable pair.

In this work, a method is presented to train an SVM based on iterative filtering by reweighting Lagrange multipliers based on their relative importance in the optimization function. The key aspects of the algorithm are that multiple instances of the reweighting schemes can be combined to build a

Institute of Neural Information Processing, Ulm University

{markus.kaechele, friedhelm.schwenker}@uni-ulm.de

filter hierarchy and that all of them can be run in parallel. Inter-layer communication and propagation of weights assure that relevant information is passed from nodes to nodes in different layers where the information is used to compute an optimization step before weight adaptations take place and the updated information is provided for the other layers. The hierarchy is set up so that each of the higher layer filters is connected to a group of lower layer filters and therefore has more information at hand. The final SVM is a result of the information that is recursively passed to the top most filter.

In various experiments different hierarchies are analysed and comparisons with other parallel SVM implementations such as Cascasde SVM (Graf et al., 2004) or Distributed SVM (Lu et al., 2008) are presented.

References

1. H.P. Graf, E. Cosatto, L. Bottou, I. Dourdanovic, V. Vapnik. Parallel support vector machines: The Cascade SVM. In NIPS, 2004.
2. Y. Lu, V. Roychowdhury, L. Vandenberghe. Distributed parallel support vector machines in strongly connected networks. Trans. Neur. Netw., 19 (7):1167–1178, 2008.
3. E. Osuna, R. Freund, F. Girosi. An improved training algorithm for support vector machines. In Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, pages 276–285, 1997.
4. J. C. Platt. Fast training of support vector machines using sequential minimal optimization, pages 185208. MIT Press, Cambridge, MA, USA, 1999.

Boolean networks

Christoph Müssel, Ludwig Lausser

Boolean networks are a popular class of models for regulatory processes in biology. These qualitative models represent genes as simple switches that can be either active (transcribed) or inactive (not transcribed). Regulatory interactions are encoded as logical statements. Different subtypes of models exist, differing in the way the networks are simulated: Classical synchronous Boolean networks update all genes at the same time, which yields a deterministic behaviour that is comparatively easy to analyze and describe. However, the assumption that all regulatory processes take the same time is unrealistic and can lead to artifacts in the simulation. Asynchronous networks overcome this problem by updating one gene at a time, which leads to a complex non-deterministic dynamic behaviour.

We propose a new model class that provides mechanisms of incorporating different time scales, but maintains a deterministic updating scheme. This is achieved by incorporating time delays and temporal predicates. Furthermore, these models provide syntactic extensions that allow for an intuitive modeling of frequent motifs, such as semi-quantitative counting statements or regulations associated with a sustained activation of upstream factors.

Tools for the simulation and analysis of such networks have recently been implemented in the BoolNet R package. The tutorial will illustrate how such models can be established and analyzed in BoolNet.

Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

{christoph.muessel, ludwig.lausser}@uni-ulm.de

Algorithm Configuration / Tuning with R

Bernd Bischl[1], Florian Schmid[2]

In virtually every subdomain of computer science and statistics, the algorithmic performance of methods can be drastically influenced by setting so-called control, strategy or hyper-parameters. Prominent examples are solvers for hard discrete optimization problems or machine learning models. For maximum gain, such a configuration must be chosen on a "per task" or per "application domain" basis, as different applications call for different configurations. We are now in an age where the general algorithm configuration problem can be nearly fully automated, by exploiting and combining efficient black-box optimization, machine learning and parallel programming.

Machine learning, and especially classification, is an important application area for such configuration techniques, although we usually call the underlying task "model selection" or "hyper-parameter tuning". For example, the number of neighbors in the case of a k-nearest neighbor classifier or the cost parameter, kernel and kernel parameters of a support vector machine are typically user-defined and can strongly influence the prognostic performance of our resulting model. They should not be chosen according to a rule of thumb, but in a sound, data-dependent way.

We present a tutorial, which gives an introduction in the functionality of the following R-packages and their usage in single- and multi-objective parameter tuning scenarios:

The TunePareto[3] package focuses on the multi-objective tuning of algorithms by scanning the parameter space. Optimizing multiple objectives, which also can be conflicting, seldom results in a single best solution. The package uses the principle of Pareto optimality to determine a set of best

[1]Institute of Statistics, Ludwig-Maximilians-University Munich, D-80539 Munich, Germany
[2]Core Unit Medical Systems Biology, Institute of Neural Information Processing, University of Ulm, D-89069 Ulm, Germany

bernd.bischl@stat.uni-muenchen.de, florian-1.schmid@uni-ulm.de

parameter configurations. To find these configurations, techniques as Latin hypercube sampling and quasi-random sequences are applied to achieve a uniform coverage of the parameter space. For scanning complex spaces an evolutionary algorithm is available in the package. Furthermore flexible interfaces for user defined classifiers and objective functions are included. The user is supported by different visualizations in making a final decision on the parameter configuration.

The mlr[1] package offers an interface to more than 50 classification, regression and survival analysis models, and most standard resampling and evaluation procedures. Models can be chained and extended with, e.g., preprocessing operations and jointly optimized. The package allows for different optimization / configuration techniques, from simple random sampling, to iterated F-racing and sequential model based optimization. The latter two are arguably among the most popular and successful approaches for algorithm configuration nowadays. Single and multi-objective model-based optimization is implemented in the mlrMBO[2] package, which learns and exploits the relation between input parameters and output performance via non-linear regression, resulting in a speedy black-box optimization method for expensive problems. It is therefore much more general than model-selection or even algorithm configuration.

References

1. B. Bischl, D. Horn, J. Bossek, J. Richter, M. Lang: mlr- MBO: Model-Based optimization and algorithm configuration. http://www.github.com/berndbischl/mlrMBO, 2014
2. B. Bischl, M. Lang, J. Richter: mlr: Machine learning in R. http://www.github.com/berndbischl/mlr, 2014
3. C. Müssel, L. Lausser, M. Maucher, H.A. Kestler: Multi-objective parameter selection for classifiers. Journal of Statistical Software 46(5), 127, 2012

A Web Application for Generating Benchmarking Data

Rainer Dangl, Friedrich Leisch

Introducing new methods of model validation in unsupervised learning requires a lot of testing. New algorithms need to be thoroughly validated before they are put to use on real world problems. Hence, benchmarking plays an important part in the development process. For this purpose, artificial data is generally used. Usually one would create a data set from scratch that should illustrate the capabilities of the new method - yet a more practical approach would be to use data utilized in previous studies (or to share the newly developed data sets with others) in order to facilitate the comparison of methods by using the same data for testing.

The talk will focus on the ongoing development of a web application that offers on the one hand the ability to download data sets that were used in previous benchmarking experiments and on the other hand the possibility to upload a newly created experimental setup. It will be possible to choose from metric scaled data, ordinal data (both of which have been implemented) and functional data (still in development).

The app uses the server environment provided by R package shiny embedded in a custom HTML interface. shiny has been developed by RStudio and emulates reactive programming paradigms in R and allows easy implementations of web applications.

References

1. T. Hothorn, F. Leisch, A. Zeileis, K. Hornik. The Design and Analysis of Benchmark Experiments. Journal of Computational and Graphical Statistics, 14(3), 675-699, 2005.

Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences, Vienna Peter-Jordan-Strasse 82, 1190 Vienna, Austria

{rainer.dangl, friedrich.leisch}@boku.ac.at

Subsampling versus bootstrap in resampling-based model selection for multivariable regression

Riccardo De Bin[1], Silke Janitza[1], Willi Sauerbrei[2], Anne-Laure Boulesteix[1]

In statistical practice, the analyst often faces the problem of choosing which variables should be included in the final model among the numerous potentially important variables collected in the study. Usually, variable selection procedures such as backward elimination, stepwise regression or all-subset approaches are used, although it is well known that they have several shortcomings, such as a high instability and a possible bias in the parameter estimates. In this context, with instability we refer to the sensitivity of a model to small changes in the data, which may modify the set of selected variables. The selection criterion, usually represented by the significance level related to a test on the parameters or information criteria such as AIC or BIC, plays a central role. In order to investigate the model stability and to provide better insight into the variable selection procedure, methods based on bootstrap resampling have been presented in the literature. By using the bootstrap technique it is possible to generate pseudo-samples which can be seen as perturbed versions of the original data. These pseudo-samples can be profitably used to identify the instability in the models obtained by a stepwise selection procedure: for example, Sauerbrei & Schumacher (1992) perform this analysis using backward elimination. The results, obtained in terms of frequency of inclusion of the variables in models derived from the pseudo-samples (inclusion frequency), allow to have a better feeling on the final model, on the importance of the different variables and on their interrelationship of being selected. Recent studies, however, have highlighted some issues related to the use of bootstrap pseudo-samples, in particular the tendency to select too many variables (see Janitza et al., 2014, for an up-to-

[1]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Germany
[2]Department of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany

boulesteix@ibe.med.uni-muenchen.de

date literature overview). Alternatives such as subsampling have been taken into consideration, and profitably applied in the context of model stability (Meinshausen & Bühlmann, 2006, 2010). The aim of our study is to provide a detailed comparison between bootstrap and subsampling in the context of model selection for multivariable regression based on inclusion frequencies, as first proposed by Gong (1982) and later extended to consider interrelationship of variable inclusion by Sauerbrei & Schumacher (1992). In particular, the use of subsampling in this framework has not been extensively investigated and contrasted with the original bootstrap approach. Here we investigate in two real data examples how the use of bootstrap and subsampling affects the model selection procedure in terms of ability of selecting one or few prominent candidate models, sparsity and prediction ability of the selected models, and ability of yielding accurate rankings of the variables based on their inclusion frequencies.

References

1. G. Gong. Some ideas on using the bootstrap in assessing model variability. In Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface. Springer New York, 1982.
2. S. Janitza, H. Binder, A.-L. Boulesteix. Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications. Tech. Rep. 163, Department of Statistics, University of Munich, 2014.
3. N. Meinshausen, P. Bühlmann. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34, 1436–1462, 2006.
4. N. Meinshausen, P. Bühlmann. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417–473, 2010.
5. W. Sauerbrei, M. Schumacher. A bootstrap resampling procedure for model building: application to the Cox regression model. Statistics in Medicine 11, 2093–2109, 1992.

Semantic clustering

Johann M. Kraus, Ludwig Lausser, Hans A. Kestler

Cluster analysis is an important technique of explorative data mining. It refers to a collection of statistical methods for learning the structure of data by solely exploring pairwise distances or similarities in feature space. However distance information can become useless as dimensionality increases. In this context, subspace clustering supports the search for meaningful clusters by including dimensionality reduction in the clustering process. We present a rapid way of preparing distance matrices for arbitrary subspaces. This method has shown to be fast enough to run standard cluster algorithms exhaustively for all feature combinations of small or medium sized datasets (Kraus et al., 2014). Using robustness analysis via resampling (Kraus et al., 2011) we are able to identify a set of stable candidate subspace cluster solutions. Based on this exhaustive clustering algorithm, we introduce a method from semantic technology to identify possible candidate subspaces that can be used for construction of new explanatory hypotheses.

References

1. J.M. Kraus, C. Müssel, G. Palm, H.A. Kestler: Multi-objective selection for collecting cluster alternatives. Computational Statistics, 26(2):341–353, 2011.
2. J.M. Kraus, L. Lausser, H.A. Kestler: Exhaustive k-nearest-neighbour subspace clustering. Journal of Statistical Computation and Simulation, 2014, DOI: 10.1080/00949655.2014.933222.

Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

{johann.kraus, ludwig.lausser, hans.kestler}@uni-ulm.de

Semantic multi-classifier systems

Ludwig Lausser, Florian Schmid, Johann Kraus, Axel Fürstberger, Hans A. Kestler

The interpretability of classification models is essential in the process of selecting biomarkers and developing diagnostic models. In high-dimensional settings, the interpretability of a model is directly related to the construction of a feature set (signature) a classifier is operating on. This signature can give hints towards the processes leading to a particular categorisation. Nevertheless, purely data driven feature selection is often affected by different forms of uncertainty and the derived signatures do not perfectly fit a given high-level interpretation. External information about the dependencies of measurements must be incorporated to increase a signatures interpretability.

In our approach, we incorporate meta-information in the training process of multi-classifier systems. This is done by training base learners on known signatures that are related to higher-level terms. By fusing these base classifiers a final prediction is made. The constructed model is interpretable in a sparse selection of terms. As the number of interpretable signatures is still high we focus on the sets related directly to the topic of interest. The selection of terms is performed by using semantic information as available in the Gene Ontology (Ashburner et al., 2000).

References

1. M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock. Gene ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29, 2000.

Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

{ludwig.lausser, florian-1.schmid, johann.kraus, axel.fuerstberger, hans.kestler}@uni-ulm.de

Fold change classifiers

Ludwig Lausser, Hans A. Kestler

Fitting a classification model to high-dimensional data typically incorporates the construction of a low dimensional feature set (signature) that reflects the major characteristics of the different categories. For many model or concept classes, this process can be seen as learning an ensemble of low dimensional base learners.

A frequently used class of base learners is the class of single threshold classifiers (rays). These base learners classify a sample according to a fixed threshold on a single feature. Prominent ensemble types that utilize single threshold classifiers are random forests, boosting ensembles or the set covering machine. The concept class of single threshold classifiers shows a high interpretability but it is unable to detect interactions of features. It remains questionable if it is the optimal choice for all kind of data.

Here, we discuss an alternative concept class of base learners which we call the concept class of fold change classifiers. The decision criterion of this concept class is based on a relative comparison of two gene expression levels of a single sample. In comparison to single threshold classifiers, the new concept class avoids the usage of globally fixed thresholds. This structural modification makes the concept class invariant to global scaling.

Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

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

An Statistical approach for Modelling of Low Frequency Oscillations in Electricity Networks

Dirk Surmann, Sebastian Krey, Uwe Ligges, Claus Weihs

Keeping up the service under all conditions is one of the major concerns in the operation of an electrical system. In a modern highly loaded electricity network with distributed energy generation from renewable energies this task is very complex. With increasing loads and frequent changes of the power feed in the control of low frequency power oscillations will become a critical aspect of the network operation. These power oscillations are a result of the network structure and the current operational setting. They can consume a large part of the network bandwidth. This makes a detailed monitoring of these oscillations necessary.

In this work we present a method to model low frequency oscillations in the highest voltage layer of an eletricity network with a system of connected mechanical harmonic oscillators. The resulting system of differential equations uses only easily measurable data from a small number of selected network nodes. This allows an easy integration in the monitoring system of the transmission system operators. We verify our very promising results by comparison to a well established and much more complex commercial simulation system used at the institute of Energy Systems, Energy Efficiency and Energy Economics of TU Dortmund University.

For the selection of the measurement nodes we rely on clustering results of the network graph into regions based on the current network topology and static information about the electrical characteristics of the network components. All these calculations are connected with the Co-Simulator of our interdisciplinary research group which allows to study the interaction of different protection and control systems under realistic conditions.

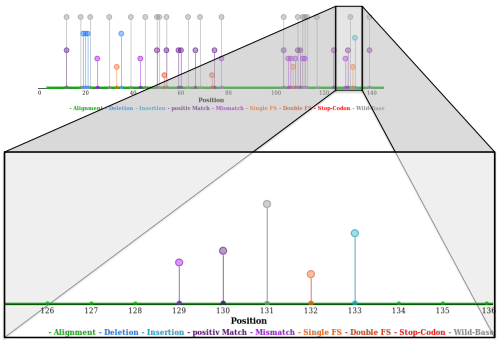
Technische Universität Dortmund, Fakultät Statistik, Vogelpothsweg 87, 44221 Dortmund, Germany

{surmann, krey, ligges, weihs}@statistik.tu-dortmund.de

Interactive zoom-able alignment graphs for pairwise wild base nucleotide protein alignments

Axel Fürstberger, Hans A. Kestler

Sequence alignment is a widely used tool for the analysis of sequencing data. The text output of alignment algorithms can be transformed into graphical representations. We developed a new tool ≫ ZAG ≪ to generate interactive zoom-able alignment graphs for pairwise alignment of wild base nucleotide sequences and protein sequences based on output of the SWAT algorithm and the dygraphs framework.
With a special focus on wild base nucleotide sequences and protein sequence alignment, this representation gives a plain and clear overview of the alignment and allows a quick summery of the data. It also offers the possibility to zoom into regions of interest and take a closer look at the aligned subsequence and specific positions.



Core Unit Medical Systems Biology, Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

{axel.fuerstberger, hans.kestler}@uni-ulm.de

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

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

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

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

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

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

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

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

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

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

96-06 *Manindra Agrawal, Thomas Thierauf*
Satisfiability Problems

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

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

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

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

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

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

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

97-01 *Jochen Messner*
Pattern Matching in Trace Monoids

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

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

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

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

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

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

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

97-09 *Christian Heinlein*
Grundlagen von Interaktionsausdrücken

97-10 *Christian Heinlein*
Graphische Repräsentation von Interaktionsausdrücken

97-11 *Christian Heinlein*
Sprachtheoretische Semantik von Interaktionsausdrücken

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

99-03 *Uwe Schöning*
On the Complexity of Constraint Satisfaction

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

99-05 *Johannes Köbler, Wolfgang Lindner, Rainer Schuler*
Derandomizing RP if Boolean Circuits are not Learnable

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

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

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

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

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

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

2000-04 *Wolfgang Gehring*
Ein Rahmenwerk zur Einführung von Leistungspunktsystemen

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

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

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

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

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

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

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

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

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

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

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

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

2002-02 *Walter Guttmann*
Deriving an Applicative Heapsort Algorithm

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

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

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

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

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

2003-05 *Christian Heinlein*
Dynamic Class Methods in Java

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

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

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

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

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

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

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

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

2005-03 *Walter Guttmann, Markus Maucher*
Constrained Ordering

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

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

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

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

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

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

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

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

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

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

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

2008-08 *Henning Wunderlich*
Covers have structure

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

2008-10 *Henning Wunderlich*
On span-P^{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 Betz, Frank Raiser, Thom Frühwirth*
A Complete and Terminating Execution Model for Constraint Handling Rules

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

2010-03 *Patrick Frey*
Case Study: Engine Control Application

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2013-06 *Andreas Steigmiller, Birte Glimm*
Nominal Schema Absorption

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

2013-08 *Daniel Ott, Dr. Alexander Raschke*
Evaluating Benefits of Requirement Categorization in Natural Language Specifications for Review Improvements

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

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

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

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

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

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm