technische universität
dortmund

SFB 876

Technical report for Collaborative Research Center SFB 876

Providing Information by Resource-Constrained Data Analysis

Dezember 2014

Part of the work on this report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis"

Speaker: Prof. Dr. Katharina Morik
Address: Technische Universität Dortmund
Fachbereich Informatik
Lehrstuhl für Künstliche Intelligenz, LS VIII
D-44221 Dortmund

Inhaltsverzeichnis

SFB 876 Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Olaf Spinczyk

A New Look at Regularization for Probabilistic Graphical Models

Nico Piatkowski
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
nico.piatkowski@tu-dortmund.de

Probabilistic graphical models can simulate and classify high dimensional, heterogeneous data and serve as underlying formalism in the data analysis of biology, physics, computer vision, natural language processing and others. Their parameter dimension is a function of the treewidth of the data's conditional independence graph and the data domain. Even if most dependencies are ignored and a pairwise approximation is considered, the dimension remains large, e.g., it exceeds millions. Hence, the sample complexity is large, models tend to overfit the data and the learned models are hard to communicate. Classic ℓ_1-regularization approaches can not be applied without changing the underlying graphical structure, which is not an option if certain dependencies should be handled explicitly by the model. A recent advance in this area has shown, that a combination of regularization and reparametrization can lead to sparse models for spatio-temporal data. This is achieved by explicitly encoding knowledge about the spatio-temporal nature of the data into the model parameters. We discuss related approaches and how these techniques can be generalized to other data domains. Furthermore, we present a first promising experimental result.

The exponential family of densities arises naturally as the maximum entropy distribution that reproduces data's empirical marginals. We consider a discrete[1] (multivariate) random variable $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ that takes realizations $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ from the discrete set[2] $\mathcal{X} = \bigotimes_{i=1}^{n} \mathcal{X}_i$, with $\boldsymbol{x}_i \in \mathcal{X}_i$. The probability mass function $p_{\boldsymbol{\theta}} : \mathcal{X} \to (0, 1)$

[1] Restricting ourselves to discrete models is just for notational convenience. The ideas presented here do apply to models with continuous variables as well.
[2] The $\bigotimes$-symbol denotes the Cartesian product.

of $\boldsymbol{X}$ is given by $p_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = exp\left(\langle \boldsymbol{\theta}, \phi_G(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right)$, whereas $A : \mathbb{R}^d \to \mathbb{R}$ is the log-partition function, $\boldsymbol{\theta} \in \mathbb{R}^d$ is a d-dimensional parameter vector and $\phi_G : \mathcal{X} \to \mathbb{R}^d$ maps realizations $\boldsymbol{x}$ to a d-dimensional feature vector. The map ϕ_G is fully determined by a graph $G = (V, E)$ that encodes conditional independence between the components of $\boldsymbol{X}$ (see [8]). In the most general case, the dimension d is a multivariate polynomial in the variables state space sizes, i.e., $d = \sum_{C^* \in \mathcal{C}(G)} \prod_{v \in C} |\mathcal{X}_v|$, where $\mathcal{C}^*(G)$ is the set of maximal cliques (fully connected subgraphs). In practice, often only the vertices and edges of G are modeled. Nevertheless, large state spaces or large graphs will still lead to millions of parameters. Classic ℓ_1-regularization [4] can not be applied directly, since the estimation of the graph and the estimation of the parameters are coupled: regularization based graph estimation approaches (e.g. [3]) perform maximum a posteriori parameter estimation, whereas a Laplace prior pushes small parameter values towards 0. The remaining non-zero weights represent the estimated graph G. Obviously, a subsequent estimation of model parameters with ℓ_1-regularization can not detect additional sparsity, since the 0s were *already found* by the step that estimated the graph. Hence, new methods are required to remove redundancies and to find sparse models.

Related work. In certain special cases, the dimension can be mostly reduced, i.e., simple Ising models from computer vision and physics have basically one single parameter for arbitrary large graphs [8]. The repetitive structure of linear chain conditional random fields allows the use of so-called factor templates [7] which do also decouple the dimension from the graph size. Both approaches are task-specific and hard-coded into the model. In the θ-Markov random field (θ-MRF) [2], parameters of correlated classification tasks are connected via spatial and semantic edges to build a graphical model over the parameters. However, the model is formulated for classification tasks on image (or at least spatial) data and not for generative models or arbitrary data domains. The general concept of instantiating conditional independence assertions on the parameter space is called hyper-Markov models [1]. A recent approach, the spatio-temporal random field (STRF) [5], is based on the temporal dependence between parameters and finds a sparse parametrization if the data has a temporal dimension. Nevertheless, inference is slightly more complex in such models and the same technique can not be applied if the data contains no temporal dimension. Overall, each of the aforementioned approaches assumes the existence of some special prior on the parameters, which is hard-coded into the respective approach. This raises the question, if methods exists that detect such latent structures automatically and independent of the underlying data domain. In what follows, we present a first step into this direction.

Ex post model quantization. The basic idea is to aggregate similar model parameters—such procedures are also known as *quantization*. Algorithm 1 performs a quantization of MRF model parameters by a 1-dimensional k-means clustering of the maximum-likelihood estimate $\boldsymbol{\theta}^*$ with a subsequent replacement of each parameter by the nearest cluster

Algorithm 1: MRF with ex post model quantization

input : Data set $\mathcal{D}$, graph G, $k \in \mathbb{N}$
output: Quantized parameters $\boldsymbol{\theta}^{**}$

$\tilde{\boldsymbol{\mu}} \leftarrow \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \phi_G(\boldsymbol{x})$; // compute sufficient statistic from data
$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^d} \langle \boldsymbol{\theta}, \tilde{\boldsymbol{\mu}} \rangle - A(\boldsymbol{\theta})$; // estimate MRF maximum-likelihood parameters
$\boldsymbol{\rho}^* \leftarrow \arg\min_{\boldsymbol{\rho} \subset \mathbb{R}, |\boldsymbol{\rho}|=k} \sum_{i=1}^{d} \min_{r \in \boldsymbol{\rho}} (\boldsymbol{\theta}_i^* - r)^2$; // k-means on $\boldsymbol{\theta}^*$
foreach $i \in \{1, 2 \ldots, d\}$ **do**
    $\boldsymbol{\theta}_i^{**} \leftarrow \arg\min_{r \in \boldsymbol{\rho}^*} (\boldsymbol{\theta}_i^* - r)^2$; // replace each parameter by the nearest center
return $\boldsymbol{\theta}^{**}$;

center. This procedure is similar to assuming a Gaussian mixture prior over the parameters. The quantized parameters $\boldsymbol{\theta}^{**}$ are no longer a maximum-likelihood estimate. However, it can be shown that the *model quantization error* $|\hat{\ell}(\boldsymbol{\theta}^{**}) - \hat{\ell}(\boldsymbol{\theta}^*)|$ is bounded by $\|\boldsymbol{\theta}^{**} - \boldsymbol{\theta}^*\|_2(\|\tilde{\boldsymbol{\mu}}\|_2 + \sqrt{|\mathcal{C}^*(G)|})$ with $\hat{\ell}(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \tilde{\boldsymbol{\mu}} \rangle - A(\boldsymbol{\theta})$.

The algorithm shares some important properties with STRF and θ-MRF: small differences in the parameters are removed (regularized) and hence, certain parts of the model are forced to use the same parameters. Notice that cliques with exactly the same parameters will not necessarily induce the same marginal probabilities on the corresponding (sets of) variables, since those are also influenced by the graphical structure. Furthermore, the number of different parameters can not grow with the size of the graph, since it is bounded to k. This might lead to a very memory efficient representation if the mapping from the index set $i \in \{1, 2 \ldots, d\}$ to the k cluster centers can be stored efficiently.

Evaluation. We evaluate the simple model quantization (Alg. 1) on the Iris[3] data set. Each of the attributes is discretized into 6 bins of equal width. We consider a pairwise model based on the fully connected graph K_5. Parameter estimation is done via gradient descent with constant stepsize 0.1 for 100 iterations. In our experiments, loopy belief propagation never diverged on this graph. Table 1 shows, for increasing k, leave-one-out cross validated estimates of the negative average log-likelihood on training set $-\hat{\ell}(\boldsymbol{\theta}^{**})$ and test instance $-\hat{\ell}(\boldsymbol{\theta}^{**})_{test}$, the model quantization error on the training set and the corresponding error bound (see above), and the prediction accuracy on the test instance. The unquantized model has a leave-one-out cross validated negative avg. log-likelihood of 4.20259 ± 0.01426 on the training set, 4.48299 ± 2.06178 on the test set, and an accuracy of 0.94667 ± 0.22545. For the sake of reproducibility, the k-means are initialized deterministically such that $\boldsymbol{\rho}_i = (i-1) \cdot d/k$, $1 \leq i \leq k$. Similar results have been achieved with integer graphical models [6] where the parameter set is restricted to $\{0, 1, \ldots, K\}$, which can also be interpreted as model quantization. It is subject to future research to

[3] Iris at the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/Iris`

Table 1: Leave-one-out cross validated results of Alg. 1 on the Iris data for several k.

k	2	4	8	16	32	64	128
$\hat{\ell}(\boldsymbol{\theta}^{**})$	±0.01735 5.38057	±0.04181 4.63347	±0.01949 4.43078	±0.01542 4.38101	±0.01649 4.37721	±0.06001 4.29104	±0.03605 4.22685
$\hat{\ell}(\boldsymbol{\theta}^{**})_{test}$	±1.45645 5.60374	±1.77172 4.86885	±1.76937 4.66813	±1.78936 4.58627	±1.75879 4.64079	±1.83471 4.63676	±1.97952 4.49653
$\lvert\hat{\ell}(\boldsymbol{\theta}^{**}) - \hat{\ell}(\boldsymbol{\theta}^{*})\rvert$	±0.01767 1.20086	±0.04592 0.45376	±0.02083 0.25106	±0.01444 0.20129	±0.01719 0.19749	±0.05745 0.11159	±0.03869 0.04810
Bound	±0.08451 17.88793	±0.23846 10.70129	±0.06983 5.79908	±0.15298 4.02000	±0.07494 3.45896	±0.45273 2.46291	±0.03327 1.44740
Test acc.	±0.25028 0.93333	±0.19662 0.96000	±0.21163 0.95333	±0.21163 0.95333	±0.21163 0.95333	±0.21163 0.95333	±0.22545 0.94667

find better bounds, to integrate the clustering directly into the training procedure and to speed up inference with quantized models. Notice that the very same approach can be applied to other parametric models like support vector machines as well.

References

[1] A. Philip Dawid and Steffen L. Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.

[2] Congcong Li, Ashutosh Saxena, and Tsuhan Chen. θ-mrf: Capturing spatial and semantic structure in the parameters for scene understanding. In *Advances in Neural Information Processing Systems 24*, pages 549–557, 2011.

[3] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[4] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML, New York, NY, USA, 2004.

[5] Nico Piatkowski, Sangkyun Lee, and Katharina Morik. Spatio-temporal random fields: compressible representation and distributed estimation. *Machine Learning*, 93(1):115–139, 2013.

[6] Nico Piatkowski, Sangkyun Lee, and Katharina Morik. The integer approximation of undirected graphical models. In *ICPRAM 2014 - Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*, pages 296–304, 2014.

[7] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[8] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Knowledge Transfer to Reduce Labelling Effort

Christian Pölitz
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
christian.poelitz@tu-dortmund.de

An expensive resource in applying Machine Learning methods are labelled data for the training of a model. For classification models for instance like Support Vector Machine, we need labelled data to learn a decision function that can label new data in the same way as the training data. To reduce this effort we try to leverage labelled training data from slightly different tasks. We concentrate on classifications of small texts to possible meanings. For instance product reviews can be in a positive of negative tone. In case we want to learn a classification models that can decide whether a new review is positive or negative we need training data of reviews that have already been labelled.

In order to avoid new labellings we try to reuse labelled data that are a bit different. Such data could be for instance reviews about different product as the ones we want to use for prediction. But the usual assumption for most of the Machine Learning tasks is that the training data used to learn a model has the same distribution as the test data on which the model is applied. In order to overcome this issue we try to find data representations that are invariant across data sources.

We assume to have two data sets with (possible large) difference in distribution. On the one hand, we have data from a source domain S that is distributed via p_s together with label information y distributed via $p_s(y|x)$. On the other hand, we also have data from a target domain T that is distributed via p_t with no label information. We define a transfer learning task as to use the source domain together with its label information to find a classifier that labels the target domain best.

From bounds on the expected error on a target domain T using only training data from a given source domain S, we learn that for transfer learning to be successful, we need at

least two things. First, the probability distributions of the two domains must be similar. Hence, $D(p_s, p_t)$ is small, for D any measure of discrepancy of distributions. And second, the difference of the hypothesis from both domain must be small. This can be directly read from the following bound by Ben-David et al. [1] from Theorem 1 page 155:

$$\epsilon_t(h) \leq \epsilon_s(h) + D(p_s, p_t) \tag{1}$$

$$+ \min \left(\int |h_s(x) - h_t(x)| p_s(x) dx, \int |h_s(x) - h_t(x)| p_t(x) dx \right) \tag{2}$$

The bound tells, that the expected error ϵ_t of any hypothesis h on the target domain can be bounded by the expected error ϵ_s on the same hypothesis trained on the source domain, the difference in distribution of target and source domain, and the expected difference of any two hypothesis h_s and h_t from the source and target domain. Our goal for transfer learning now will be to minimize this bound by finding a suitable data representation.

Based on several observations many approaches have been proposed to find a proper representation of the data from both domains to account for a good domain adaptation or transfer of knowledge.

One line of research tries to find such feature representations of different data sets that are invariant for both data set distributions. Low dimensional feature representations are used to capture this invariances. These representations are for instance extracted via dimension reduction methods like (kernel) PCA.

In order to find the low dimensional representation for the transfer of the knowledge of the labels, we find projection operators that map all data onto a suitable subspace. This projection operator P is defined as $P : H \to H$ with H the feature space of the data. In case H is the Euclidean space the operator is a matrix such that $P^T \cdot P = I$. In case H is a infinite dimensional Hilbert space the operator is defined as expansion of the data in the feature space.

The operator is found by minimizing the discrepancy between the source and target domain. Formally we define the optimization task as:

$$\min_{P \in H} D(p_s(P \cdot x), p_t(P \cdot x)) \tag{3}$$

$$s.t. P^T \cdot P = I$$

In case the data lies in a Euclidean space we perform this optimization directly on the set of all possible projection matrices. If the data lies in an infinite dimensional Hilbert space the projection P maps onto the data points. This means $P \cdot x = a \cdot X$ for X the

being the data matrix $[x_1, \cdot, x_n]$. Since the x_i are infinite dimensional elements of a Hilbert space we cannot explicitly use them. This is why we have to assume H is a Reproducing Kernel Hilbert Space (RKHS) with a Kernel K such that $K = X \cdot X^T$. Now we need to find A such that $D(p_s(A \cdot K^{\frac{1}{2}}), p_t(A \cdot K^{\frac{1}{2}}))$ is minimized.

In [2] we propose to minimize D by mapping onto the principle components of the covariance operator of all data (source and target domain). By this we effectively reduce the variance of the data by projecting onto a subspace where most of the data is concentrated. This also reduces the discrepancy between the distributions. In this case, the projection operator P is found via Kernel PCA as proposed by [4]. Using the standard scalar product, the same method can be used in the Euclidean space too.

In the previous approach, we simply used all avail source data that to estimate the projection operator and all labels to train the classifier. An interesting question with respect to resource limitation is whether we can find the best samples from the source domain to transfer knowledge. Although we expect to have access to large amounts of source data, by sampling the potential most useful ones for the knowledge transfer, we might be able to converge faster and reduce computational effort.

We investigate two ways of sampling from the source domain. First, we propose strategies to reduce the amount of computation and storage by greedily selecting data samples from the sources to efficiently find the invariant subspace. In the paper [2] we propose to use distance based sampling strategies. This results in reducing the number of source data points to estimate the invariant subspace. Second, we also propose to sample the labels from the source domain that are most promising for the knowledge transfer. In [3] we combine active sampling strategies with transfer learning to identify good source domain samples. We argue that the distance of the texts is a good measure of how appropriated texts from different domains are for the training of a classifier for a certain target domain. We calculate the distance of texts to domains as the distance to a latent subspace of the corresponding target domain. Finally, we define an active sampling strategy that integrates this distance measure to choose potentially useful texts from different domains for the training of an SVM that is applied to a target domain where no label information are available.

In the future we plan to extend the subspace and train data sampling to multi feature spaces. In a multi kernel learning approach we find the optimal combination of kernels respectively subspaces, to align optimally between the kernels and source labels.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010.

[2] Christian Poelitz. Projection based transfer learning. In *Workshops at ECML 2014 (to be published)*, Nancy, France, 2014.

[3] Christian Pölitz. Distance based active learning for domain adaptation. In *ICPRAM 2015 - Proceedings of the 4rd International Conference on Pattern Recognition Applications and Methods (to be published)*.

[4] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Advances in kernel methods. chapter Kernel Principal Component Analysis, pages 327–352. MIT Press, Cambridge, MA, USA, 1999.

Declarative Event-Based Data Acquisition in Ubiquitous System Software

Jochen Streicher
Lehrstuhl für Informatik 12
Technische Universität Dortmund
jochen.streicher@tu-dortmund.de

Automatically optimizing data acquisition tasks for mobile devices requires them to be automatically analyzable. Existing data collection infrastructures either lack this requirement, or are too restricted, especially regarding deep OS data. The approach presented here combines functional power with simplicity and analyzability. While the genericity has its cost, leaving room for further optimization, the first results are promising.

1 Problem

Data collection from mobile devices, especially from their system software, is subject to special requirements regarding energy and privacy. Our goal is to automatically analyze, optimize and fuse data acquisition tasks with respect to these requirements. Existing flexible data acquisition infrastructures like Funf [2] are either too restricted regarding their configurability (e.g., to (de-)activation of data sources or their sampling frequency), while others, like SystemTap or our own infrastructure *MobiDAC*, offer a greater degree of flexibility, including data preprocessing, but are not amenable to the transformation of data collection tasks according to our goals. Thus, an approach is necessary that combines flexibility with analyzability. A first step into this direction was our data modeling approach that captures the existing data sources in an extensible way. [4] It serves as a basis to provide data acquisition tasks with semantics, allowing to chose concrete data providers according to energy and quality properties. The second step is a declarative and thus analyzable query language.

```
Select(
       Filter(
              Join(
                  GenStream(net.device["eth0"].onRx),
                      processes.process.tcomm,
                      Stream.net.onRx.socket == Join.processes.process.sockets
              ), processes.process.tcomm == "firefox"
       ), net.onRx.size
)
```

Listing 1: Example of a query that keeps track of the sizes of network packets that are received by all firefox processes from a specific NIC.

2 Approach

The query language is based on Aurora. [1] It allows to create streams of data tuples from permanently available data *sources, events*, which occur and deliver context information only at specific points in time, and the more complex operating system *objects*, which provide relational data. These may be subject to various operations already known from relational databases and data stream management systems, namely *filtering* of whole tuples and *projection*. Furthermore, relational data from *objects* may be *joined* to the data stream. Listing 1 shows a query that creates an *event* stream from received network packets, *join*s the receiving process *object* via the matching socket, *filter*s for all "firefox" processes and *select*s the packet size.

The system accepts and processes queries from clients that may reside in kernel as well as in user space and delivers the generated data back to the client. The actual acquisition of data from events and data structures is the job of *data providers*, which are assigned to the elements of the data model. If a query is received from a client, the contained data model elements are identified and the respective providers are found and set up for data acquisition. Providers may, again, reside in kernel or user space. An example for user space system software data is the *foreground application* on Android-based devices. Providers may be registered at runtime, thereby extending the data model. The data delivered by providers is not restricted to already existing data structures and events. They may also act as clients, and require data from other providers to generate data for the model elements, for which they registered. Queries may use data from both kernel and user space. Since both clients and providers may reside in kernel or user space, the core components exist, virtually symmetrically, both in the kernel and in the respective process's address space. The details of synchronization and communication between kernel and user space components, however, are beyond the scope of this document.

Scenario	Load		duration [s]			latency [µs]
			Baseline	SystemTap	Declarative Approach	
1	SysBench	avg	780	780	780	48.9
		stddev	.0742	.0938	0.075	222
	Linux build	avg	1140	1140	1150	25.4
		stddev	5.36	3.79	3.27	11.7
2	SysBench	avg	809	812	847	13.4
		stddev	0.111	0.561	5.60	5.04
	Linux build	avg	1250	1300	1439	12.2
		stddev	2.85	4.31	4.10	5.13

Table 1: Preliminary quantitative evaluation results. The duration is based on 30 runs for each avg/stddev pair. The latencies are collected from one run for each avg/stddev pair.

3 Results

The declarative approach was evaluated regarding latency and overhead of query execution and compared to *SystemTap* using two different data acquisition tasks: 1.) Capture incoming network packets, *filter* for a specific source IP address, and *select* their size and 2.) on process termination, *join* the processes *object* and select the time spent in kernel mode. While the latency is not important for pure data acquisition, it comes into play when the acquired data is directly used for system adaptation upon critical events. The evaluation platform was a desktop machine with an Intel Core i5 processor and Ubuntu Server 14.04. The clock frequencies of the four cores were fixed at 3.4GHz each. The vanilla Linux kernel 3.14.17[1] was used for the implementation. Different load scenarios were used, both without (baseline) and with the described data acquisition scenarios. The load scenarios consist of SysBench[2] (CPU and user mode only) and a full Linux kernel build (same as above, CPU and I/O activity). For the first scenario, a second machine connected with a direct Gigabit Ethernet connection sent TCP packets at full speed.

4 Conclusion

While the approach is promising, there are still necessary features missing, like multiple providers for data sources, aggregates, or splitting and combining streams to allow query fusion. We plan to integrate an existing DSMS to build on for this purpose. Query optimization, especially with regard to minimizing the amount of address-space-crossing

[1] configured with the Debian 3.14-0.bpo.2 standard configuration and default options where necessary
[2] http://sysbench.sourceforge.net

data, should lead to better results. Further optimization potential mostly concerns consistency and relational joins: If a join predicate does not resolve to a single object ID (i.e., the primary key of a database row), a full join has to iterate over all instances, while keeping the according lock. An *index,* the database way to accelerate this, might not make sense for the frequently updated kernel data structures. Processing several tuples with one snapshot of the kernel state is probably a better idea if one can live with the impact on *consistency*. This is either possible by caching a relational snapshot for a certain amount of time. Caching the stream tuples instead might be even more efficient, but increases *latency.* Regarding kernel state consistency, we are anyway faced with the same problems as a pure relational interface to kernel data. [3] Operating system data structures are actively used and constantly changed. Ensuring the consistency that is expected from databases would cause severe contention and hamper the operating system's normal activities. In most cases, the consistency or invalid references are not even a problem. Sometimes we might not even *need* the consistency of a *single* data structure. If this case it is a bad idea to hamper normal OS operation and increase contention by ensuring it. For example, joining the values of one or more process objects to a data stream does not necessarily require the protection of the whole process list. In this case, the consistency requirements may already be inferred from the query structure. In other cases, the query language should allow to express the necessary level of consistency.

References

[1] Daniel J Abadi, Don Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stan Zdonik. Aurora: a new model and architecture for data stream management. *The VLDB Journal—The International Journal on Very Large Data Bases*, 12(2):120–139, 2003.

[2] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643 – 659, 2011. The 9th Annual IEEE Intern. Conference on Pervasive Computing and Communications (PerCom 2011).

[3] Marios Fragkoulis, Diomidis Spinellis, Panos Louridas, and Angelos Bilas. Relational access to unix kernel data structures. In *Proceedings of the Ninth European Conference on Computer Systems*, EuroSys '14, pages 12:1–12:14, New York, NY, USA, 2014. ACM.

[4] Jochen Streicher. Data modeling of ubiquitous system software. Technical Report 7, TU Dortmund University, 8 2014.

Subproject A2
Algorithmic aspects of learning methods in embedded systems

Christian Sohler Jan Vahrenhold

Speed-Up Techniques for Orthogonal Range Clustering

Ebrahim Ehsanfar
Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
ebrahim.ehsanfar@tu-dortmund.de

1 Abstract

We present a fast techniques for orthogonal range clustering using summary-based data structures. Given a range space $(X, \Re)$, and a set, $P \in X$, of n points in d-space, the aim of orthogonal range clustering is to find a k-clustering of the points within any arbitrary range $Q \in \Re$. We expedite this process by pre-storing some data into a data structure. This enables us to compute a summary of range in the query time. To obtain the summary, we add only a small portion of the data to each subtree of the data structures. We show that the cost of the clustering based on the summary is only slightly higher than that of of the clustering of the original records.

2 Introduction

Responding to the user's queries is the primary function of a database system. Generally, queries can be divided into *reporting* and *aggregation* queries. Reporting queries involve collecting records that satisfy query conditions. Here, the database is only used as a data retrieval tool. With *big data*, the result of such queries could contain a huge number of records. Therefore, even the simple task of counting records can be very time consuming. In contrast, aggregation queries produce an aggregate result for a particular feature of data. Often statistical information about the data (e.g., records mean or distribution) is more interesting than the actual records. Particularly, when the size of data is very large, special aggregation queries can give a useful summary of data. The advantage

of summaries is that they demonstrate the main properties of the original data, but are much smaller in size.

In this paper, we focus on summaries that are generated by queries that involve optimizing an objective function with respect to a similarity metric (e.g., Euclidean distance). To illustrate such queries further, consider the following example. Suppose that using a large database of temperature records one wants to find a set of locations across the world that are optimal for collecting global warming data. These locations must have the shortest distance to critical areas, i.e., places where the average annual temperature is below 0 degrees. Therefore, the following query is posed to the database:

(Q): What is a set of locations for collecting global warming data such that each location has the shortest distance to any place whose average annual temperature is below 0 degrees?

The standard way of executing the above query is to first find all records and then perform an optimization to find the locations that satisfy the constraints. However, since in large databases processing such queries is computationally intensive, a solution that has a low time complexity is highly desirable. To date, no study has specifically investigated this problem. The most related works are explained in the next section.

To address this shortcoming in the literature, we propose an approach that using an approximation technique provides a near-optimal solution with a sublinear time complexity. More specifically, we present a method called *orthogonal range clustering* that, instead of evaluating all records, clusters them with respect to an orthogonal range. Given an orthogonal range, the goal of this method is to return, instead of the entire query result, only a small subset of the records that represent a summary of the data. Moreover, we propose that by embedding these summaries into a database index, aggregation queries can be answered efficiently at query time with a time complexity proportional to the size of the summary itself, not the size of the naive query results. We discuss the problem more formally as follow:

Let D be a database containing N records. Each record $p \in D$ is associated with a query attribute $A_q(p)$ and a summary attribute $A_s(p)$. A summary query identifies a range constraint $[q_1, q_2]$ and the the goal is to return a summary on the A_s attribute of all records whose A_q attribute is within the range. For example, in the query (Q) above, A_q is coordinate of the points and A_s is temperature. Note that A_s and A_q could be the same attribute. Our objective is to build an index on D such that a clustering summary query can be answered efficiently. like any other indexing problems, the primary measures are the size of space that the structure uses and also the query time.

Going back to the standard solution, one can always retrieve the entire $P(q) = P \cap q$, and then find the cluster centers C from $P(q)$ using an standard clustering algorithm e.g. k-means. Throughout this paper, we use RAM computational model. Specifically, our contribution to the state-of-the-art of query processing research is as follows:

- Given a set $P \subset \mathbb{R}$, we construct a linear-size data structure for P with a query time of $O(\frac{k}{\epsilon})$. Given a range q, this structure computes a $O(k)$ size summary for clustering the records in $P \cap q$.
- We propose a method for generating data summaries for cluster range queries in multi-dimensional spaces, using data structures tailored for orthogonal range clustering. These data structures, in the best case, have a size of $O(n + n\log(\frac{k}{\delta\epsilon}))$ and the query time of $O(n^{1-\frac{1}{d}} + n\log(\frac{k}{\delta\epsilon}))$.
- We use several data sets including Berkeley Earth (with more that 32,000,000 records in 4 dimensions) to empirically evaluate our range clustering method and compare the results with the existing standard method. In doing so, we use several linear-size (i.e., $O(k \cdot n)$) data structures for multi-dimensional clustering range queries that approximate clustering quality with a significantly high accuracy. We show that with a very small summaries for each subtree, (e.g. even of size $k = 5$), the results are significantly reasonable.

3 Related works

Working on orthogonal range clustering is significantly related to the indexing for aggregation queries. Generally in this regard, most aggregates can be performed simply using a binary tree in one dimension. It is enough to store the aggregate of all records in the associate subtree for any internal node. Using this structure, an aggregation query can be answered in $O(\log n)$ time [4].

In higher dimensions, the problem has been studied in computational geometry and database systems areas [4] [1] [5]. Most of those solutions are based on space-partitioning hierarchies e.g. partition trees where an internal node stores the aggregate for its subtree .

4 Our Method

The objective of our method is to speed up orthogonal range clustering. To this end, we add coreset summaries to range searching data structures such that, when performing a search, the summaries are reached before the individual data records. Doing so enables us to avoid traversing the entire data structure to answer queries, and thus, the query time reduces. In what follows, we explain in more detail how this can be done.

It is known that clustering corsets are mergeable [3]. That is, given two coresets of two disjoint data sets, the union of the coresets is a coreset of the union of the data sets. Using this decomposability property, one can partition a set into different subsets,

compute a coreset for each subset in parallel, and then union the resulting coresets which would yield a good summary of the original data. Given this, in order to solve a range query, we represent the data as a collection of disjoint canonical subsets, and to each of them assign a coreset. Then, the returned result is the summary of the data within the range, which is the union of the coresets corresponding to the canonical subsets within the range.

More formally, suppose that we construct a binary tree (or any other range searching tree) to represent the data T, where each leaf contains a single record or a set of records denoted as s. For each internal node u of T, let T_u denote the subtree of T rooted at u. We attach to u an $(k, \epsilon) - coreset$ of the A_s attribute of all records stored in the subtree below u. Since each $(k, \epsilon) - coreset$ has size s_ϵ, and the number of internal nodes is $O(N)$, the total size of the structure is $O(N \cdot s_\epsilon)$. To answer a query $[q_1, q_2]$, we perform a search on T. It is well known that any range $[q_1, q_2]$ can be decomposed into $O(log(N))$ disjoint canonical subtrees T_u [2]. Therefore, we decompose the query into $O(\log N)$ canonical subtree, and retrieve the $(k, \epsilon) - coreset$ attached to the root of each of them. Then, we return the union of the retrieved results, which is an $(k, O(\epsilon)) - coreset$ of the entire in-range data. The total query time is thus the time required to union the $O(log(N))$ coresets. The returned summary can be used as a reasonably good representative of the data for k-clustering problem.

References

[1] Martin Burger, Vincenzo Capasso, and Daniela Morale. On an aggregation model with long and short range interactions. *Nonlinear Analysis: Real World Applications*, 8(3):939–958, 2007.

[2] M. De Berg, O. Cheong, and M. van Kreveld. *Computational geometry: algorithms and applications*. Springer, 2008.

[3] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300. ACM.

[4] Yufei Tao, Cheng Sheng, C Chung, and J Lee. Range aggregation with set selection. 2013.

[5] Jeffrey Scott Vitter and Min Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. In *ACM SIGMOD Record*, volume 28, pages 193–204. ACM, 1999.

On clustering time series under the Fréchet distance

Amer Krivošija
Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
amer.krivosija@tu-dortmund.de

The problem of clustering of the time series under the Fréchet distance is studied, as the Fréchet distance captures well the similarities between the curves and does not depend on their complexity. The k-center and k-median problems are studied and the first $(1+\varepsilon)$-approximation algorithms for them are given. Their running time is almost linear in terms of the input size. These results will be published in [6].

Introduction

A *time series* is a recording of a signal that changes over time. It represents a sequence of discrete measurements of a continious signal. The signal/data can be multidimensional, but we limit ourselves to the univariate case. Examples of such data are stock market values, temperature and tide measures, number of queries on search engines, etc.

Formally, it is a series $s_1, \ldots, s_m$ of measurements of a signal $[0,1] \to \mathbb{R}$, assumed that the time scale is $[0,1]$, what can be easily extended to any other domain, and that the measurements were done in the timely order. A time series can be also observed as mapping $\tau : [0,1] \to \mathbb{R}$ by linearly interpolating $s_1, \ldots, s_m$ in temporal order. We call τ a curve of complexity m, and (s_i, t_i), $i \in \{1, \ldots m\}$ its vertices, where s_i is the value of the measurement, and t_i is its time stamp.

The clustering is very important tool to analyze the time series. Common approach is to observe the series as a point $(s_1, \ldots s_m)$ in m-dimensional Euclidean space, when any of the known clustering algorithms may be applied. Disadvantages of such approach

are that the series must have the same complexity, and the measurements have to be made synchronizely, what is often hard to obtain. The other possible approach is to measure distance between the curves by Fréchet distance, that successfully solves these two difficulties and depends only on the shape of the curves.

In the literature the Fréchet distance between two curves τ and π is often illustrated as the minimal length of the leash that enables that a man, running along τ, walks a dog on the leash running along π, where both man and dog are not allowed to run backwards and the speeds of both of them may vary. Formally it is defined as follows:

Let $\mathcal{H}$ be the set of continuous and increasing functions $f : [0,1] \to [0,1]$ with the property that $f(0) = 0$ and $f(1) = 1$. For two given input curves $\tau : [0,1] \to \mathbb{R}$ and $\pi : [0,1] \to \mathbb{R}$, their *Fréchet distance* is defined as

$$d_F(\tau, \pi) = \inf_{f \in \mathcal{H}} \max_{t \in [0,1]} \|\tau(f(t)) - \pi(t)\|, \tag{1}$$

The Fréchet distance between two time series is defined as the Fréchet distance of their corresponding signals. The function f that realizes the Fréchet distance is called a *matching*.

The most often used algorithm to calculate the Fréchet distance is one by Alt and Godau [2]. The best algorithm to calculate the Fréchet distance was given by Buchin et al. [3], and it requires quadratic running time in terms of curve complexities.

It remains for us to formulate our problem:

We denote with Δ_l the set of all univariate time series of complexity at most l. Given a set of n time-series $P = \tau_1, \ldots \tau_n \subseteq \Delta_m$ and parameters $k, \ell \in \mathbb{N}$, we define a (k,ℓ)-*clustering* as a set of k time-series $C = c_1, \ldots c_k$ taken from Δ_ℓ which minimize one of the following cost functions:

$$\text{cost}_\infty(P, C) = \max_{i=1,\ldots n} \min_{j=1,\ldots k} d_F(\tau_i, c_j)\,. \tag{2}$$

$$\text{cost}_1(P, C) = \sum_{i=1,\ldots n} \min_{j=1,\ldots k} d_F(\tau_i, c_j)\,. \tag{3}$$

We refer to these clustering problem as (k,ℓ)-*center* and (k,ℓ)-*median* respectively.

On signatures

To capture the critical points of the input time series, and therefore to reduce the their complexity, we introduce the concept of *signatures*. The idea to find the low-complexity cluster centers that well describe the input set is old. We want to use a technique similar to the idea of shortcutting used for partial curve matching in [4, 5]. Our signatures have

the strict hierarchical structure, that enables us to calculate them efficiently, both in the case when the error δ or goal complexity k is given. We prove also that the signatures always exist and give the algorithms to calculate them.

Definition 1 (δ-signature)**.** *We define the δ-signature of any polygonal curve $\tau : [0,1] \to \mathbb{R}$ as a curve π, defined by a series of values $0 = t_1 < \cdots < t_k = 1$ as the linear interpolation of $\tau(t_i)$ in the order of the index i, and with the following properties, for $1 \leq i \leq k-1$:*

(i) (non-degeneracy) if $i \in [2, k-1]$ then $\tau(t_i) \notin \langle \tau(t_{i-1}), \tau(t_{i+1}) \rangle$,
(ii) (direction-preserving)
if $\tau(t_i) < \tau(t_{i+1})$ for $s < s' \in [t_i, t_{i+1}]$: $\tau(s) - \tau(s') \leq 2\delta$, and
if $\tau(t_i) > \tau(t_{i+1})$ for $s < s' \in [t_i, t_{i+1}]$: $\tau(s') - \tau(s) \leq 2\delta$,
(iii) (minimum edge length)
if $i \in [2, k-2]$ then $|\tau(t_{i+1}) - \tau(t_i)| > 2\delta$, and
if $i \in \{1, k-1\}$ then $|\tau(t_{i+1}) - \tau(t_i)| > \delta$,
(iv) (range) for $s \in [t_i, t_{i+1}]$:
if $i \in [2, k-2]$ then $\tau(s) \in \langle \tau(t_i), \tau(t_{i+1}) \rangle$, and
if $i = 1$ and $k > 2$ then $\tau(s) \in \langle \tau(t_i), \tau(t_{i+1}) \rangle \cup \langle \tau(t_i) - \delta, \tau(t_i) + \delta \rangle$, and
if $i = k$ and $k > 2$ then $\tau(s) \in \langle \tau(t_{i-1}), \tau(t_i) \rangle \cup \langle \tau(t_i) - \delta, \tau(t_i) + \delta \rangle$, and
if $k = 2$ then $\tau(s) \in \langle \tau(t_1), \tau(t_2) \rangle \cup \langle \tau(t_1) - \delta, \tau(t_1) + \delta \rangle \cup \langle \tau(t_2) - \delta, \tau(t_2) + \delta \rangle$,

where $\langle a, b \rangle$ denotes the interval $[\min(a, b), \max(a, b)]$.

Our contribution

Our algorithms are first $(1+\varepsilon)$-approximation algorithms for the k-center and k-median problems. The main results that will be published in Driemel et al. [6] are following two theorems.

Theorem 2. *Let $\varepsilon > 0$ and $k, \ell \in \mathbb{N}$ be given. There exists a constant $t_{k,\ell,\varepsilon}$ such that given a set of curves $P = \{\tau_1, \ldots, \tau_n\}$ we can compute a $(1+\varepsilon)$-approximation to the optimal (k, l)-center clustering of P in time $O\big(nm \log\big(\frac{1}{\varepsilon}\big) + t_{k,\ell,\varepsilon} nm\ell \log(m\ell) \log\big(\frac{1}{\varepsilon}\big)\big)$.*

The following theorem on k-median clustering leans to the work of Ackermann *et al.* [1], after we have proven that the modified sampling property holds for the Fréchet distance and our problem definition (3) when the sample size depends on l.

Theorem 3. *Given a set of curves $P = \{\tau_1, \ldots, \tau_n\} \subset \Delta_m$ and a parameter $\varepsilon > 0$, there exists an algorithm that with constant probability returns a $(1+\varepsilon)$-approximation to the (k, ℓ)-median problem for input instance P, and that has running time in*

$$O(n 2^{O(k m_{\varepsilon/3,\lambda,\ell} \log(\frac{k}{\varepsilon} m_{\varepsilon/3,\lambda,\ell}))} m\ell \log m\ell),$$

where $m_{\varepsilon,\lambda,\ell}$ is a constant depending only on ε, λ and ℓ.

For constant ε, k and ℓ, the running time of our algorithms is therefore $\widetilde{O}(nm)$.

The following lemma claim tha the clustering of the univariate curves is NP-hard. Thereby we reduce our problem to [7].

Lemma 4. *Let $\delta > 0$ be a given parameter. The (k, ℓ)-center and (k, ℓ)-median clustering δ-decision problems under Fréchet distance are NP-hard. Furthermore, the (k, ℓ)-center problem is NP-hard to approximate within a factor of 2.*

The Fréchet distance is a pseudo metric, as the reflexivity property of a metric does not hold, since there may be two different functions that have distance 0. If such functions are grouped into equivalence classes, we obtain a metric space such that holds:

Lemma 5. *Let $(X, d_F(\cdot, \cdot))$ be the metric space, where X is a set of all one-dimensional time-series of complexity l and d_F be the Fréchet distance metrics. Then it holds for the doubling dimension d of the metric space (X, d_F) that:*

(i) if l is not bounded, d is not bounded;
(ii) if l is finite and $l \geq 4$, d is not bounded;
(iii) if $l = 2$, then $d = 2$, and if $l = 3$, then $d = 3$.

References

[1] M. R. Ackermann, J. Blömer, and C. Sohler. Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms*, 6(4):59:1–59:26, 2010.

[2] H. Alt and M. Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5:75–91, 1995.

[3] K. Buchin, M. Buchin, W. Meulemans, and W. Mulzer. Four soviets walk the dog-with an application to alt's conjecture. *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1399–1413, 2014.

[4] K. Buchin, M. Buchin, and Y. Wang. Exact algorithms for partial curve matching via the Fréchet distance. In *Proceedings of the 20th ACM-SIAM Symposium on Discrete Algorithms*, pages 645–654, 2009.

[5] A. Driemel and S. Har-Peled. Jaywalking your dog – computing the Fréchet distance with shortcuts. *SIAM Journal of Computing*, 42(5):1830–1866, 2013.

[6] A. Driemel, A. Krivošija, and C. Sohler. Clustering time series under the fréchet distance. *Computing Research Repository (arXiv)*, to appear, 2014.

[7] N. Megiddo and K. J. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13(1):182–196, 1984.

Online Preemptive Scheduling with Deadlines and Utilization Maximization

Chris Schwiegelshohn
Lehrstuhl für Effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
chris.schwiegelshohn@tu-dortmund.de

We address a basic online scheduling problem with preemption and deadlines on parallel identical machines. We partition the instance space by characterizing the deadlines with the help of the so called stretch factor and by using limitations of the jobs' processing times. As in most online problems, jobs are submitted over time and a decision whether to accept or to reject a job is required at its submission before any other job is considered. A job can only be accepted if it and all previously accepted jobs can be completed on time. Our objective is the maximization of machine utilization. We apply competitive analysis to determine the performance of algorithms. While greedy acceptance of jobs achieves an optimal competitive factor for the deterministic single machine model, we show that in case of parallel identical machines a more complex approach produces better results.

1 Introduction

We consider a basic online scheduling problem in a parallel identical machine environment that supports preemption and migration. Jobs are submitted over time and have a deadline. As in most online problems, we must immediately decide for every submitted job whether it is accepted or rejected, that is, we cannot postpone our decision until we have received information about future job submissions even if two jobs have the same submission time. We can only accept a job if it and all previously accepted jobs can be completed at or before their respective deadline. In scheduling literature, the weighted total number of late jobs ($\sum w_j U_j$) is the most common deadline related objective. There

are two specific variants of this problem: the first variant assumes that all weights are 1 which corresponds to the total number of while the second sets each job weight to be identical with the processing time of the job ($w_j = p_j$). In this paper, we address the second problem and describe our problem as $P_m|r_{j,\text{online}}, \text{pmtn}, \text{migration}|\sum p_j U_j$ using the three-field-notation.

Informally our objective function describes the amount of resources (machine-time-product) that is requested but cannot be granted. The objective function has practical relevance as a resource provider typically considers the amount of resources consumed by a job when determining the price of its execution. Since a resource provider is likely more interested in the received income rather than the lost income, we consider the maximization of $\sum p_j(1 - U_j)$ instead of the minimization of $\sum p_j U_j$. Competitive analysis introduced by Sleator and Tarjan [9] is typically used to evaluate the performance of online algorithms but of limited use for this problem formulation. For instance, consider a simple instance of our problem in a single machine scenario: The first job request arrives at time 0. The job has processing time 1 and a tight deadline of 1. If we reject the job and no further requests are submitted then our competitive ratio will be 0. But if we accept the job and immediately afterwards receive a single additional job with a very large processing time Δ and a tight deadline Δ as well then the competitive ratio is $1/\Delta$ as an optimal solution rejects the first job and accepts the second one. This straightforward example (which can easily be extended to parallel and randomized settings) shows that there is no online algorithm with an acceptable competitive factor for our problem.

In order to obtain more meaningful results, we can partition the problem space and apply the analysis on every partition separately. In our problem we determine for each job the difference between deadline and submission time. Then we consider the smallest ratio between this difference and the processing time of the corresponding job. This smallest ratio is referred to as the the stretch or slack factor f for $f > 1$. We assume that the stretch factor is known a priori and can be used to determine an appropriate online algorithm. In practice such assumption is reasonable since the resource provider usually can choose a stretch factor as part of the system policy.

1.1 Related Work

Variants of our problem have been discussed in the context of real time scheduling. Baruah and Haritsa [1] address online scheduling of independent jobs with slack factor and preemptions on a uniprocessor system. In their paper, they present the generalized ROBUST (Resistance to Overload By Using Slack Time) algorithm. Their algorithm guarantees an effective processor utilization (EPU) of $1+\frac{1}{f-1}$ during any overload interval. The effective processor utilization is closely related to our competitive factor. DasGupta and Palis [2] base their hard real-time scenario on the concept of Baruah and Haritsa and extend it to parallel identical machines that allow preemption without migration. They

show that the competitive factor $1 + \frac{1}{f-1}$ of a uniprocessor cannot be improved for their model of parallel identical machines and that a simple greedy acceptance policy (always accepting a job if it can be accepted) achieves this competitive factor.

Other theoretical scheduling papers address similar online problems without preemptions. Lipton and Tomkins [7] present randomized algorithms for non-preemptive interval scheduling on a single machine if there are only two different lengths for the intervals. Goldman et al. [3] generalize the model of Lipton and Tomkins by introducing a stretch factor and prove a competitive factor of $1 + \frac{\lceil f-1 \rceil}{f-1}$ for two different processing times. Goldwasser [4] distinguishes between exactly two distinct processing times or arbitrary processing times and gives tight deterministic competitive factors of $1 + \max\left(\frac{\lceil f-1 \rceil+1}{\lceil f-1 \rceil}, \frac{\lfloor f-1 \rfloor+1}{f-1}\right)$ and $2 + \frac{1}{f-1}$, respectively. Lee [6] further studies the case of small stretch factors $f \leq 2$ and gives a randomized algorithm for a single machine with a competitive factor of $O\left(\log \frac{1}{f-1}\right)$ based on a deterministic algorithm for parallel identical machines.

In the offline variants of our problem all information on release dates, deadlines, and processing times are available to determine the schedule. Without considering any restrictions, these problems are weakly NP-hard, see Lawler [5]. We can solve the problem $P|\text{pmtn}|L_{\max}$ in polynomial time by considering it in reverse time direction ($P|r_j, \text{pmtn}|C_{\max}$) and applying the Longest Remaining Processing Time (LRPT) first algorithm, see Pinedo [8]. Therefore, we can decide immediately after the submission of a new job whether it is possible to *legally* execute this job ($L_{\max} \leq 0$) or not ($L_{\max} > 0$).

1.2 Our Contribution

We extend the parallel scheduling problem by allowing preemptions with migrations, that is a job being processed on a machine A may be halted and later resumed with the same remaining workload on a different machine B. In this setting the lower bounds of DasGupta and Palis [2] no longer apply. In particular, for small $f \leq 2$ we show an exponential improvement with respect to their bounds for the deterministic parallel setting. Our deterministic online algorithm is given in Algorithm 1. The exact value of $e(m, f)$ is given from the evaluation of a piecewise step function. Moreover, we show that a greedy algorithm is not optimal and, for $f \leq 2$, can be as much as an exponential factor worse. A precise statement of all our results for arbitrary f and m is out of the scope of this short abstract. We therefore are content with the following excerpt.

Theorem 1. *For any number of machines m and any stretch factor f, Algorithm 1 is a tight deterministic algorithm. As m tends to infinity, the competitive ratio converges to* $f \cdot \ln \frac{f}{f-1}$.

Algorithm 1: Online Saving Acceptance

input: Currently submitted job J_j

Include J_j into the list of accepted jobs.

Apply job splitting if required.

forall the deadlines $d_i \geq d_j$ of accepted but not completed jobs **do**

  if $\sum_{J_h | d_h \leq d_i} p_h > d_i \cdot e(m, f)$ **then**

    remove J_j from the list of accepted jobs

    return Job J_j is rejected.

return; Job J_j is accepted.

References

[1] S.K. Baruah and J.R. Haritsa. Scheduling for overload in real-time systems. *IEEE Trans. Computers*, 46(9):1034–1039, 1997.

[2] B. DasGupta and M.A. Palis. Online real-time preemptive scheduling of jobs with deadlines on multiple machines. *Journal of Scheduling*, 4(6):297–312, 2001.

[3] S.A. Goldman, J. Parwatikar, and S. Suri. Online scheduling with hard deadlines. *Journal of Algorithms*, 34(2):370 – 389, 2000.

[4] M.H. Goldwasser. Patience is a virtue: the effect of slack on competitiveness for admission control. In *Proceedings of the tenth annual ACM-SIAM symposium on discrete algorithms*, SODA, pages 396–405, Philadelphia, PA, USA, 1999. Society for Industrial and Applied Mathematics.

[5] E.L. Lawler. Recent results in the theory of machine scheduling. In A. Bachem, M. Groetschel, and B. Korte, editors, *Mathematical programming: the state of the art (Bonn, 1982)*, pages 202–234. Springer, Berlin, 1983.

[6] J. Lee. Online deadline scheduling: multiple machines and randomization. In *Proceedings of the fifteenth annual ACM symposium on parallel algorithms and architectures*, SPAA, pages 19–23, New York, NY, USA, 2003. ACM.

[7] R.J. Lipton and A. Tomkins. Online interval scheduling. In *In Proceedings of the fifth Annual ACM-SIAM symposium on discrete algorithms*, SODA, pages 302–311. Society for Industrial and Applied Mathematics, 1994.

[8] M.L. Pinedo. *Scheduling: Theory, Algorithms, and Systems*. Springer Science+Business Media, forth edition, 2010.

[9] D.D. Sleator and R.E. Tarjan. Amortized efficiency of list update and paging rules. *Commun. ACM*, 28(2):202–208, 1985.

Subproject A3
Methods for Efficient Resource Utilization in Machine Learning Algorithms

Jörg Rahnenführer Peter Marwedel

Advances in Semi-offline Scheduling Schemes for Hard Timing-constrained Software

Jan Kleinsorge
Entwurfsautomatisierung für Eingebettete Systeme
Technische Universität Dortmund
jan.kleinsorge@tu-dortmund.de

Worst-case time (WCET) analyses for single tasks are well established and their results ultimately serve the purpose of providing execution time parameters for schedulability analyses. Besides WCET analysis, an important problem is maximum blocking time (MBT) analysis which is essential in deferred preemption schedules for the selection of preemption points. Among the most pressing problems in this context is the need for good path analyses, which are a fundamental bottleneck for selecting these points. Current state of the art relies on ILP-based or severely constrained explicit path analyses, both of which are unsatisfactory in general.

The major advance in our research of this year is to propose a general explicit path analysis to compute maximum blocking times, specifically for scheduling policies with deferred preemption. The proposal improves the current state of the art significantly for both WCET and MBT analysis, as it is efficient, accurate, easily extensible and computes path lengths between all program points, without imposing any artificial constraints, and under a general flow bound model, unmatched by other existing explicit path analyses, while significantly outperforming the ILP-based approach.

In embedded software systems, timing analysis and feasibility analysis of multi-task scheduling are key elements to reason about its timing behavior. For hard real-time systems, static *worst-case execution time* (WCET) analysis can provide safe upper time bounds for the uninterrupted execution of separate tasks. Schedulability analysis in turn asserts the applicability of a scheduling policy under given processing resources that are reflected by WCET.

Task preemptions cause task interruptions that are not reflected by the WCET of an isolated task alone. Direct and indirect context switch costs have to be taken into account. Direct costs are ones that result, for example, from saving and restoring execution context and from disruption of pipelined execution. The latter are caused by content changes to system caches due to execution of other tasks. These *cache-related preemption delays* (CRPD) are indirect because they do not necessarily occur right at the preemption point, after a task is resumed, but have a potential effect on the execution time at a later program point, when cache contents are actually requested. They are highly variable and potentially vastly dominate overall preemptions costs. Direct costs, on the other hand, are comparably small and can often be safely bounded with a constant.

CRPD can be bounded by statically analyzing access patterns of the cache by the preempting tasks and the reuse pattern by the preempted task. However, in *fully preemptive scheduling*, neither points in control flow nor points in time can be accurately determined at which preemptions occur. This means only worst-case assumptions can be made about the number of preemptions and the CRPD, which leads to significant overestimations. Overall, static timing predictability is low. The most important advantage is that fully preemptible tasks yield low average response times for higher priority tasks.

Non-preemptive scheduling represents the other extreme. On the one hand, timing predictability is very high, since the CRPD, as a major source of inaccuracy, does not have to be taken into account. On the other hand, some task sets that have been schedulable under fully preemptive scheduling potentially have no feasible schedule under this policy. The reason for this is the blocking time that lower priority tasks now impose on tasks of higher priority ready to execute.

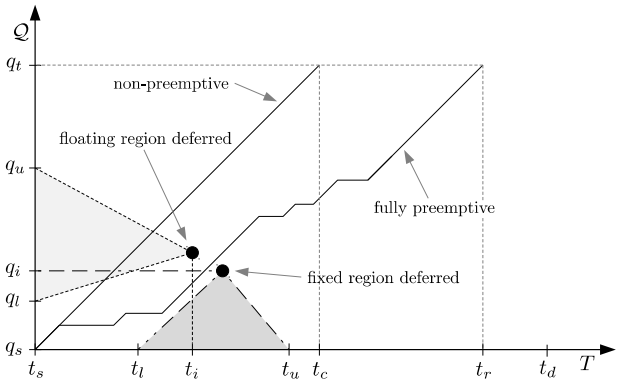


Figure 1: Relation and performance trade-offs of non-preemptive, fully-preemptive and deferred preemption schemes

The compromise is to employ scheduling policies with only limited preemption — so called *deferred preemption scheduling* — which either feature floating or fixed non-preemptive regions. If preemptions are time-triggered, then it is in general not possible to know the exact execution context in which a preemption occurs. Hence the name *floating region*.

Since strict timing guarantees can be given on the length of such a region, existing schedulability tests can directly be applied. The downside is, again, the lack of accuracy for the CRPD. *Fixed regions* are established by adding explicit preemption points to the program code. Although program context can now be determined much more accurately at the point of preemption, it is hard to determine the point in time of the preemption, and thus, the length of these regions. Moreover, to significantly reduce the actual and estimated costs of fixed regions, preemption points have to be chosen carefully.

Figure 1 illustrates the respective trade-offs of the different scheduling schemes.

To support fixed preemptive regions, *maximum blocking time* (MBT) analysis is required which computes the longest paths between potential preemption points. The goal is not to find any longest path between two program points, but the longest path to a specific program point, such that it only occurs once on this path — the most distant preemption points. In general, the lack of tools to tighten the relation between execution time and program context is a major obstacle in enhancing the accuracy of static multi-task analyses. In this context, WCET analysis and MBT analysis allow to map program points onto execution time and vice versa. Improving the accuracy of either potentially enhances worst-case analysis of all three kinds of preemption policies mentioned.

The focus of our research is the design of a general and efficient algorithm for MBT analysis, advancing the state of the art in explicit path analysis to facilitate deferred preemption schedules. It is designed not only to serve the mere theoretic purpose to enable such analyses at all, it is also technically simple.

The resulting approach has been published in [2].

We briefly summarize the results here. We performed runtime measurements on the Mälardalen WCET benchmark suite (MRTC [1]), a representative set of benchmarks for hard real-time systems, hence representing particularly resource-constrained software.

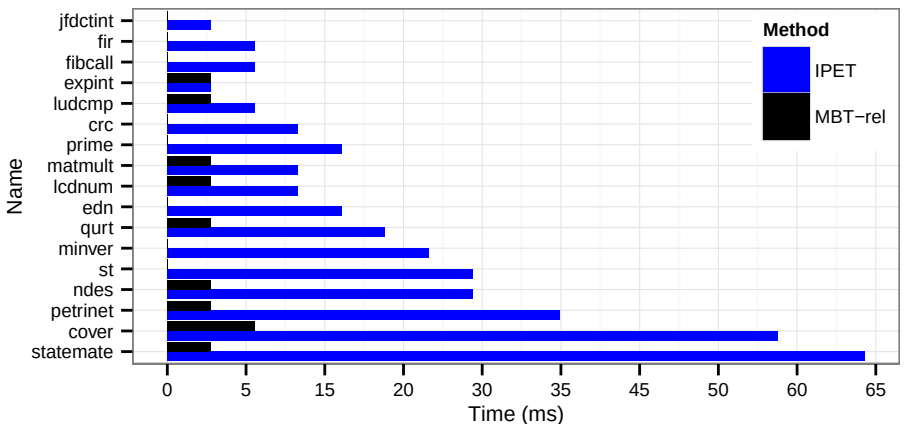


Figure 2: Runtimes on real-time benchmarks

Fig. 2 shows the results for a random subset of MRTC benchmarks at a sampling resolution of 1ms. Our approach (*MBT-rel*) significantly outperforms the ILP-based approach (*IPET*) in all use cases. We solve the MBT problem from the source to all reachable nodes below 1ms in some cases.

Our approach dominates alternative approaches in terms of generality and accuracy. MBT now allow for precise preemption point placement in general CFG. To completely replace IPET in practice for MBT as well as for WCET analysis, an efficient way to model mutual exclusion of paths is required, and generalized flow bounds must be further discussed in all detail. We will address these limitations specifically and propose an analysis framework that has the potential to yield significant gains in performance and accuracy for static multi-task timing analysis.

References

[1] Mälardalen WCET benchmark suite. `http://www.mrtc.mdh.se/projects/wcet/benchmarks.html`.

[2] Jan Kleinsorge and Peter Marwedel. Computing Maximum Blocking Times with Explicit Path Analysis under Non-local Flow Bounds. In *Proceedings of the International Conference on Embedded Software (EMSOFT 2014)*, EMSOFT 2014, New Delhi, India, oct 2014.

Memory Optimization for the R Language

Helena Kotthaus
Computer Science 12
TU Dortmund University
helena.kotthaus@tu-dortmund.de

Dynamic languages like R are increasingly used to process large data sets. Here, the R interpreter induces a large memory overhead due to wasteful memory allocation policies [1, 3]. If an application's working set exceeds the available physical memory, the OS starts to swap, resulting in slowdowns of several orders of magnitude. Existing R optimizations are mostly based on dynamic compilation or native libraries. Both methods are futile when the OS starts to page out memory. So far, only a few, data-type or application specific memory optimizations for R exist. To remedy this situation, we present a page sharing approach for R that significantly reduces the interpreter's memory overhead [2].

1 Memory Saving Strategies

The R interpreter assumes that objects are laid out in contigous blocks in memory. It uses a copy-on write mechanism to defer copy for call-by-value semantics. When an object modification happens, an duplicate function working on object-level granularity is used. An object is commonly a vector that span multiple pages. The R interpreter produces unnecessary duplicates, when only some of the object pages are modified.

Our optimization reduces this overhead by duplicating on page-level granularity. Here only parts of the object are copied instead, sharing the same memory pages between multiple objects as long as they are not modified. This is illustrated in Figure 1. On the left side the situation before duplication is shown. For a duplication of this object the standard interpreter would need five new physical pages when modification happens.

As can be seen on the right side of this figure our optimization reduces this to a single allocated page H' which includes the header of the object. Pages that are unmodified are only virtually copied and shared. To know how many times a page is shared we use reference counters shown below the physical pages.

To enable memory optimizations within the standard R interpreter we employed a custom memory allocator which uses a memory management scheme similar to standard virtual memory schemes used in OS kernels. When the internal function of the R interpreter is called to allocate an R object, our optimized interpreter can select between the default allocator and our custom allocator for page sharing. Given that not every object has sharing potential, the custom allocator is only enabled for objects with a size that indicates potential for sharing. R objects smaller than two pages are therefore passed to the default allocator.

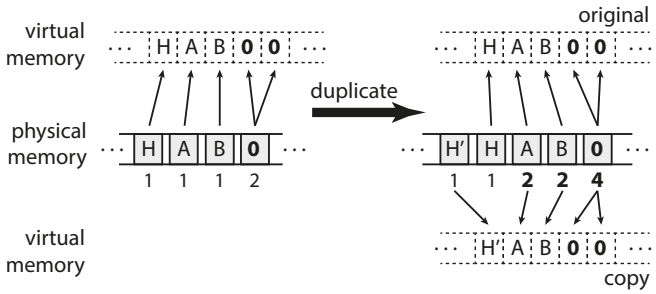


Figure 1: Duplication with page sharing [2]

Besides this refinement and the duplication on page-level granularity, we also optimized the object allocation. Here we used a page filled with zeros called global zero page. This zero page is also shown in Figure 1 where it is shared. When the custom allocator is asked to allocate an object, physically this object only consists of a single non-shared page H for the header followed by a shared page which is the global zero page. This global page avoids the zero-initialization of allocated objects. With our optimization, memory is not only saved during allocation, we also use a dynamic refinement that saves memory by reducing identical content. We constructed a restricted deduplication algorithm. This optimization checks for pages identical to the existing global zero page and frees them. Compared to scanning for arbitrary duplicated content, this scan has less overhead as it can stop as soon as a non-zero element is found. This content check is only triggered after garbage collection to reduce the number of checks.

Our scheme is transparent to the interpreter's memory management, requiring only small changes in memory allocation and freeing, as well as in the duplicate function. Compared to general approaches like memory deduplication used in the kernel samepage merging [5] or compression which lack knowledge about memory behavior of the runtime environment, we can reduce the overhead since we pro-actively avoid the generation of page duplications by sharing pages right from the start.

2 Memory Consumption and Runtime Evaluation

We compared the R Interpreter including our optimizations against the standard interpreter. Therefore we used a set of 15 real-world benchmarks [4]. For the memory analysis we measured several parameters. Since peak memory usage does not represent information about changing memory usage over time, we also calculated an average memory usage by measuring the peak usage during one-second intervals and averaging these peaks over the entire runtime. Figure 2 shows the results of peak and average memory usage. The 100% baseline represents the memory usage of the standard R interpreter. Even though our optimization results in a slight increase of peak or average memory usage for three of the benchmarks, all others gain from our optimizations in both the peak and average memory usage. The geometric mean over all fifteen benchmarks shows a reduction of peak memory usage by 13.6% and a reduction of average memory usage by 18.0% [2].

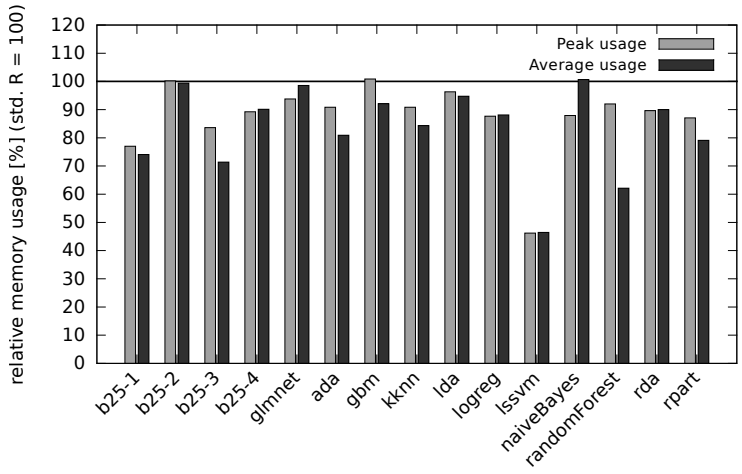


Figure 2: Relative memory usage with page sharing compared to standard R (lower is better) [2].

We also evaluated the runtime, here our optimization incurs an overhead of 5.3% on average [2]. When the RAM in the system is too small to hold all data required, there are situations where we can also reduce the runtime instead of adding overhead.

The best case is shown in Figure 3 which represents the memory-over-time profile of lssvm. We limited the system to 1GB of RAM instead of increasing the data set size since runtime does not scale linearly with the data size. For lssvm we gained a speed up of a factor of 5 by saving enough memory to avoid swapping, which makes the overhead of our optimization irrelevant. This shows that reducing the memory consumption can improve the runtime for memory–hungry benchmarks. In turn, it allows processing larger data sets.

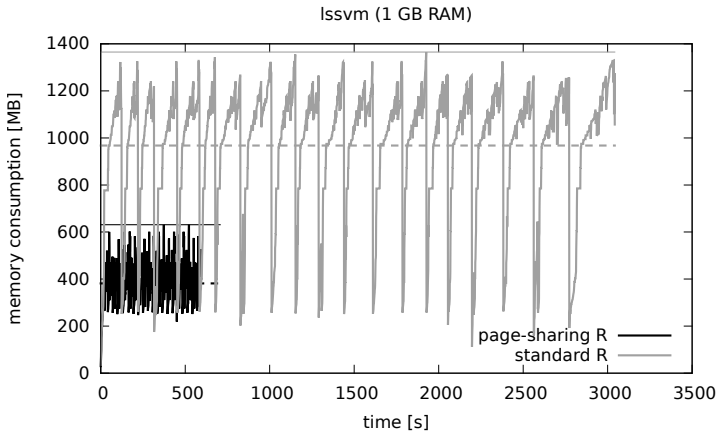


Figure 3: Memory consumption over time profile for lssvm with page sharing vs standard R. Lines at the top mark the respective peak memory usage, dotted lines mark the average memory usage [2].

We presented an application transparent memory optimization that employs page sharing at a memory management layer between the R interpreter and the operating system's memory management. By concentrating on the most rewarding optimizations we avoid the high runtime overhead of existing generic approaches. For future work, we plan to enable possibilities for concurrent optimization of multiple applications with large memory footprints. This could result in a better utilization of multicore systems since it allows to execute a larger number of applications simultaneously without forcing the OS to swap.

References

[1] Kotthaus H., Korb I., Lang M., Bischl B., Rahnenführer J., Marwedel P., Runtime and Memory Consumption Analyses for Machine Learning R Programs. Journal of Statistical Computation and Simulation. 2014.

[2] Kotthaus H., Korb I., Engel M., Marwedel P., Dynamic Page Sharing Optimization for the R Language. Proceedings of the 10th Symposium on Dynamic Languages. 2014.

[3] Kotthaus, H. et al.: Performance Analysis for R: Towards a Faster R Interpreter. In Abstract Booklet of the International R User Conference. 2014.

[4] Lang M., Kotthaus H, BenchR: Set of Benchmark of R. TU Dortmund University. https://github.com/allr/benchR 2014.

[5] Arcangeli A., Eidus I., Wright C., Increasing memory density by using KSM. In Proceedings of the Ottawa Linux Symposium. 2009.

Automatic model selection for high-dimensional survival analysis

Michel Lang
Statistical Methods in Genetics and Chemometrics
Technische Universität Dortmund
lang@statistik.tu-dortmund.de

Over the past years many models for the analysis of high-dimensional survival data have been developed. Unfortunately, this field of research lacks benchmark studies and thus only very little is known about the real predictive performance of most models. Here the computational demands are the major roadblock towards a comprehensive comparison: With more than 20 000 covariates, a single model fit can require hours of CPU time, but hundreds of such model fits are required. This has two reasons – First, to obtain an unbiased performance measure, repeated validation on independent test sets is mandatory. And second, every statistical model has hyperparameters which have to be tuned in order to assess a meaningful and representative prediction performance estimate. Both issues can be solved by a modern black box optimization technique.

There is extreme need for objective and reproducible comparison studies for analyses of high dimensional genomic data. A first approach can be found in [6] where iterated F-racing [7] was used to tune feature selection filters and survival models simultaneously. The application of this algorithm configuration on four lung cancer data sets led to models which dominated all typical models for high-dimensional survival analysis. The resulting concordance indices are summarized in Table 1. The analysis of the optimization path allowed to derive even more important information about filters, models and data sets as well.

To follow up the promising results we now aim at a even broader comparison study. Therefore ten lung cancer and thirty breast cancer data sets have been diligently accumulated and with the help of clinicians carefully pre-processed. Figure 1 gives an

Data set	CoxBoost	CoxPH	mboost	Ridge	Configurator
GSE31210	0.24	0.35	0.24	0.23	0.18
GSE4573	0.47	0.45	0.46	0.47	0.44
GSE37745	0.49	0.54	0.46	0.49	0.42
Jacob	0.33	0.34	0.33	0.33	0.31

Table 1: Mean concordance indices (less is better) for baseline models CoxBoost (Likelihood-based boosted Cox PH Model), Cox PH (simple Cox PH model with upstream feature filter), mboost (Componentwise boosted Cox PH model), Ridge (Ridge-regularized Cox PH model) with model resulting from iterated F-racing. All cross-validated with three folds.

exemplary overview over the survival times of the extended set of lung cancer data sets.

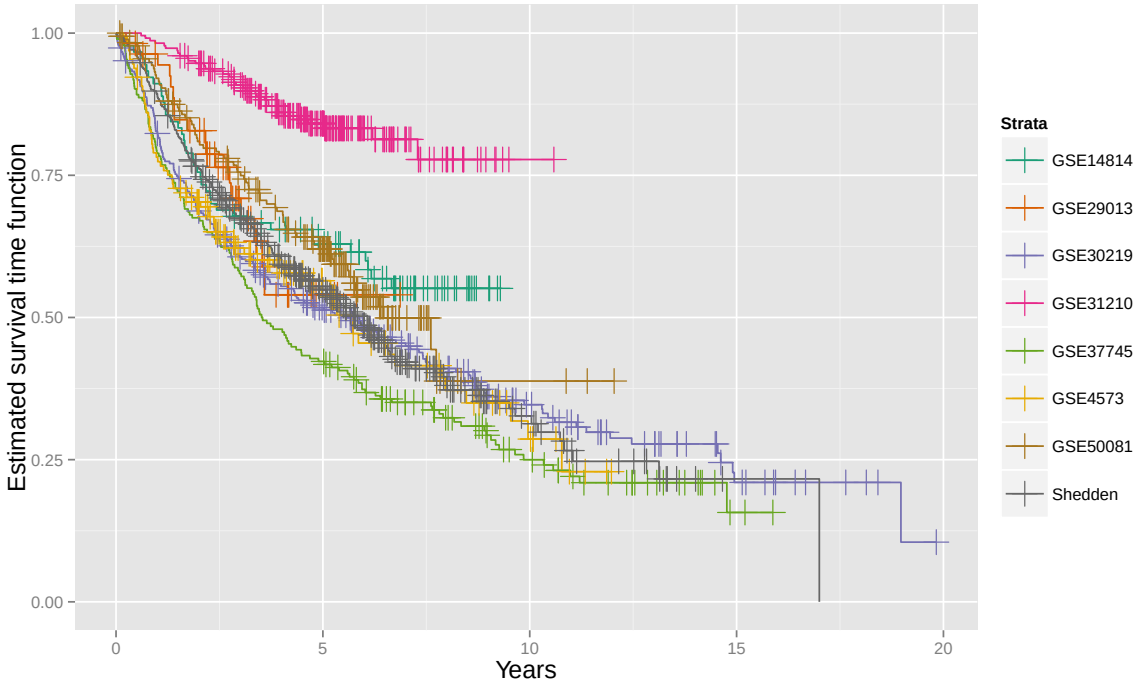


Figure 1: Estimated survival times for a selection of lung cancer data sets.

To simplify benchmark studies for the future, all popular implementations of survival models have been included into the `R` package `mlr` [4]. This package offers a unified interfaced for more than 90 machine learning algorithms which are hosted on the comprehensive R archive network (CRAN). Many convenience functions allow a similar but more programmatically workflow like RapidMiner or Weka.

Optimization and benchmark studies influenced the development of `mlr` in multiple ways. As a result all key requirements required for such expensive computer experiments are now met:

- Configurable and fail-safe error handling.
- A reflection system which allows programming on tasks, learners, performance measures and resampling techniques.
- Extensive pre-processing capabilities. Especially imputation of missing values are tremendously important for algorithm configuration.
- Wrapping learning algorithms to extend their functionality, e. g. models can be wrapped into a filter method in order to fuse a new, custom learning algorithm.
- Support for parallelization. `parallelMap` [3] gives the user fine-grain control over the parallelization using a register for parallel levels. `BatchJobs` [1] for the parallelization on high-performance computing clusters is just one of the five supported backends.

Furthermore `mlr` interfaces many tuners, i. e. the iterated F-racing using the `irace` package, are already built-in and can also be wrapped to create new learners. Model-based optimization (MBO) [5] – which will be employed for the next benchmark study – is supported via the add-on package `mlrMBO` [2]. MBO uses a regression approach to model the performances of algorithms in their parameter space. New configurations are suggested by scoring possible points in the parameter space using the expected improvement criteria: A configuration is more interesting the better the performance and the less is known about the respective area (i. e., the higher the estimated variance). We plan to extend MBO to estimate computational resources and use these estimates for efficient scheduling.

Because of the clear heterogeneity between cohorts, the algorithm configuration is performed on each data set separately. Yet we think it might also be a worthwhile endeavor to try to find model configurations for a whole domain of survival analysis data sets, e. g., all lung cancer gene expression data.

References

[1] B. Bischl, M. Lang, O. Mersmann, J. Rahnenführer, and C. Weihs. Computing on high performance clusters with R: Packages BatchJobs and BatchExperiments. Technical Report 1/2012, TU Dortmund University, 2012. Available at: `http://sfb876.tu-dortmund.de/PublicPublicationFiles/bischl_etal_2012a.pdf`.

[2] Bernd Bischl, Jakob Bossek, Daniel Horn, and Michel Lang. mlrmbo: Model-based optimization for mlr. R package version 1.0.

[3] Bernd Bischl and Michel Lang. parallelMap: Unified interface to some popular parallelization back-ends for interactive usage and package development, 2014.

[4] Bernd Bischl, Michel Lang, Jakob Richter, Jakob Bossek, Leonard Judt, Tobias Kuehn, Erich Studerus, and Lars Kotthoff. mlr: Machine learning in r. R package version 2.3.

[5] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proceedings of LION-5*, page 507–523, 2011.

[6] Michel Lang, Helena Kotthaus, Peter Marwedel, Claus Weihs, Jörg Rahnenführer, and Bernd Bischl. Automatic model selection for high-dimensional survival analysis. *Journal of Statistical Computation and Simulation*, 85:62–76, 2014.

[7] M. López-Ibáñez, J. Dubois-Lacoste, T. Stützle, and M. Birattari. The irace package, iterated race for automatic algorithm configuration. Technical Report TR/IRIDIA/2011-004, IRIDIA, Université Libre de Bruxelles, Belgium, 2011.

Analysis of high dimensional toxicological data

Eugen Rempel
Statistical Methods in Genetics and Chemometrics
Technische Universität Dortmund
rempel@statistik.tu-dortmund.de

Problem statement
In the last years the number of available datasets with high-dimensional measurements from molecular biology has increased drastically. A typical challenge is the presence of a large number of variables, often in the thousands, compared to a small number of experiments (samples), typically at most a couple hundred. In these experiments the expression (activity) or abundance of thousands of genes or proteins is measured on a genome-wide scale. The resulting data enable a better understanding of the underlying biological processes being triggered by the environmental factors. Some applications in this field originate from toxicological research. Here, one of the goals is to obtain improved models for toxicant response on the genomic level. This knowledge would enable researchers to simulate the biological cellular processes *in silico* thus reducing the number of animal experiments.

Goal
This project is based on a cooperation with Prof. Dr. Jan Hengstler from IfADo (Leibniz-Institut für Arbeitsforschung an der TU Dortmund). In a toxicological study embryonic human stem cells were treated in vitro with different compounds of two different types (mercurial compounds and histone deacetylase inhibitors [HDACi]). Then genome gene expression was measured in the treated cells. The major goal was to classify the types of compounds based on the expression data. In detail, the data set contains 4-5 technical replicates of six representatives of each type and three representatives of the control compounds. After the subtraction of controls we would like to distinguish between mercurials and HDACi's.

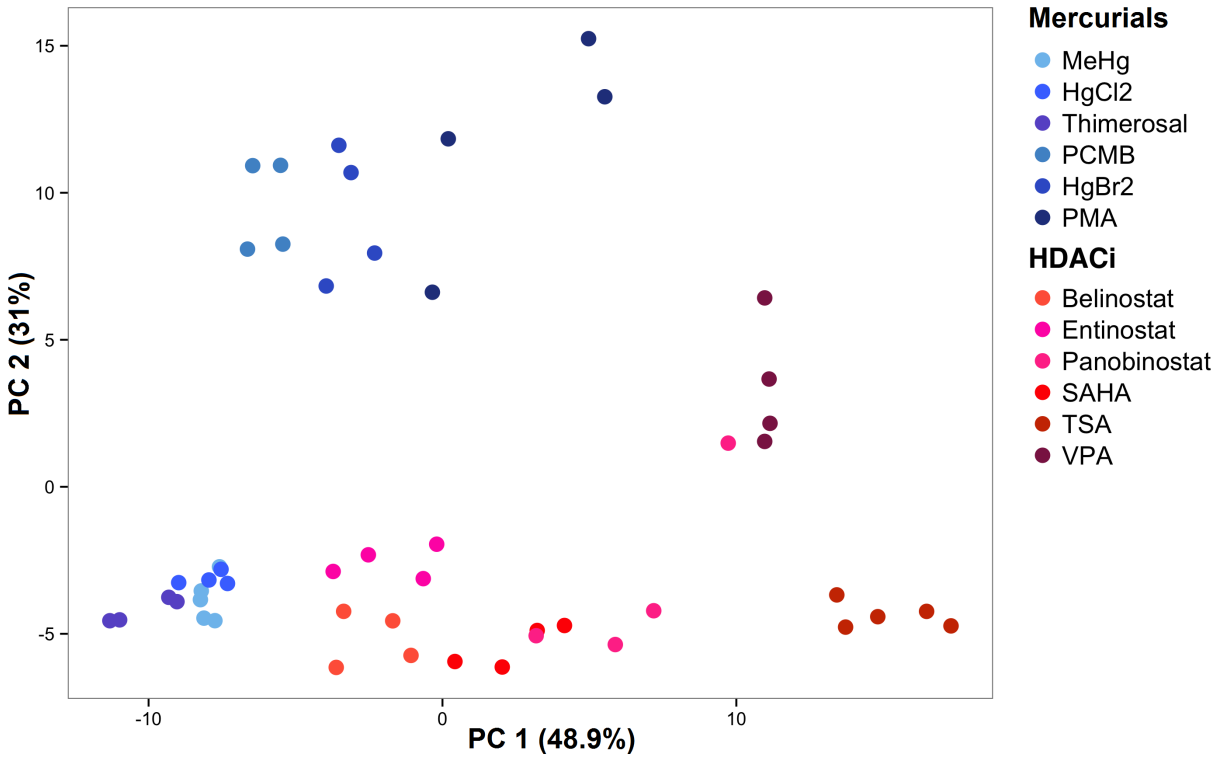


Figure 1: Principal components plot

Analysis

We have applied a large analysis pipeline to evaluate the influence of compound on the gene expression changes, including exploratory data analysis, identification of differentially expressed genes with adjusted t-test like statistics, and identification of differential pathway activity (gene set overrepresentation analysis). For example, we performed the principal component analysis to visualize the results of the classification in Figure 1.

Classification

To predict the type of compounds we have applied support vector machine. We have used technical replicates of one (two, three, four and five) compound(s) as a test set and trained the learner on the train set of left out compounds. To obtain the reliability score of classification we used the version SVM with the probabilistic output. The results of leave-one- and leave-two-out classification are presented in Figure 2. The numbers are the predicted probability for being an HDACi. The main diagonale contain the probabilities for leave-one-classification. All compounds were correctly predicted. The other probabilities represents the classification results for leave-two-out approach. Three combinations

	Belinostat	Entinostat	$HgBr_2$	$HgCl_2$	MeHg	Panobinostat	PCMB	PMA	SAHA	Thimerosal	TSA	VPA
Belinostat	0.72	0.43	0.72	0.75	0.73	0.67	0.69	0.62	0.62	0.70	0.74	0.66
Entinostat	0.25	0.73	0.75	0.74	0.80	0.59	0.76	0.76	0.70	0.76	0.75	0.74
$HgBr_2$	0.06	0.03	0.03	0.06	0.05	0.04	0.10	0.24	0.05	0.04	0.03	0.03
$HgCl_2$	0.08	0.11	0.10	0.08	0.16	0.09	0.09	0.13	0.08	0.12	0.09	0.10
MeHg	0.14	0.14	0.17	0.23	0.15	0.14	0.17	0.21	0.14	0.17	0.11	0.14
Panobinostat	0.96	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	1.00	0.99
PCMB	0.03	0.02	0.08	0.02	0.03	0.01	0.02	0.04	0.02	0.02	0.01	0.02
PMA	0.23	0.25	0.75	0.23	0.35	0.21	0.27	0.31	0.33	0.37	0.25	0.31
SAHA	0.57	0.79	0.87	0.87	0.83	0.86	0.84	0.89	0.85	0.86	0.79	0.81
Thimerosal	0.06	0.07	0.07	0.10	0.11	0.08	0.06	0.05	0.07	0.06	0.06	0.06
TSA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
VPA	0.95	0.92	0.95	0.90	0.96	0.96	0.95	0.96	0.90	0.96	0.86	0.95

Figure 2: Predicted probabilities

(PMA in absence of $HgBr_2$, Belinostat in absence of Entinostat and vice versa) were falsely predicted. These results are an indication for a possible application of machine learning approaches for the in vitro classification of toxicological compounds. These findings are to be submitted. The previous results are published in [1],[2],[3].

Literature

[1] Krug, Anne K., et al. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. Archives of toxicology, 2013, 87. Jg., Nr. 1, S. 123-143.

[2] Waldmann, Tanja, et al. Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. Chemical research in toxicology, 2014, 27. Jg., Nr. 3, S. 408-420.

[3] Balmer, Nina V., et al. From transient transcriptome responses to disturbed neurodevelopment: role of histone acetylation and methylation as epigenetic switch between reversible and irreversible drug effects. Archives of toxicology, 2014, 88. Jg., Nr. 7, S. 1451-1468.

Subproject A4
Resource efficient and distributed platforms for integrative data analysis

Peter Marwedel Olaf Spinczyk Christian Wietfeld

Resource Cost Evaluation of Fault Tolerance in an Embedded Operating System

Christoph Borchert
Department of Computer Science 12
TU Dortmund
christoph.borchert@tu-dortmund.de

The increasing ubiquity of resource-constrained embedded systems leads to a demand for dependable hardware and software systems. This conflicts with the trend of energy-efficient hardware, implemented in a few nanometer CMOS technology that operates at reduced voltages. To compensate for cheap and potentially unreliable hardware, the operating-system software layer can incorporate fault-tolerance mechanisms. This paper evaluates the resource costs of such mechanisms and serves as a starting point for deriving a formal resource cost model.

1 Introduction

Errors in main memory are one of the primary hardware problems for failures of today's computer systems. Recent studies on current DRAM technology (DDR2 and DDR3) confirm an approximate fault rate of 0.066 FIT[1]/Mbit [4], and the ever increasing demand for higher memory capacities worsens this reliability problem.

In general, the operating-system (OS) *kernel* is the most important piece of software with regard to dependability, as all other software components depend on the OS. Surprisingly, in spite of their impact on total system resiliency and – compared to the rest of the system – their very small memory footprint, state-of-the-art OS kernels are not equipped

[1]FIT (Failure In Time) is defined as 1 failure per billion hours of operation

with software-based protection against memory errors. This fact may be contributed to the tedious task of manually implementing software-based error-detection mechanisms (EDMs) and error-recovery mechanisms (ERMs) in the OS kernel. An unmanageable amount of source code needs to be rewritten and verified. In particular, special care has to be taken that the inherent kernel synchronization is not violated.

However, the conceptual *algorithms* used for memory-error handling remain similar across the different source-code locations in the kernel code, for example, inserting a checksum to protect important data structures. *Only the adaptation* of such algorithms to the particular data structures depends on the respective source-code locations. Modern programming paradigms, such as *Aspect-Oriented Programming* with *AspectC++* [3], offer means to express crosscutting concerns in a single modular source-code file. Thus, an EDM/ERM can be implemented (and verified) *once*, and can then be applied *often* to several OS code locations. The adaptation of the EDM/ERM to the specific OS code locations is carried out by the AspectC++ compiler using its feature to *introspect* the source code at compile time. Thereby, EDM/ERM implementations can even be *reused* for multiple operating systems.

In this paper, I will discuss the resource costs of several aspect-oriented EDM/ERMs applied to the *embedded Configurable operating system (eCos)* [2].

2 Aspect-Oriented Fault-Tolerance Mechanisms

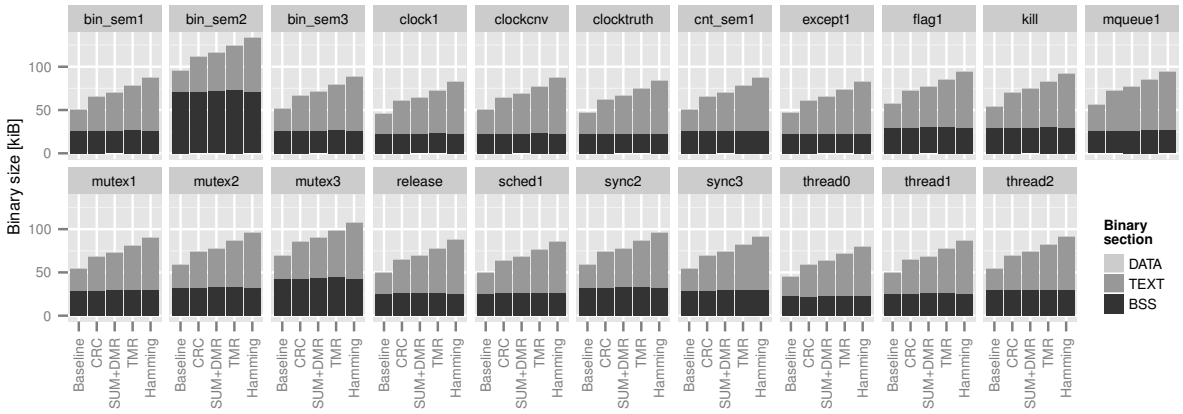
The eCos kernel is hardened by *Generic Object Protection (GOP)* [1] to increase the robustness against memory faults. GOP is implemented in the AspectC++ language, and can be selectively applied to the data structures of object-oriented software. In this study, the process scheduler and thread data structures of eCos are hardened. These data structures are the most important for correct execution of the OS [1]. GOP can be configured at compile time to insert one of the following EDM/ERMs into the protected data structures:

- *Cyclic Redundancy Check (CRC)* code
- CRC code plus object replication (CRC+DMR)
- Two's complement addition checksum plus object replication (SUM+DMR)
- Double replication, yielding *Triple Modular Redundancy (TMR)*
- Hamming Code

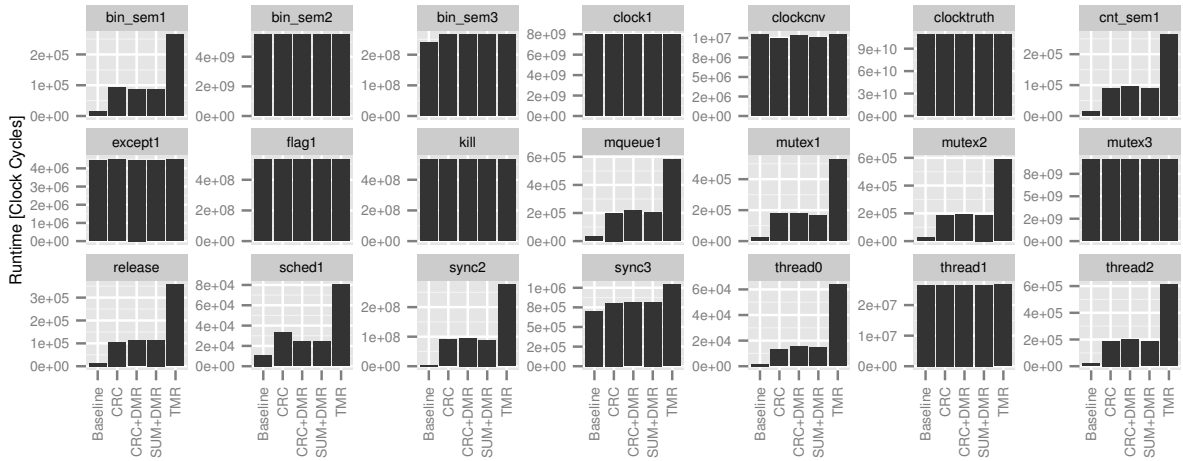
Each EDM/ERMs comes at different static and dynamic costs that are evaluated in the following section.

3 Evaluation

For the resource cost evaluation of the aforementioned EDM/ERMs, I selected the eCos kernel test suite as benchmark. This test suite consists of 21 test programs that use eCos' threads and that are implemented in C++. Figure 1 (a) shows the static binary size of these benchmarks without any protection (Baseline) and with the EDM/ERMs discussed in the previous section. The DATA segment stays almost constant in size for all variants, and the same applies to the BSS segment, which grows at most by 3.6 % compared to the baseline. On the other hand, the code size (TEXT segment) increases clearly: The CRC code adds 58 % on the average, while SUM+DMR and CRC+DMR increase the code size by 74 % and 79 %, respectively. TMR and HAMMING are even more costly, consuming 105 % and 146 % more code size, on the average.



(a) Code-size costs of the different EDM/ERMs



(b) Runtime costs of the different EDM/ERMs

Figure 1: Resource cost evaluation of Generic Object Protection [1]

The runtime costs, measured on a contemporary Intel Core i7-M620 notebook running at 2.666 GHz, are shown in Figure 1 (b). The results can be classified into two categories, depending on the frequency of the process-scheduler invocation: Benchmarks that bombard the scheduler with scheduling requests suffer from high runtime overhead, for example, the `sync2` benchmark. Contrastingly, infrequent scheduler invocations result in negligible runtime overhead compared to the baseline. The total runtime of those benchmarks infrequently invoking the scheduler dominates the whole test suite, so that the aggregated runtime overhead totals at only 0.09 % for the SUM+DMR, CRC and CRC+DMR protection variants, followed by TMR and Hamming with 0.23 % and 1.75 % respectively.

4 Conclusions and Future Work

The evaluation of the resource costs for reliability shows that the costs for the static binary sizes are similar between different programs, and, thus, are easy to estimate. However, the runtime costs vary extremely, depending on the application code. The next step is to model the application behavior, in order to derive an approximation for the runtime overhead in advance. Further, the CPU runtime is coupled with the energy consumption of an embedded device. Hence, the model for the runtime can be refined to estimate the energy consumption as well. Therefore, detailed energy measurements have to be conducted to verify this assumption.

References

[1] Christoph Borchert, Horst Schirmeier, and Olaf Spinczyk. Generative software-based memory error detection and correction for operating system data structures. In *43nd IEEE/IFIP Int. Conf. on Dep. Sys. & Netw. (DSN '13)*. IEEE, June 2013.

[2] Anthony Massa. *Embedded Software Development with eCos*. Prentice Hall Professional Technical Reference, 2002.

[3] Olaf Spinczyk and Daniel Lohmann. The design and implementation of AspectC++. *Knowledge-Based Systems, Special Issue on Techniques to Produce Intelligent Secure Software*, 20(7):636–651, 2007.

[4] Vilas Sridharan and Dean Liberty. A study of DRAM failures in the field. In *Int. Conf. for High Perf. Computing, Networking, Storage and Analysis (SC '12)*, pages 76:1–76:11, Los Alamitos, CA, USA, 2012. IEEE.

KRATOS - A Resource Aware, Tailored Operating System

Markus Buschhoff
Fakultät für Informatik, Lehrstuhl 12
Technische Universität Dortmund
markus.buschhoff@tu-dortmund.de

In an endeavor to design a system that is both highly customizable and resource-aware, KRATOS[1] was constructed as a best-practices approach on small operating system design: It combines the flexibility of software product-lines, the separation of software features using an aspect-oriented language (AspectC++), and new approaches for describing non-functional behaviors within the feature model and the driver level.

1 Introduction

To overcome limitations in resources like memory and energy, the program code of deeply embedded devices needs to be as compact, efficient and specialized as possible. This is a contradiction to the general demand for multi-purpose functionalities in operating systems. A solution to this dilemma are software product lines, which allow to configure a program by defining a subset of interdependent software fragments, so called *features*, to most exactly suit a single purpose.

As a foundation for KRATOS we have chosen AOStubs[2], an aspect oriented operating system which was developed here at LS12 for didactical reasons. AOStubs is very small and efficient, and it is already ported for upcoming platforms within the A4 project (the inBin platform [2] based on TI's MSP430 processors). Additionally, it was developed in AspectC++ [4], an aspect-oriented language which already provides adequate measures to separate features on the source-code level. AOStubs already implements multi-threading

[1] KRATOS is a Resource Aware, Tailored Operating System
[2] AspektOrientiertes STUdenten-BetriebsSystem

using a pre-emptive round-robin scheduler, and a sophisticated two-level interrupt handler system (prologue/epilogue model) for hardware drivers.

2 Product Line Model

To implement a product line model on top of AOStubs, the code first has been restructured to suit our demands. For example, the architecture-dependent code was parted from the generic code, thus allowing us to configure the operating system for multiple hardware platforms. Currently, KRATOS runs on a set of MSP430-based chips, the Raspberry PI platform, and is being ported to the XEN x86 hypervisor using paravirtualized drivers.

In the next step, we constructed a language for feature descriptions. A feature description contains a human-readable description for a certain feature, and a set of dependencies towards other features. A feature itself is identified by a path in a feature-tree, e.g., the feature to set up the "DCO" clock on an MSP430FR-CPU is called "/Architecture/MSP430FR/CLOCK/DCO". All parent nodes in the tree are automatically required features. Features may require other features to be active or inactive, and a feature also can be exclusive within its branch. Further requirements can be described as logical expressions using AND, OR and NOT operators. Additionally, features may contain a set of variables that need to be configured by the user and will be passed to the program code at compile time. Lastly, there are instructions for the compiler on how to handle an activated feature. Here, the C++ and AspectC++ files that implement the feature are defined. The feature description language is based on the extensible markup language (XML).

The reason for creating a new language for product line setups is extensibility. Currently, an extension is planned to declare the resource consumption of a feature within its feature description. This will allow for optimizing a set of features towards a better non-functional behavior.

By now, the feature description language gets translated for an existing product-line software named *kconfig*, which originally was used to configure the Linux kernel. *Kconfig* includes graphical user interfaces for different platforms. Since *kconfig* only has a subset of the KRATOS product line functionalities, we are planning to build an own set of configuration tools based on our feature description language.

The output of a configuration process is a set of variables for the *gnu make* tools and a configuration header file that contains the values of feature-variables as preprocessor macros for the source-code. Finally, a highly specialized version of KRATOS, tailored to the configured architecture, network infrastructure and peripheral devices can be compiled. The configuration of the system can be very detailed: CPU timings, bus settings

for peripheral devices, usage of IRQ[3]/Polling/DMA[4] based transfers for each device and much more can be set-up statically to generate a very compact source-code.

3 State Aware Drivers

Since hardware-drivers are the interface between hardware and software, they are a starting point for describing the non-functional behavior of peripheral hardware within the software layer of a system. A classical driver implementation for an operating system implements an application programming interface (API) for higher software layers, sends signals via hardware buses to its peripheral device, and reacts on interrupt requests (IRQs) triggered by the device. Usually, only the functional behavior of the device is represented in its driver, and this is rarely done in a fashion that resembles some sort of state-model of the device.

To model the non-functional behavior, formal descriptions like finite state-machine models (FSMs) are commonly used. Among these, the use of priced timed automata (PTA) models for resource models is a wide-spread approach [1, 3]. Because the functional and non-functional behaviors of a device are not independent, the non-functional states of such an FSM can often be mapped to functional states of the device, although the mapping is not necessarily one-to-one. The result of such a mapping is a merged state machine that models both the functional and non-functional behavior.

A step towards the prognosis of resource consumptions at run-time would require to know the non-functional states each peripheral device is in at any point of time, and to propagate this information through each software layer. This basically means, a peripheral driver must be either aware of the functional and non-functional states of a device model, or to identify its current state in the merged state machine.

As an example, a simplistic communication device can be taken. The device may have several functional states, such as *idle*, *send* and *receive*, and a non-functional model with annotated costs (e.g., energy consumption) on all send, receive and idle transitions and states. A classical communication driver would not explicitly know what states exist, how they are connected, and what the cost model looks like. It would simply react on send and receive API calls and interrupt requests. A state-aware driver, however, does know the state machine. The API for such a driver would call the driver to take certain transitions to enforce certain states on the device. It could also request the current driver state and the cost model for all states and transitions. An abstraction layer on top of such a driver would then be able to translate standard API calls for sending and receiving data into transition calls, but also was able to request the energy consumption of these calls.

[3] IRQ: Interrupt Request
[4] DMA: Direct Memory Access

An implementation concept for state-aware drivers within the KRATOS operating system can be found in the master thesis of Robert Rapczynski, which will be submitted at the time of the publication of this report.

References

[1] Markus Buschhoff, Christian Günter, and Olaf Spinczyk. A unified approach for online and offline estimation of sensor platform energy consumption. In *Proceedings of the 8th International Wireless Communications and Mobile Computing Conference (IWCMC '12)*, pages 1154 –1158, August 2012.

[2] Jan Emmerich, Moritz Roidl, Tobias Bich, and Michael ten Hompel. Entwicklung von energieautarken, intelligenten ladehilfsmitteln am beispiel des inbin. *Logistics Journal*, 2012, 2012.

[3] C. Seceleanu, A. Vulgarakis, and P. Pettersson. Remes: A resource model for embedded systems. In *Engineering of Complex Computer Systems, 2009 14th IEEE International Conference on*, pages 84–94, June 2009.

[4] Olaf Spinczyk. The Home of AspectC++. `http://www.aspectc.org`. [Online; accessed 18-November-2014].

Optimization of Energy Efficiency in LTE Networks

Dennis Kaulbars
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
dennis.kaulbars@tu-dortmund.de

This report gives an introduction on how to optimize the energy consumption of LTE-based mobile end devices from the network side. For this purpose, the report describes either an energy-efficient resource scheduling strategy for the uplink as also an outlook about a possible protocol parameter optimization concerning the tail time, which is part of the discontinuous reception scheme of LTE.

1 Introduction

The world of modern mobile communication is growing faster than ever. The market of end devices offers modern smartphones with even-larger displays, thinner bodies and ever-growing CPU performance. This also encourages network operators to roll-out mobile technologies as LTE providing high data rates and low latencies optimized for innovative real-time services and data-hungry applications such as video and audio streaming. This also means that the devices' need for energy also increases, which place higher demands on the batteries implemented inside the devices. As a consequence, even new smartphones suffer from short battery lifetimes of only a few hours.Beside manufacturers, network operators can also try to counter this problem by adapting the resource scheduling and protocol behavior of LTE in order to optimize the energy efficiency of active end devices inside the mobile cell. This report describes two approaches on how the energy-efficiency of smartphones can be enhanced by LTE networks: an energy-efficient scheduling approach called CoPoMo scheduler and a parameter optimization of the tail times, which are part of the discontinuous reception (DRX) scheme.

2 Energy-Efficient Scheduling of Uplink Data Traffic

The proposed CoPoMo scheduler focuses on the optimization of energy efficiency receptively the enlargement of communication time for a group of UEs sending data in the uplink (from device to base station), which means that the scheduler tries to allocate the available radio resources in a way that all active UEs are able to take part on the communication process as long as possible. We use CoPoMo [1] to formulate the statement that the energy consumption of an UE can be optimized by allocating as much resource blocks (RBs) in the frequency domain to the UE as possible, because then the UE is able to transmit a pre-defined amount of data in a shorter time and afterwards it can switch to the energy-saving *IDLE* mode. The scheduling scheme derived from this approach performs the following steps (for one transmit time interval (TTI), N_{total} RBs have to be allocated to n UEs):

1. Sort UEs to be scheduled by remaining battery capacity in a descending manner (battery capacity is shared with eNode B on application layer)
2. For each UE, calculate a weight w_k, $k \in \{1..n\}$ describing the priority of each device, which means that the UE with lowest remaining battery capacity gets highest priority: $w_k = \frac{\sum_{i=k}^{n} \frac{1}{i}}{n}$
3. Calculate the number of provided RBs for each UE in current TTI:

$$N_k = \begin{cases} \lfloor N_{total} \cdot w_k \rfloor, & k \in \{1..(n-1)\} \\ N_{total} - \left(\sum_{k=1}^{n} \lfloor N_{total} \cdot w_k \rfloor \right), & k = n \end{cases}$$

4. Allocate N_k contiguous RBs to UE_k. The next free RB with the highest channel quality indicator (CQI) can be chosen as the center of the group of RBs for corresponding UE

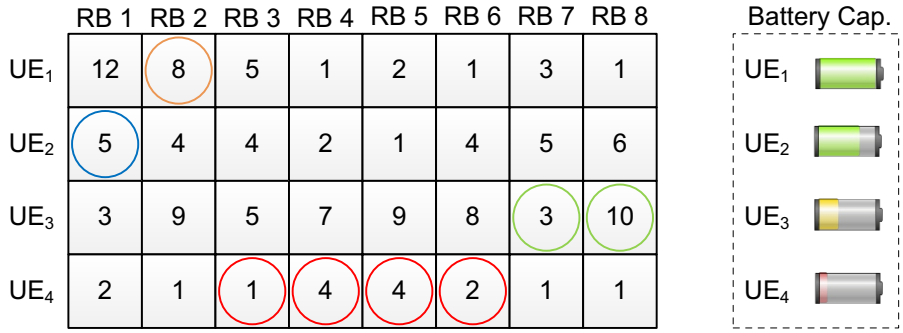


Figure 1: Illustration of CoPoMo scheduler for allocating 8 Resource Blocks to 4 Devices

Figure 1 shows an illustration of allocating $N_{total} = 8$ RBs to a group of $n = 4$ devices. UE_4 has the lowest, UE_3 the highest remaining battery capacity of the devices inside the

group. The scheduler allocates the highest number of RBs to UE_4, the lowest number of available RBs to UE_1. The amount of allocated resources is inversely proportional to the remaining battery capacity of the UE, which meets the statement concerning the energy-efficiency of the end devices and the aim of the scheduler to enhance the group communication time.

The proposed CoPoMo scheduler has the following properties:

- UEs with lower remaining battery capacity gets a higher scheduling priority in terms of energy efficiency than UEs with a higher remaining battery capacity
- The scheme allocates only contiguous RBs meeting the contiguity constraint for uplink scheduling schemes
- By using CQI as a basis for choosing the center of a group of RBs, the scheduler tries to use the limited spectrum efficiently

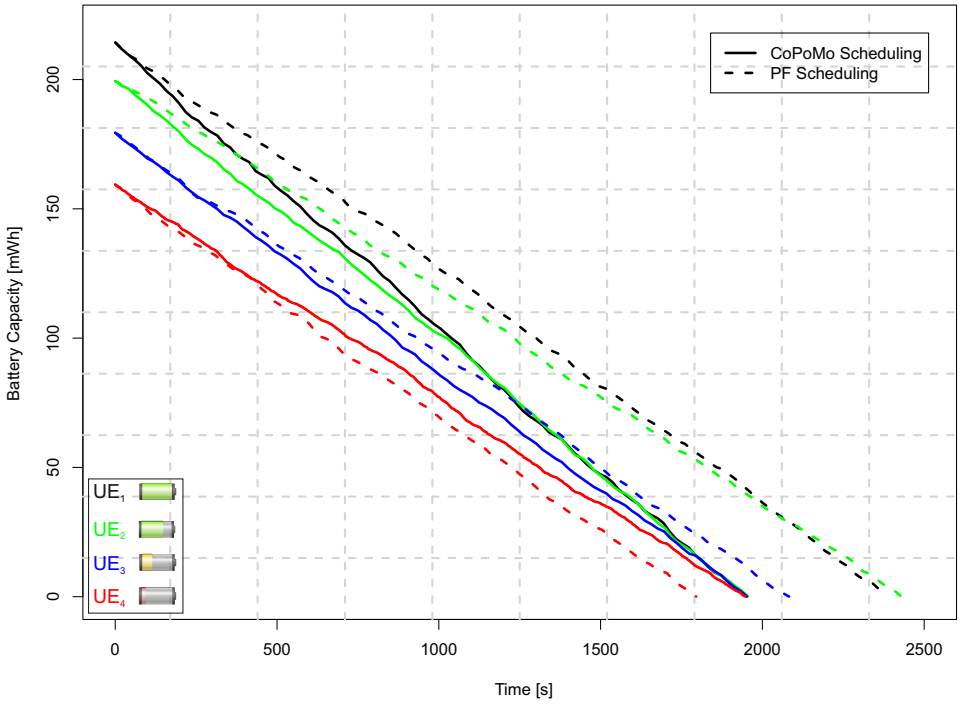


Figure 2: Effect of CoPoMo scheduler on battery capacity

The effect of the CoPoMo scheduler for a scenario with four UEs with different remaining battery capacities is illustrated in Figure 2. The traditional proportinal fair (PF) scheduler does not take care about the remaining battery capacity of the UEs which have to be scheduled and, as a consequence, the discharge curves of the UEs decline nearly parallel. On the other hand, the CoPoMo scheduler takes the remaining battery capacity of each UE as a basis for the resource allocation process. This results in the fact that the amount of RBs for each UE per TTI is inverse proportional to the remaining battery

capacity and due to the fact that it is more energy efficient to allocate more RBs in the frequency than in the time domain, the discharge curves of higher prioritized UEs are flatter than the ones of lower prioritized UEs. This results in the effect that all discharge curves converge against each other, which means that after some time, all UEs will have a similar remaining battery capacity. From this time on, the behavior of the CoPoMo scheduler is similar to the PF scheduler, since all UEs gets the same priority in the scheduling process leading to similar shapes of the corresponding discharge curves. In the next phase of the A4 research project, the CoPoMo scheduler will be analyzed quantitatively under different scenarios with single and multiple cell environments as well as compared to other scheduling schemes.

3 Outlook: Tail Times and Energy Efficiency

The Discontinious reception (DRX) concept of LTE is used to minimize the power consumption of an UE by introducing micro-sleeps between several sending processes. For this purpose, the UE's radio unit distinguishes between two states: The first state called *RRC_CONNECTED* is used if the UE is active, whereas the *RRC_IDLE* state is used if the UE is inactive [2]. When the UE is in the *RRC_CONNECTED* mode and a transmission has been finished, the UE waits for a pre-defined duration called *tail time* before it switches back to the *RRC_IDLE* state. This method enables the UE to fast resume a transmission if another packet should be sent. It was already found out that the duration of the tail time has a huge impact on the device's energy consumption [3]. In order to quantitatively analyze this impact in combination with different devices, transmission frequencies and service rates, the DRX scheme can be adapted to CoPoMo in order to be able to carry out the proposed studies.

References

[1] B. Dusza, C. Ide, L. Cheng, and C. Wietfeld. CoPoMo: a context-aware power consumption model for LTE user equipment. *Transactions on Emerging Telecommunications Technologies*, 24(6):615–632, 2013.

[2] M. Gupta, S.C. Jha, A.T. Koc, and R. Vannithamby. Energy impact of emerging mobile internet applications on LTE networks: issues and solutions. *Communications Magazine, IEEE*, 51(2):90–97, February 2013.

[3] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck. An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, SIGCOMM '13, pages 363–374, New York, NY, USA, 2013. ACM.

New Random Frequency Hopping Policy for Femtocell Networks

Markus Putzke
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
Markus.Putzke@tu-dortmund.de

Typical interference mitigation of femtocells is based on self-organizing resource allocations. In this way, the air interface needs to be periodically scanned and radio parameters are appropriately tuned. There are situations where such interference scanning is not applicable or where it results in high complexity. Exemplary situations are femtocell initializations and mobile femtocells where the interference environment changes very fast. In these situations we propose to use Random Frequency Hopping inside femtocells. Existing Random Frequency Hopping in the context of Orthogonal Frequency Division Multiple Access is restricted to carrier frequencies defined on a discrete resource block grid such that only every 12th subcarrier can be allocated. In this work, we introduce a new Random Frequency Hopping policy which allows to select each subcarrier as a random carrier frequency. This enables a reduction of the interference by 50% compared to existing approaches. In order to quantify these gain, we present an analytical model and validate its results by simulations.

1 Motivation

The data traffic in mobile radio networks is dramatically increasing and exceeding nearly every forecast. As the provisioning of more bandwidth, higher modulations or new transmission technologies like Orthogonal Frequency Division Multiple Access (OFDMA) is no longer able to sustain this enormous traffic demand, the deployment of small cells is considered as a very promising approach. Due to restricted coverage areas of small cells,

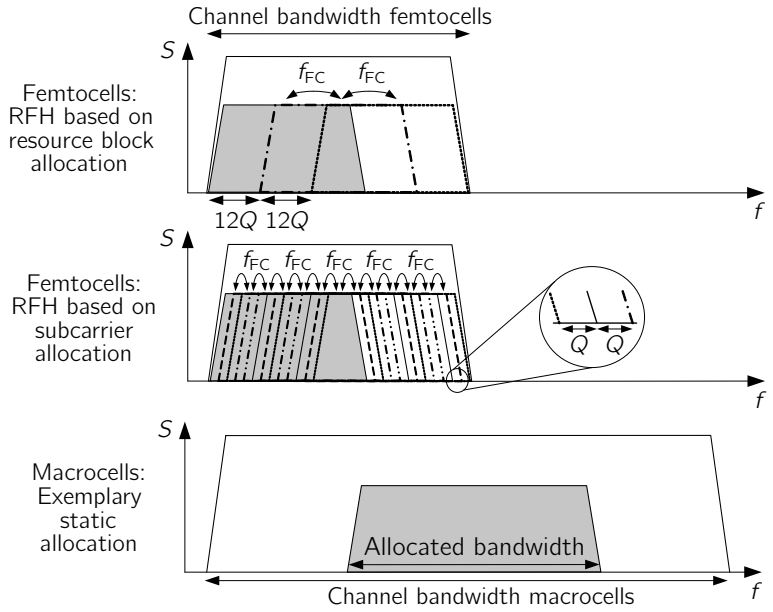


Figure 1: Random Frequency Hopping based on resource block and subcarrier allocation

the area spectrum reuse can be considerably increased. One common type of small cells are femtocells. In order to reduce capital expenditure from a service providers point of view, femtocells are typically deployed by users such that their activity in time and space is random. Femtocells reuse licensed spectrum in order to avoid new receiver structures. The combination of frequency reuse and random activity regarding femtocells introduces interference, which can only be reduced efficiently by self-organizing methods.

Self-organizing interference mitigation is based on sensing the actual radio environment and tuning frequencies as well as transmit power accordingly. Hence, knowledge of the actual interference environment is required. However, there are situations where the actual interference environment is unknown, like in femtocell initialization phases of access points after power-up. Moreover, there are situations where knowledge of the actual interference environment requires high complexity of the transmitter and the receiver, as for fast moving femtocells. Regarding the latter, the radio environment changes very fast, such that channel quality reports of users, which capture the interference environment, are already outdated when they are received by access points.

2 New Random Frequency Hopping Approach

Motivated by the described scenarios, we propose to use Random Frequency Hopping (RFH) for spectrum allocation within femtocells. RFH means that the carrier frequency changes randomly from one OFDMA symbol/ subframe to another within the channel bandwidth. Existing RFH approaches for OFDMA femtocells are restricted to carrier

frequencies which are selected from a resource block frequency grid. As shown in the upper part of Fig. 1, this implies that only every 12th subcarrier can be used as random carrier frequencies. As a consequence, existing RFH is only able to average interference across different scenarios. It suffers from heavy interference when the traffic load becomes high. Due to this, we introduce a new RFH policy which outperforms the existing one. In contrast to known RFH, the new approach allows to select each subcarrier as a random carrier frequency, cf. middle part of Fig. 1. The following analysis reveals that RFH based on subcarrier allocation is not only able to average the interference, but also to reduce the interference compared to existing RFH. Note that this is achieved without additional knowledge. The idea behind the new policy is based on our works [1–3].

3 Analytical Model and Performance Analysis

The interference reduction of the new RFH policy can be best understood from analytical modeling. Using the models derived in [2] and assuming a BPSK transmission, the interference symbols i_s of both RFH approaches are given by

$$i_s \approx \sum_{\mathrm{MC}} A \mathrm{e}^{\xi_i/2} z_{d'} \cos(2\pi Q C (s - d')) \,. \tag{1}$$

where A represents the path loss and small scale fading, ξ_i models log-normal shadowing, z_d are the transmitted symbols, d is the symbol index of the interferer, s the symbol index of the femtocell, Q the subcarrier spacing and C the cyclic prefix. The summation in (1) considers the interference from different macrocells. Since the subcarrier grid of both RFH approaches as well as that of the macrocell in Fig. 1 is the same, the interference symbols i_s are only caused by one subcarrier d' of each macrocell. This subcarrier is given by the carrier frequency difference f_x between the macro- and the femtocell: $f_x = f_{\mathrm{MC}} - f_{\mathrm{FC}} = Q(s - d')$. For RFH based on resource-block allocation, f_x is a multiple of resource blocks $f_x = 12mQ$, $m \in \mathbb{Z}$. From this and a cyclic prefix which is a multiple of 1/24, the interference power of existing RFH approaches is only a function of the number of colliding subcarriers

$$P_i = \mathbb{E}\left\{i_s^2\right\} = \sum_{\mathrm{MC}} \frac{P_{\mathrm{BS}}}{L_{\mathrm{MC}} N_{\mathrm{FC}}} \sum_{s} \delta_{sd'} \,, \tag{2}$$

where P_{BS} is the transmit power of base stations, L_{MC} the path loss to surrounding macrocells, N_{FC} the number of allocated femtocell subcarriers and $\delta_{sd'}$ the Kronecker delta. In case of the new RFH approach the carrier frequency difference is an arbitrary multiple of the subcarrier spacing $f_x = mQ$. Due to this, the interference power of half the random carrier frequencies disappears, although subcarriers collide for all of them in the same manner as in (2)

$$P_i = \begin{cases} 0, & f_x = (2n+1)Q \\ \sum\limits_{\mathrm{MC}} \frac{P_{\mathrm{BS}}}{L_{\mathrm{MC}} N_{\mathrm{FC}}} \sum\limits_{s} \delta_{sd'}, & f_x = 2nQ \end{cases} , \; n \in \mathbb{Z}. \tag{3}$$

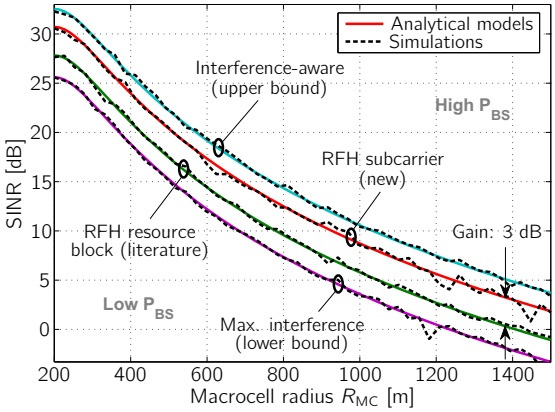


Figure 2: SINR of femtocell users as a function of the macrocell radius

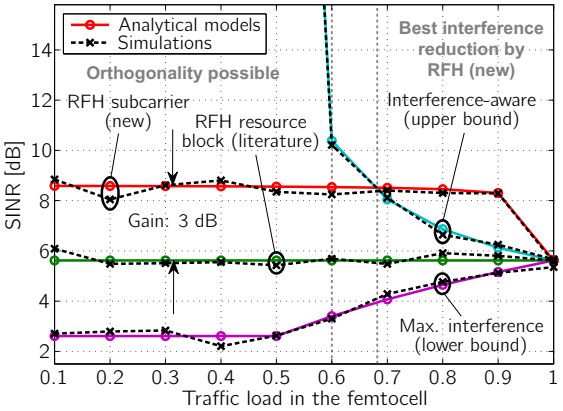


Figure 3: SINR of femtocell users as a function of the femtocell traffic load

The reason is that BPSK modulated symbols are assumed, i.e. the femtocell demodulator only considers the real part of the interference symbols in the constellation diagram. Half of the interference symbols from the new RFH approach are imaginary, whereas those of RFH from literature are always real. In other words, the new RFH policy is able to shift half of the interference symbols to an orthogonal domain. Fig. 2 and Fig. 3 show the SINR of femtocell users for both RFH approaches as well as systems having perfect knowledge of the interference (Interference aware) and systems where femto- and macrocells allocate the same subcarriers (Max. interference). Fig. 2 is a function of the macrocell size, whereas larger cell sizes result in higher transmit powers of interfering base stations, such that the SINR is monotonously decreasing. In contrast to that, Fig. 3 is a function of the femtocell traffic load. As long as the traffic load is lower than 100%, the new RFH policy increases the SINR by 3dB. This is due to an interference power reduction of 50%, since half of the random carrier frequencies from the new RFH policy result in orthogonal interference symbols with respect to BPSK demodulation, cf. (3).

References

[1] M. Putzke and C. Wietfeld. Self-Organizing Ad Hoc Femtocells for Cell Outage Compensation Using Random Freqeuncy Hopping. In *IEEE Symposium on Personal, Indoor and Radio Communications*, pages 315–320, Sydney, Australia, Sept. 2012.

[2] M. Putzke and C. Wietfeld. Self-Organizing Fractional Frequency Reuse for Femtocells Using Adaptive Frequency Reuse. In *IEEE Wireless Communications and Networking Conference*, pages 434–439, Shanghai, China, April 2013.

[3] S. Rohde, M. Putzke, and C. Wietfeld. Ad Hoc Self-Healing of OFDMA Networks Using UAV-Based Relays. *Elsevir Journal on Ad Hoc Networks*, 11(7):1893–1906, Sept. 2013.

Subproject A5
Exchange and Fusion of Information under Availability and Confidentiality Requirements in MultiAgent Systems

Gabriele Kern-Isberner Joachim Biskup

Component-based Agentmodels - concept and implementation

Patrick Krümpelmann
Faculty of Computer Science
Technische Universität Dortmund
patrick.kruempelmann@tu-dortmund.de

A variety of logical formalisms with different expressivity and computational properties have been developed for knowledge representation with the agent paradigm in mind [2]. Especially non-monotonic formalisms are designed to deal with incomplete information and to enable an agent to act in uncertain environments. Moreover, the field of research on belief change has been working for over 25 years already on solutions on how to change an agent's beliefs in the light of new information [3]. Yet, very little of the approaches developed in these two fields of research are available in actual multiagent frameworks. Our concept of component based agents and its realisation in the Angerona framework are designed to reduce this gap, to support the development of knowledge based, i. e. epistemic, agents based on logical formalisms for knowledge representation and reasoning, and to support the use of belief change operators based on belief change theory. Moreover, it facilitates the development of divers agents with respect to their architecture and knowledge representation. It allows the formation of multiagent systems comprising heterogeneous agents which interact by communicating or in a common simulated environment.

1 Concept of Compound Agents

Angerona agents are based on a concept of hierarchical, component-based agent models with the goal of capturing a variety of agent architectures in a flexible and extensible way. In the following we give an overview of the main concepts of it. In this, a general

agent instance is a tuple $(\mathcal{K}, \xi)$ comprising of an epistemic state $\mathcal{K} \in \mathcal{L}_{ES}$ from a given language $\mathcal{L}_{ES}$ and a functional component $\xi = (\circ, \mathsf{act})$. Further, we assume the set of possible actions Act and perceptions Per to be given. These might, for instance, be speech acts that are interchanged by the agents. Then, we require the operators of the functional component to be of the following types:

$$\circ : Per \times \mathcal{L}_{ES} \to \mathcal{L}_{ES} \text{ and } \mathsf{act} : \mathcal{L}_{ES} \to Act.$$

The language of the epistemic state might be a logical language, e. g. an answer set program or a conditional belief base, or a Cartesian product of (logical) languages, e. g. to represent the BDI components of an agent by the language $\mathcal{L}_B \times \mathcal{L}_D \times \mathcal{L}_I$. The epistemic state of an agent contains representations of its background knowledge about how the world works, and information coming from its perceptions, as well as its goals and know-how, and potentially more. The functional component of an agent consists of a change operator $\circ$, which adapts the current epistemic state of the agent upon reception of a perception, and an action operator act, which executes the next action based on the current epistemic state. The change of the epistemic state might involve different types of reasoning, such as non-monotonic reasoning, deliberation and means-ends reasoning. These are partially or completely based on logical inference. This means, that an agent's behavior is realized in parts by the functional component, and in parts by the knowledge representation and reasoning based on the epistemic state. To capture these different types of agents we consider the epistemic state as well as the functional component to consist of hierarchical components.

A compound epistemic state is a component, which again can either be atomic or compound. An atomic component $\mathcal{C}_a$ is an element from the components language $\mathcal{L}_{\mathcal{C}_a}$, e. g. a belief base BB from the language $\mathcal{P}(\mathcal{L}_{BB})$, such as an OCF-base or an answer set program. Belief operators of the form $Bel : \mathcal{P}(\mathcal{L}_{BB}) \to \mathcal{P}(\mathcal{L}_{BS})$ are applied by other operators to belief bases to determine the current belief set for it. A compound component is a tuple of components, $\mathcal{C} = \langle \mathcal{C}_1, \ldots, \mathcal{C}_n \rangle$, and each component is an element of its language such that the language of a compound component is a cartesian product of languages: $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}_1} \times \cdots \times \mathcal{L}_{\mathcal{C}_n}$. In particular, each component can potentially have a different representation. The interaction of the components is realized by the functional component of the agent. In particular, for an epistemic state $\mathcal{K} \in \mathcal{L}_{ES}$ and functional component $\xi = (\circ, \mathsf{act})$ the change operator $\circ : Per \times \mathcal{L}_{ES} \to \mathcal{L}_{ES}$ can be realized by a single function or by a composition of explicit sub-functions. In the latter case sub-functions are applied to the epistemic state in sequential order. Each sub-function modifies a single component or a set of components of the epistemic state. The next function operates on the epistemic state that results from the modifications of the previous functions. This concept realizes the idea of an agent cycle. Typical agent cycles as the one of the BDI architecture can be easily formalized. The resulting hierarchical agent model with compound epistemic state and compound functional component is illustrated in Figure 1. Formally, a compound functional component consists of a change operator

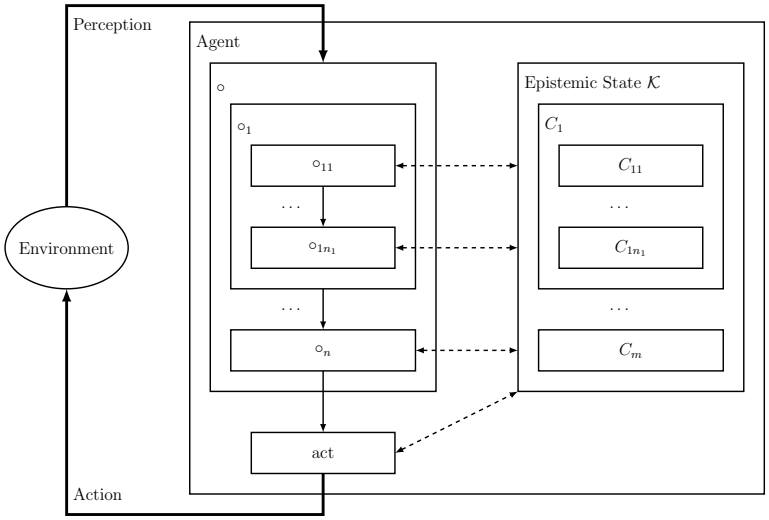


Figure 1: Hierarchical agent model

$\circ$ that is a composition of operators, i. e. $\circ =_{def} \circ_1 \cdot \ldots \cdot \circ_{n'}$, and an action function act. More details about the concept of compound agents can be found in [4].

2 Implementation of Compound Agents

Angerona agents consist of *agent components* which can be *epistemic components*, i. e. belief bases and associated operators, and other data components, or *functional components*, i. e. operators used for the agent cycle. Logic based components are based on the *belief base plug-in*. Operators for the agent cycle are based on the *operator plug-in*. The class diagram in Figure 2 illustrates the realization of the conceptual model in the Angerona framework. An Angerona agent contains an epistemic state and a list of operators. An epistemic state consists of agent components. One type of agent components are belief bases which are defined via a belief base plug-in. The belief base plug-in implements the interfaces of the Tweety library [5], in particular those for a belief base, a formula, a revision operator and a belief operator. Different belief operators might be available for the same formalism. Different agents might use the same knowledge representation formalism but different belief operators, and each agent might use different belief operators in different situations. We use, for example, families of belief operators that are ordered by their credulity, e. g. skeptical reasoning vs. credulous reasoning, in the setting of secrecy preserving agents [1]. More on the Angerona framework can be found

in [4]. Angerona is open source and available on *github*[1], as is Tweety on *sourceforge*[2].

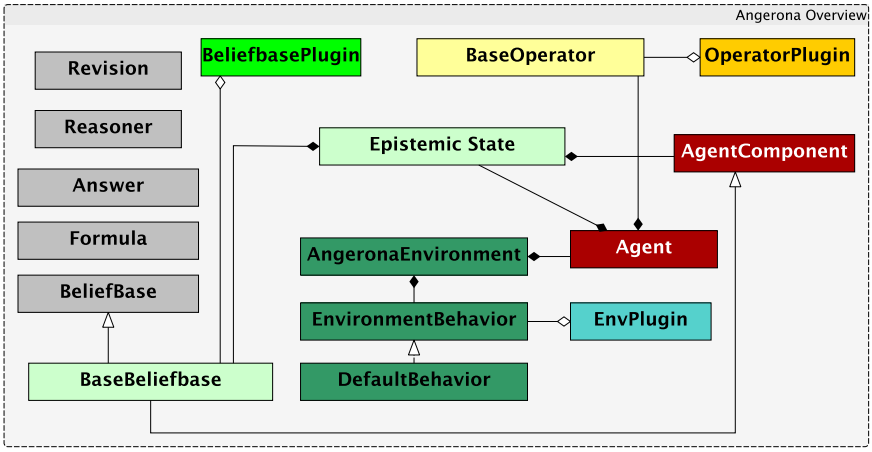


Figure 2: Angerona agent simplified class diagram

References

[1] Joachim Biskup, Gabriele Kern-Isberner, Patrick Krümpelmann, and Cornelia Tadros. Reasoning on secrecy constraints under uncertainty to classify possible actions. In Christoph Beierle and Carlo Meghini, editors, *Foundations of Information and Knowledge Systems*, volume 8367 of *LNCS*, pages 97–116. Springer, 2014.

[2] Ronald J. Brachman and Hector J. Levesque. *Knowledge Representation and Reasoning*. Elsevier and Morgan Kaufmann Publishers, 2004.

[3] Eduardo Fermé and Sven Hansson. AGM 25 years. *Journal of Philosophical Logic*, 40:295–331, 2011. 10.1007/s10992-011-9171-9.

[4] Patrick Krümpelmann, Tim Janus, and Gabriele Kern-Isberner. Angerona - a flexible multiagent framework for knowledge-based agents. In Nils Bulling, editor, *Proceedings of the 12th European Conference on Multi-Agent Systems*, volume to appear of *Lecture Notes in Artificial Intelligence*. Springer, 2014.

[5] Matthias Thimm. Tweety - a comprehensive collection of java libraries for logical aspects of artificial intelligence and knowledge representation. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR'14)*, July 2014.

[1] `https://github.com/Angerona`
[2] `http://sourceforge.net/projects/tweety/`

Efficient Computation of Inference-Proof Materialized Views Based on Weakenings

Marcel Preuß
Lehrstuhl für Informationssysteme und Sicherheit
Technische Universität Dortmund
preuss@ls6.cs.tu-dortmund.de

During the last year the main focus of my research was on the development of a novel approach provably protecting sensitive information specified by a confidentiality policy – even if an adversary employs inferences to deduce confidential information. The protection mechanism of this novel approach basically relies on the idea of weakening a database instance by means of disjunctions. This research first led to a purely generic algorithm providing all basic steps needed to compute such a weakening in a declarative way. Then, a concrete algorithm implementing an availability-maximizing flavor of this generic approach on the operational level has been proposed and this implementation has been shown to be highly efficient.

Nowadays, data publishing is ubiquitous. Governments are often legally obliged to provide data about matters of public concern, companies release project-related data to partners and even in most peoples' private lifes the sharing of data plays a major role. But usually only certain portions of some data are appropriate for being shared, as data often contains sensitive information. This applies in particular to data containing personal information, as surveyed in [8, 14].

In the area of relational databases the logic-oriented framework of Controlled Interaction Execution (CIE) can assist a database owner in ensuring that each of his interaction partners can only obtain a so-called "inference-proof view" on the owner's data [3]. An inference-proof view does not contain information to be kept confidential from the respective partner, even if this partner is an adversary trying to deduce confidential information

by drawing inferences based on his a priori knowledge and his general awareness of the protection mechanism. An example of such a protection mechanism creating inference-proof materialized views – which are suitable for data publishing – by modifying a minimum number of truth-values of database tuples has been developed in [6].

My research during the last year dealt with the development of a novel approach within the framework of CIE creating *inference-proof materialized views* suitable for data publishing and thereby *provably* enforcing a confidentiality policy without modifying any truth-values: instead, harmful database tuples are replaced by weaker knowledge in the form of disjunctions formed by ground atoms stemming from the confidentiality policy (each of which logically represents a database tuple). These disjunctions contain only true information, but weaken an adversary's possible gain in information such that the adversary is provably not able to infer protected sensitive information.

The results of this research – which are accepted for publication [4] and which will be presented at ICISS 2014 conference in Hyderabad, India – first describe a *purely generic* approach in the sense that non-trivial disjunctions of any length ≥ 2 might be employed. Then, a possible instantiation of this generic approach is presented, which aims at *maximizing availability* in the sense that only disjunctions of length 2 are seen to be admissible. For this instantiation an algorithmic treatment is given, which is based on graph clustering realized with the help of well-known maximum matching algorithms (cf. [9]). Note that this availability maximizing instantiation fully specifies the generic approach except for an admissibility criterion expressing which subsets of potential secrets of the confidentiality policy might possibly form a disjunction. This criterion should be tailored to the needs of each specific application and can be easily specified by employing query languages known from relational databases (cf. [1, 12]).

To be able to fully implement the availability-maximizing flavor to experimentally demonstrate its high efficiency – which can be even raised by computing maximum matchings with the help of a heuristic (cf. [11]) resulting only in a slight loss of availability – an example for such an admissibility criterion called *interchangeability* is provided and evaluated. Interchangeability admits only disjunctions formed by ground atoms which all pairwise differ in the same single position and do not differ in any other position. This local restriction of distortion preserves definite information about all but one position of each ground atom and *generalizes* each distorted value to a wider set of possible values. Moreover, extensions of the generic approach dealing with policies (and hence disjunctions) of existentially quantified atoms and also coping with a basic kind of an adversary's a priori knowledge in the form of ground atoms are outlined.

As an adversary is aware of which values are weakened by simply considering the disjunctions, particular attention must be paid to eliminate so-called *meta-inferences* (cf. [3]). A deduction of sensitive information is called a meta-inference, if it is obtained by excluding all possible alternative settings, under which this sensitive information is *not* valid, by

simulating these alternative settings as inputs for the algorithm generating the inference-proof views and by being able to distinguish the outputs resulting from each alternative setting from the published one. In the developed approach meta-inferences are eliminated by imposing a total order on the sentences of weakened instances.

My research for the next year will deal with those extensions of the above presented approach which are mentioned as future work in [4]. Until now, the developed approach is only able to handle an adversary's a priori knowledge which is restricted to ground atoms. But this rather simple kind of a priori knowledge does not allow for modeling commonly used semantic database constraints such as the well-known classes of Equality Generating Dependencies and Tuple Generating Dependencies (cf. [1]). Examples for achieving inference-proofness under versatile subclasses of these database constraints are given in [5, 6] and should be transferred to the current approach.

As the weakening of knowledge with the help of disjunctions corresponds to the generalization of certain values to a wider set of possible values, the developed approach is related to the well-known approaches of k-anonymization and ℓ-diversification [10, 13]. These approaches aim at preventing the re-identification of individuals based on so-called quasi-identifiers, which describe some of the individuals' properties, by generalizing these quasi-identifiers. So, it might be desireful to elaborate the connection between the developed weakening approach the approaches of k-anonymization and ℓ-diversification.

Moreover, the definition of inference-proofness underlying the approach so far only guarantees the existence of at least one "secure" alternative instance from an adversary's point of view. But in terms of enhancing confidentiality it might be desirable to strengthen this definition to always guarantee a certain number k of different "secure" alternative instances. Considering the basic ideas developed so far, this can be achieved by increasing the length of disjunctions. But until now, no algorithms constructing availability-maximizing clusters of size ≥ 3 are developed on the operational level. As the results known about the complexity of k-anonymization [2, 7] suggest that the computation of weakenings always guaranteeing more than one alternative instance is not possible in polynomial time, a formal analysis of the complexity of such an extended weakening algorithm might also be desirable.

References

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, Reading, 1995.

[2] Gagan Aggarwal, Tomás Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Anonymizing tables. In Thomas Eiter and Leonid Libkin, editors, *10th International Conference on Database Theory, ICDT 2005*, volume 3363 of *LNCS*, pages 246–258, Heidelberg, 2005. Springer.

[3] Joachim Biskup. Inference-usability confinement by maintaining inference-proof views of an information system. *International Journal of Computational Science and Engineering*, 7(1):17–37, 2012.

[4] Joachim Biskup and Marcel Preuß. Inference-proof data publishing by minimally weakening a database instance. In *10th International Conference on Information Systems Security, ICISS 2014*. to appear.

[5] Joachim Biskup and Marcel Preuß. Database fragmentation with encryption: Under which semantic constraints and a priori knowledge can two keep a secret? In Lingyu Wang and Basit Shafiq, editors, *Data and Applications Security and Privacy XXVII – 27th Annual IFIP WG 11.3 Conference, DBSec 2013*, volume 7964 of *LNCS*, pages 17–32, Heidelberg, 2013. Springer.

[6] Joachim Biskup and Lena Wiese. A sound and complete model-generation procedure for consistent and confidentiality-preserving databases. *Theoretical Computer Science*, 412(31):4044–4072, 2011.

[7] Jeremiah Blocki and Ryan Williams. Resolving the complexity of some data privacy problems. In Samson Abramsky, Cyril Gavoille, Claude Kirchner, Friedhelm Meyer auf der Heide, and Paul G. Spirakis, editors, *37th International Colloquium on Automata, Languages and Programming, ICALP 2010, Part II*, volume 6199 of *LNCS*, pages 393–404, Heidelberg, 2010. Springer.

[8] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Data Mining and Knowledge Discovery. CRC Press, Boca Raton, FL, 2011.

[9] Bernhard Korte and Jens Vygen. *Combinatorial Optimization: Theory and Algorithms*. Algorithms and Combinatorics. Springer, Heidelberg, 5th edition, 2012.

[10] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[11] Jakob Magun. Greedy matching algorithms: An experimental study. *ACM Journal of Experimental Algorithmics*, 3(6), 1998.

[12] Raghu Ramakrishnan and Johannes Gehrke. *Database Management Systems*. McGraw-Hill, Boston, MA, 3rd edition, 2003.

[13] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[14] Raymond Chi-Wing Wong and Ada Wai-Chee Fu. *Privacy-Preserving Data Publishing – An Overview*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, San Rafael, CA, 2010.

Subproject B1
Analysis of Spectrometry Data with Restricted Resources

Sven Rahmann Jörg Ingo Baumbach

An Online Peak Extraction Algorithm for Ion Mobility Spectrometry Data

Dominik Kopczynski
Bioinformatics, Computer Science XI
TU Dortmund University, Germany
Dominik.Kopczynski@tu-dortmund.de

Ion mobility (IM) spectrometry (IMS), coupled with multi-capillary columns (MCCs), has been gaining importance for biotechnological and medical applications because of its ability to measure volatile organic compounds (VOC) at extremely low concentrations in the air or exhaled breath at ambient pressure and temperature. Ongoing miniaturization of the devices creates the need for reliable data analysis on-the-fly in small embedded low-power devices e.g. the Raspberry Pi. We present the first fully automated online peak extraction method for MCC/IMS spectra. Each individual spectrum is processed (with a time restriction of 100 ms) as it arrives, removing the need to store a whole measurement of several thousand spectra before starting the analysis, as is currently the state of the art.

Introduction

Ion mobility (IM) spectrometry (IMS), coupled with multi-capillary columns (MCCs), MCC/IMS for short, has been gaining importance for biotechnological and medical applications. With MCC/IMS, one can measure the presence and concentration of volatile organic compounds (VOCs) in the air or exhaled breath with high sensitivity; and in contrast to other technologies, such as mass spectrometry coupled with gas chromatography (GC/MS), MCC/IMS works at ambient pressure and temperature. A detailed description of the MCC/IMS device and especially the data is given in [1]. State-of-the-art software only extracts peaks when the whole measurement is available, which may take up to 10 minutes because of the pre-separation of the analytes in the MCC. However, storing the whole measurement is not desirable or possible when the memory and CPU power is restricted. A method is introduced that extracts peaks and estimates a parametric representation while the measurement is being captured. During the first step, a current spectrum is beeing modeled as a sum

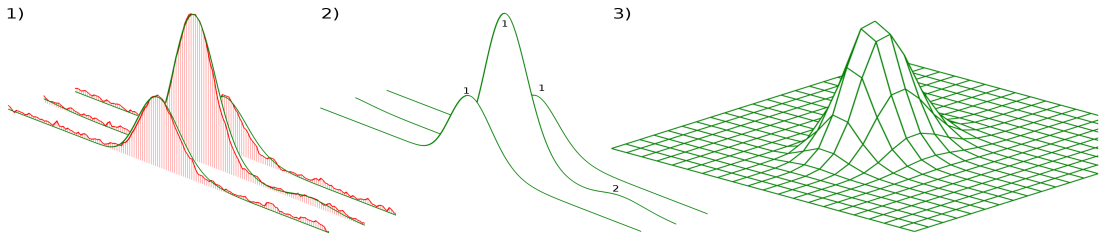


Figure 1: Three steps for online peak extraction: (1) reducing spectrum into set of weighted statistical distributions; (2) build chains by aligning consecutive spectra; (3) estimate all over parameters for two-dimensional peaks individually.

of weighted statistical distributions. This step is referred to *spectrum reduction*. In the second step, consecutive reduced spectra are aligned, in order that models with similar modes are put together into a peak chain. Finally in step three, every peak chain is processed to estimate parameters for a two-dimensional model. Figure 1 illustrates all steps.

Spectrum Reduction

The algorithm scans for peaks, starting at the left end ofspectrum S, by sliding a window of width ρ across S and fitting a quadratic polynomial to the data points using non-linear least square [3, Chapter 10] within the window. Assume that the window starts at index β and ends at $\beta + \rho$, the latter being not included. The value of ρ is determined by the grid opening time d_{grid}, the maximum drift time of the spectrum D_{last} and the number of data points in the spectrum, $\rho := d_{\text{grid}}/D_{\text{last}} \cdot |D|$ data point units. Let $f(x;\theta) = \theta_2 x^2 + \theta_1 x + \theta_0$ be the fitted polynomial. A window is being considered as a *peak window*, if the following conditions are fulfilled:

- the extreme drift time $D_{\text{x}} = \theta_1/(2\theta_2)$ ranges within the interval $[D_\beta, D_{\beta+\rho}]$.
- $f(D_{\text{x}};\theta) \geqslant 2\,\sigma_{\text{R}}$
- The polynomial is convace, i.e., $\theta_2 < 0$.

The first condition can be more restricted to achieve more reliable results, by shrinking the interval towards the center of the window. When no peak is found, the moving window is shifted one index forward. If a peak is detected, the window is shifted half the window length forward before the next scan begins, but first the peak parameters and the reduced spectrum are computed. Given the drift time D_{x} of maximal intensity in the window, the mode descriptor of the peak is simply $m = D_{\text{x}} \cdot f_{\text{ims}}$, where f_{ims} is the scaling constant that converts drift times (in ms) into IRMs (in Vs/cm^2). Given the mode, the other peak parameters can be inferred. Spangler *et al.* empirically derived that the width $w_{1/2}$ of the drift time interval between the point of maximal intensity and the point of half the maximum intensity (assuming peak symmetry) is $w_{1/2} = \sqrt{(11.09\,\mathcal{D}\,d)/V_{\text{d}}^2 + d_{\text{grid}}^2}$, where $\mathcal{D}$ is the diffusion coefficient,

d the mean drift time of the compound, V_d the drift velocity. Using the Einstein relation , $\mathcal{D}$ can be computed as $\mathcal{D} = k\mathcal{K}_B\mathcal{T}/q$, where k is the ion mobility, $\mathcal{K}_B$ the Boltzmann constant, $\mathcal{T}$ the absolute temperature and q the electric charge. From this, the standard deviation can be computed as follows $\sigma = \omega_{1/2}/2.3548 \cdot f_{\text{ims}}$, remark that in a Gaussian curve $\omega_{1/2} \approx 2.3548\sigma$. Empirically, the mean is found to be $\mu' = \left(d + \sqrt{(4.246 \cdot 10^{-5})^2 + d^2/585048.1633}\right) \cdot f_{\text{ims}}$. Having computed the peak descriptors, they are converted into the parameters (μ, λ, o) of the Inverse Gaussian parameterization. The scaling factor v for the peak is $v = f(d;\theta)/g(d \cdot f_{\text{ims}}; \mu, \lambda, o)$. The model function is subtracted from the spectrum, and the next iteration is started with a window shifted by $\rho/2$ index units. For each spectrum, the output of this step is a *reduced spectrum*, which is a set of parameters for a mixture of weighted Inverse Gaussian models describing the peaks.

Aligning two Consecutive Reduced Spectra

Given for each spectrum a set of peak parameters, the question arises how to merge the sets $P = (P_i)$ and $P^+ = (P_j^+)$ of two consecutive spectra. For each peak P_i, the Inverse Gaussian μ_i, λ_i, o_i, the peak descriptors μ_i', σ_i, m_i (mean, standard deviation, mode) and the scaling factor v_i are stored, and similarly so for the peaks P_j^+. The idea is to compute a global alignment similar to the Needleman-Wunsch method [2] between P and P^+. Now, an alignment score between P_i and P_j^+ and a score leaving a peak unaligned (i.e., a gap) has to be specified. Therefore the ratio between the probability that model P_j^+s position is described by P_i parameters i.e. $g(m_j^+;\, \mu_i, \lambda_i, o_i)$ and the probability that model P_is position plus offset δ is described by P_i parameters i.e. $g(m_i + \delta;\, \mu_i, \lambda_i, o_i)$ is considered. Two models are being considered as unaligned, when the ratio drops below 1. Here, offset is $\delta := d_{\text{grid}}/2.3548 \cdot f_{\text{ims}}$. Computing the alignment now in principle uses the standard dynamic programming approach. However, while using log odds for *deciding* whether to align P_i to a gap, the *score* is zero. Thus, the borders of $(S_{i,j})$ are initialized to zero. The computation of $S_{i,j}$ for $i \geqslant 1$ and $j \geqslant 1$ is described as follows

$$\zeta_{i,j} = \log\left(\frac{g(m_j^+;\, \mu_i, \lambda_i, o_i)}{g(m_i + \delta;\, \mu_i, \lambda_i, o_i)}\right), \qquad S_{i,j} = \max\begin{cases} S_{i-1,j-1} + \zeta_{i,j}, \\ S_{i-1,j}, \\ S_{i,j-1}. \end{cases}$$

The alignment is obtained with a traceback, recording the optimal case in each cell, as usual. There are three cases to consider.

- If P_j^+ is not aligned with a peak in P, potentially a new peak starts at this retention time. Thus model P_j^+ is put into a new peak chain.
- If P_j^+ is aligned with a peak P_i, the chain containing P_i is extended with P_j^+.
- All peaks P_i that are not aligned to any peak in P^+ indicate the end of a peak chain at the current retention time.

All completed peak chains are forwarded to the next step, two-dimensional peak model estimation.

Estimating 2-D Peak Models

Let $C = (P_1, \ldots, P_n)$ be a chain of one-dimensional Inverse Gaussian models. The goal of this step is to estimate a two-dimensional peak model (product of two one-dimensional Inverse Gaussians) from the chain, or to reject the chain if the chain does not fit such a model well. Potential problems are that a peak chain may contain noise 1-D peaks, or in fact consist of several consecutive 2-D peaks at the same drift time and successive retention times. First the peak height vector $h = (h_i)_{i=1,\ldots,n}$ is set up at the individual modes; $h_i := v_i \cdot g(m_i; \mu_i, \lambda_i, o_i)$ and in parallel the vector $r = (r_i)$ of corresponding retention times of the models. To identify noise chains, an affine function of r to h is fitted by finding (θ_0', θ_1') to minimize $\sum_i (h_i - \theta_1' r_i - \theta_0')^2$ (linear least squares). Then the cosine similarity between $\ell = (\ell_i)$ with $\ell_i := \theta_1' r_i + \theta_0'$ and h is being computed. If it exceeds 0.99, there is no detectable concave peak shape and the chain is discarded as noise. Otherwise, similarly to Section "Spectrum Reduction", a fitting of quadratic polynomials $h_i \approx \theta_2 r_i^2 + \theta_1 r_i + \theta_0$ is proceeded to appropriately sized windows of retention times. The index $i^* := \arg\max_i(h_i)$ of the highest signal is recorded and checked for following peak conditions: (1) As a minimum peak height $h_{i*} \geqslant 5\sigma_R/2$ is required. (2) As a minimum peak width ρ (size of the moving window), $\rho = (25 + 0.01 \cdot R_{i*})$ is used. The lower bound of 25 was explained above, but with increasing retention time R_i, the peaks become wider. This was found empirically by manually examining about 100 peaks in several measurements and noting a linear correlation of peak retention time and width. When the peak conditions are satisfied, the descriptors for an Inverse Gaussian are computed as follows: $v = -\theta_2(\theta_1/(2\theta_2))^2 + \theta_0$, $\sigma = \sqrt{v/(2|\theta_2|)}$, $m = -\theta_1/(2\theta_2)$, $\mu' = m + 0.1\sqrt{m}$. Having computed the descriptors, the model parameters (μ, λ, o) are computed. Ideally, a single shifted Inverse Gaussian model is provided from the chain. However, to optimize the model fit and to deal with the possibility of finding several peak windows in the chain, the EM algorithm is used to fit a mixture of Inverse Gaussians to (h_i) and finally only the dominant model from the mixture is taken as the resulting peak. To achieve descriptors in IRM dimension and the volume v,the weighted average over all models within the chain is computed for every descriptor.

References

[1] M. D'Addario, D. Kopczynski, J. I. Baumbach, and S. Rahmann. A modular computational framework for automated peak extraction from ion mobility spectra. *BMC Bioinformatics*, 15(1):25, 2014.

[2] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453, 1970.

[3] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

Subproject B2
Resource optimizing real time analysis of artifactious image sequences for the detection of nano objects

Peter Marwedel Heinrich Müller Alexander Zybin

Multi-Objective Design Space Exploration for GPGPU Programs

Pascal Libuschewski
Lehrstuhl 12
Technische Universität Dortmund
pascal.libuschewski@tu-dortmund.de

In this report a multi-objective design space exploration for Graphics Processing Units (GPUs) is presented. Different objectives and constraints can be considered, e.g. the run time and energy-consumption in conjunction with a real-time constraint and a lower bound for the result that is calculated by the given GPU program. Beside the local processing on GPUs an offloading approach was developed, where the work or part of the work can be offloaded to a server via the Long Term Evolution (LTE) wireless network.

1 Introduction

This work presents a multi-objective design space exploration for GPUs. For any given General Purpose computing on Graphics Processing Units (GPGPU) application, a Pareto front of best suited GPUs can be calculated. The objectives can be chosen according to the demands of the system, for example energy efficiency, run time and real-time capability. As mobile processing has other demands than processing on a desktop or high performance system, different processing options can be identified. When a mobile device needs to process large amount of data and an energy supply is not available, offloading is a good option to save energy. Therefore the presented work makes use of a simulator for the Long Term Evolution (LTE) wireless network. With this simulator it is possible to identify offloading strategies for different environments, which can then be used dynamically to save energy on mobile devices.

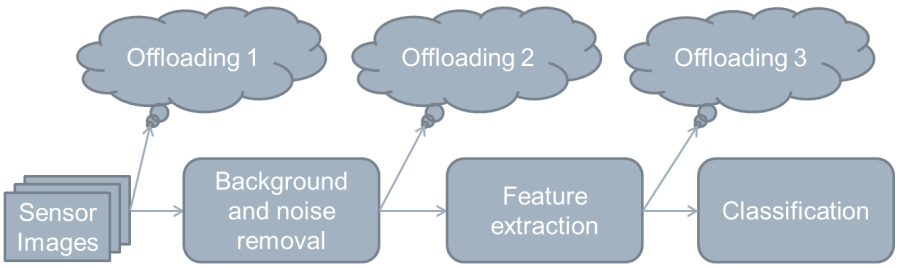


Figure 1: Offloading example. Different steps in the pipeline where the remaining calculation can be processed on a server. The work is offloaded via the Long Term Evolution (LTE) wireless network.

2 Multi-Objective Energy-Aware Design Space Exploration

The automatic design space exploration framework, published in [4], was extended to a multi-objective optimization which can also model constraints as real-time processing and lower bounds for the detection quality. An important factor for a GPU design space exploration is the evaluation time. Two approaches have been developed to reduce the evaluation time without sacrificing accuracy. With a checkpointing approach current states of the GPU processing can be saved and loaded within the simulator. This reduces the evaluation time significantly. The second approach is the so called quick start approach, where some arrays in the GPU memory can be left uninitialized if this does not influence the further processing. These extensions of the design space exploration framework are accepted for publication in [3].

The framework was also extended to make use of the GPUSimPow [5] simulator as an alternative to the GPUGPU-Sim [1] simulator in conjunction with the GPUWattch simulator. It was also extended to use a LTE simulator, which enables a simulated offloading to be part of the optimization. This approach was published in cooperation with the A4 project in [2]. The whole work or parts of the work can now be offloaded to a server within the simulation. This enables the design of mobile devices without a GPU or a device with a GPU where offloading is used in addition to save energy. Also tradeoffs can be identified, the processing on a local GPU is usally faster than offloading work to a server, but consumes more energy. Depending on the environment and if a power supply is given or not, the work can be offloaded and the best offloading point can automatically identified with the proposed design space exploration framework. In figure 1 an offloading example can be seen. In the processing pipeline different offloading points are defined. The work is processed locally on the GPU and the remaining work is offloaded and processed by a server, which afterward sends back the result of the processing back to the mobile device. Another cooperation was done with the A2 project to classify the signal and background time series of the virus detection pipeline, published in [6].

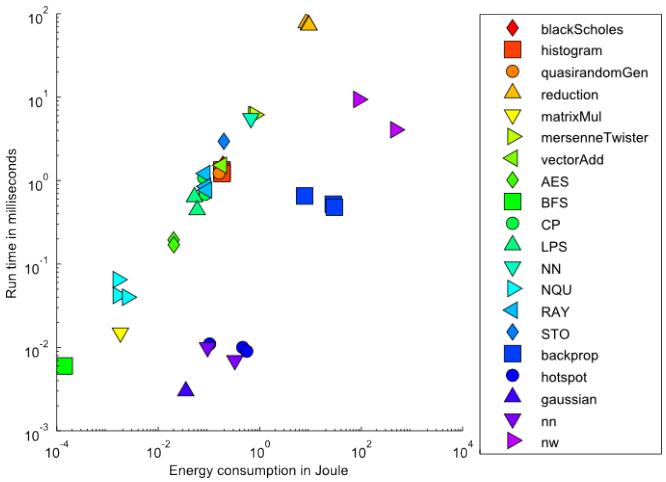


Figure 2: Results for the multi-objective optimization. Pareto fronts for different programs and different GPU architectures.

The evaluation of the framework was done on a wide variety of medical, physical, biological and industrial programs. The programs were running both OpenCL and CUDA code and the GPU architectures were modeled from energy efficient mobile GPUs up to high performance GPUs. Even GPUs that do not yet exist were modeled. As can be seen in figure 2, the tradeoffs between energy consumption and run-time were identified for the inspected programs.

In order to investigate the validity of the quick start and checkpointing approach measurements were done on a reduced number of frames and a prediction was done for a greater number of frames. This prediction was compared to the actual evaluated measurement. It could be shown that the error for both approaches is below 1.5% while the processing time was reduced by at least 80%. If a better accuracy is needed, the relative error could easily be reduced to values below 0.05% by processing more frames.

3 Conclusion and Future Work

A framework for an automatic, multi-objective energy-aware design space exploration of GPGPUs has been developed. This framework can be used in a wide variety of applications. As the framework was extended to simulate offloading via the LTE network and to make the processing of larger data possible, the framework can be used for small mobile devices up to large high performance systems. The objective is not limited to power- or energy efficiency. The detection rate of the virus detector, number of cycles or cycles per Watt can also be used. Another important application is the development of new GPUs, as the configuration is not limited to existing GPUs and can easily be adapted to new GPU hardware.

The presented approach is the basis for further research: The automatic design space exploration framework will be used to identify the tradeoffs between an energy efficient computation and the detection quality of the virus detector. To test the framework on actual hardware an Odroid XU-3 board will be used for image pre-processing or for the processing of the full pipeline. The Odroid board has the built-in ability to measure the power consumption of the CPU, the RAM and the GPU. With this measurements a hardware/software co-design will be carried out, where the tradeoffs between energy consumption and run-time will be identified. As in the second phase of the project, the distributed processing on mobile devices becomes more important, the framework can also be used in more complex offloading scenarios.

References

[1] A. Bakhoda, G.L. Yuan, W.W.L. Fung, H. Wong, and T.M. Aamodt. Analyzing cuda workloads using a detailed gpu simulator. In *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on*, pages 163–174, 2009.

[2] Pascal Libuschewski, Dennis Kaulbars, Dominic Siedhoff, Frank Weichert, Heinrich Müller, Christian Wietfeld, and Marwedel Peter. Multi-objective computation offloading for mobile biosensors via the lte network. In *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare: MobiHealth 2014*, 2014.

[3] Pascal Libuschewski, Peter Marwedel, Dominic Siedhoff, and Müller Heinrich. Multi-objective energy-aware gpgpu design space exploration for medical or industrial applications. In *Proceedings of the Workshop on Computational Intelligence Techniques for Industrial and Medical Applications: CITIMA 2014*. IEEE Computer Society, 2014.

[4] Pascal Libuschewski, Dominic Siedhoff, and Frank Weichert. Energy-aware design space exploration for gpgpus. *Computer Science - Research and Development (CSRD)*, 2013.

[5] Jan Lucas, Sohan Lal, Michael Andersch, Mauricio Alvarez-Mesa, and Ben Juurlink. How a single chip causes massive power bills GPUSimPow: A GPGPU power simulator. In *Performance Analysis of Systems and Software (ISPASS), IEEE International Symposium on*, pages 97–106. IEEE, 2013.

[6] Dominic Siedhoff, Hendrik Fichtenberger, Pascal Libuschewski, Frank Weichert, Christian Sohler, and Heinrich Müller. Signal/background classification of time series for biological virus detection. In *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014. Proceedings*. Springer, 2014.

Multi-Objective Aware Parallelization For Embedded MPSoC

Olaf Neugebauer
Computer Science 12
TU Dortmund University
olaf.neugebauer@tu-dortmund.de

This report covers two works focusing on efficient utilization of embedded multiprocessor systems. First, a communication optimization for embedded heterogeneous multiprocessor system-on-chip (MPSoC) is presented. Here, genetic algorithms were used to optimize the communication between concurrently executed tasks of parallel applications. We conducted several experiments with real world benchmarks to analysis the capabilities of our approach. The results are evaluated on different simulated embedded platforms. The second part presents the modified preprocessing pipeline of the PAMONO virus detection software ported to an embedded system enabling future work on virus detection software on real embedded hardware.

Platforms

This section gives a brief overview of the embedded platforms and the runtime system we are using in both parts of this report. Task creation and management is implemented

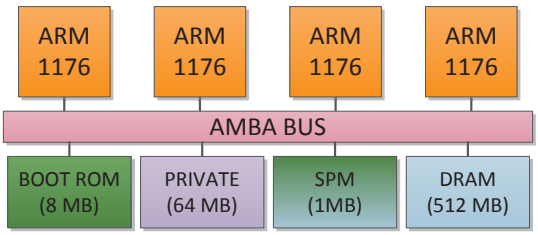


Figure 1: Target platform (Caches and memory controllers hidden)

using a lightweight runtime library utilizing R^2G [4] in combination with the RTEMS [6] real time operating system. The homogeneous and heterogeneous platform composed of ARM processors are created and simulated using Synopsys's Virtualizer [7]. Both platforms are composed of 4 processors and 4 memories connected by a bus as shown in Figure 1. In the case of the homogeneous system, all processors where clocked with 500 MHz and in the heterogeneous case two with 500 MHz, one with 250 MHz and one with 100 MHz. Energy consumption is measured utilizing a high-level model. For this purpose we implemented a Metrix component attached to our simulation environment which is triggered on every instruction, bus, cache and memory access and accumulates the corresponding energy values during measurement. For the processor we use energy values depending on the current processor's frequency. The energy consumption for the bus is a static value for each access. Memory values were obtained from CACTI [5].

Multi-Objective Aware Communication Optimization for Resource-Restricted Embedded Systems

Creating efficient parallel software for current embedded multicore systems is a complex and error-prone task. Numerous approaches try to parallelize and map sequential applications to a multicore platform but waste significant optimization potential. Parameters critical for the performance of software on such a platform, like communication performance between cores and the speed of different memories in the memory hierarchy, are often not considered in existing publications.

Using PICO (*Parallelism Implementer and Communication Optimizer*) in conjunction with the PAXES (*PArallelism eXtraction for Embedded Systems*) parallelizer [2, 3] we analyze the advantage of integrated multi-objective optimization. We use several benchmarks, and multiple platforms of homogeneous and heterogeneous nature to evaluate our approach. We observed that the solutions utilize different data exchange capabilities of the platform which improve performance in terms of runtime and energy consumption.

Our approach assumes a parallelized application (annotated C source code) and analyzes how various communication techniques influence the performance on different platforms. Transferred to a high-level representation, the challenge is to find a good mapping of task input/output and communication nodes to provide hardware/software implementations with respect to resource restrictions like energy consumption.

To analyze the benefits of using various communication mechanisms we implemented a set of different software FIFO prototypes. All implementations use a common API which makes it easy to add new communication methods into our approach. Those prototypes can be mapped onto different memories like DRAM or SPM. Several modifications were added to reduce the optimization overhead introduced due to simulation-based optimization.

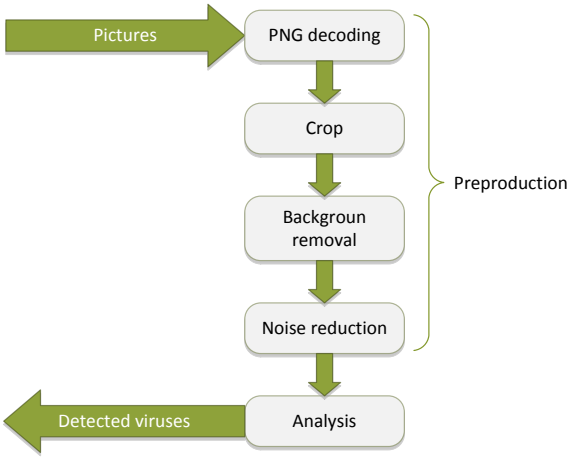


Figure 2: Pamono pipeline

Our approach is able to optimize runtime, energy and memory performance using a high-level cost model by using genetic algorithms. The results show that our algorithm is able to optimize parallelized applications and provides more detailed performance numbers compared to PAXES. We were able to reduce energy consumption by around 35% for JPEG compared to the sequential application. Furthermore, runtime reductions of roughly 55% were achieved for this benchmark on the homogeneous platform. The optimization algorithm achieves a reduction of energy consumption by about 15% compared to the naive communication implementation even for suboptimal parallelized applications like the presented Spectral benchmark.

PAMONO Virus Detection On Embedded Devices

To achieve a mobile usage of the PAMONO sensor and virus detection software the first step is to execute the detection software on an embedded system. Thus, in the first phase, we modified the preprocessing of the processing pipeline (cf.Figure 2). This light wight implementation of the preprocessing can be executed on our simulated homogeneous and heterogeneous ARM-based platforms including the real time operating system. Especially this real time operating system required a huge modification to the existing application. We conducted several experiments to analyze possible task partitionings of the sequential application. These parallel versions were evaluated on the above mentioned platforms. As a result, a parallelized preprocessing pipeline was created. In the next phase we plan to adopt the rest of the processing pipeline to embedded systems and to evaluate them on real hardware.

Conclusion and Future Work

We developed an integrated multi-objective aware parallelization infrastructure. By using genetic algorithms we are able to further optimize the parallelized solution. Here, we focused on the necessary communication between concurrently executed tasks. For homogeneous and heterogeneous target platforms, we used a simulator-based approach inspired by real world platforms combined with a real time operating system. By using this infrastructure we adopted the exiting preprocessing pipeline of the PAMONO virus detection software onto both platforms.

In the future, we will add a real embedded platform Odroid-XU3 [1]. This system is based on Samsung's Exynos5422 processor where four Cortex A15 processors are combined with four Cortex A7 processors. In addition, this platform allows direct monitoring of the power consumption of the different processors, memory and graphics card. In combination with our framework and the modified virus detection software this enables analysis on real hardware.

References

[1] Hardkernel Odroid-XU3. `http://www.hardkernel.com/main/products/prdt_info.php?g_code=G140448267127`, December 2014.

[2] Daniel Cordes, Michael Engel, Olaf Neugebauer, and Peter Marwedel. Automatic Extraction of pipeline parallelism for embedded heterogeneous multi-core platforms. In *Proc. of CASES*, 2013.

[3] Daniel Cordes, Olaf Neugebauer, Michael Engel, and Peter Marwedel. Automatic Extraction of Task-Level Parallelism for Heterogeneous MPSoCs. In *Proc. of ICPP*, 2013.

[4] Andreas Heinig. R2G: Supporting POSIX like semantics in a distributed RTEMS system. Technical Report 836, TU Dortmund, Faculty of Computer Science 12, Dortmund, December 2010.

[5] Naveen Muralimanohar, Rajeev Balasubramonian, and Norman P Jouppi. CACTI 6.0: a tool to model large caches. Technical report, HP Laboratories, 2009.

[6] RTEMS. RTEMS Operating System | Real-Time and Real Free. `http://www.rtems.com/`, 2014.

[7] Synopsys. Virtualizer, Virtual Prototyping Solution. `http://www.synopsys.com`, 2014.

Recent Advances in Analysis Methods for PAMONO Biosensor Data

Dominic Siedhoff
Lehrstuhl für Graphische Systeme
Technische Universität Dortmund
dominic.siedhoff@tu-dortmund.de

1 Introduction

This report summarizes results from project B2, concerning the analysis of PAMONO biosensor data. They comprise an extended model of the PAMONO biosensor (section 2) which is used within an advanced version of the Synthesize/Optimize method (section 3) as proposed in last year's report. A robust method for time-series classification is discussed (section 4). Finally, future work for the second phase is listed (section 5).

2 Extended Sensor Model

The sensor model described in last year's report was extended to model the shallow depth of field of the optical system [8], which is important because the imaged surface (gold layer in Figure 1, left) is not parallel to the focal plane (green) of the camera. A spatially varying point spread function (PSF) is used to account for the decreasing focus above and below the focused line (Figure 1, center). Hence, out-of-focus variations in the nano-object appearances need no longer be captured by the user in extracting nano-object templates as seeds for synthesis. This reduces manual effort significantly. Results can be found in Table 1, where the columns with synthetic templates refer to the extended sensor model.

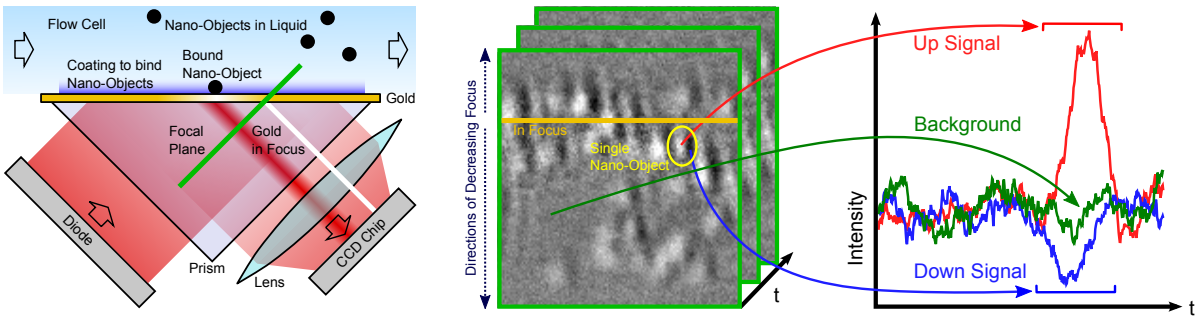


Figure 1: Scheme of PAMONO sensor with focal plane of the camera (left). PAMONO images with line of focus (center). Classes of time-series (right).

Table 1: Detection quality measures after parameter optimization [8].

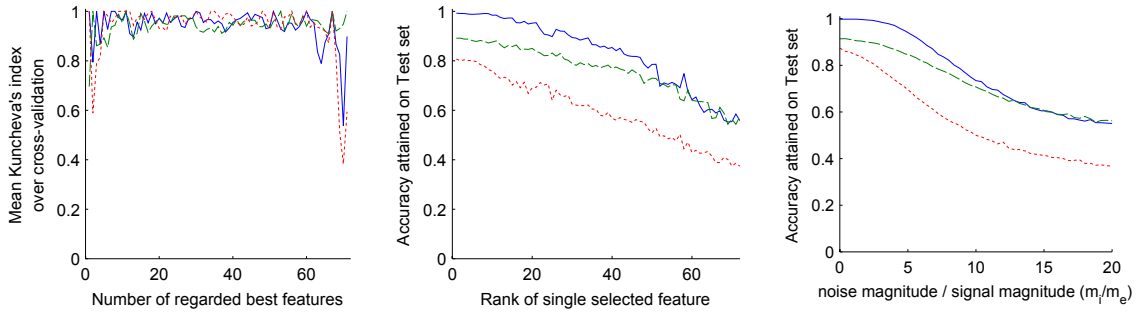
Dataset	Measure	Real Background		Synth. Background	
		Real Templ.	Synth. Templ.	Real Templ.	Synth. Templ.
Training	Precision	1.0000	1.0000	1.0000	0.9983
	Recall	0.9484	0.9533	0.9800	0.8983
Test	Precision	1.0000	1.0000	1.0000	0.9983
	Recall	0.9434	0.9266	0.9867	0.8550
Real	Precision	0.9667	0.9476	0.9527	0.9489
	Recall	0.9286	0.9539	0.9286	0.8341

3 Enhanced Synthesize/Optimize Method

Last report's Synthesize/Optimize method for automatic object detection and classification by means of data synthesis and parameter optimization was enhanced and fine-tuned towards finalization. A paper covering the entire method is pending, while partial results can be found in [8]: Here the detector was applied to real PAMONO data (last two lines of Table 1) after being trained on synthetic data of various degrees of realism. Besides that, scale space extensions of the intensity curvature features in [9] and the blob features in [3] were developed for improved classification. Validating the classification on real data with smaller nano-objects is currently in progress.

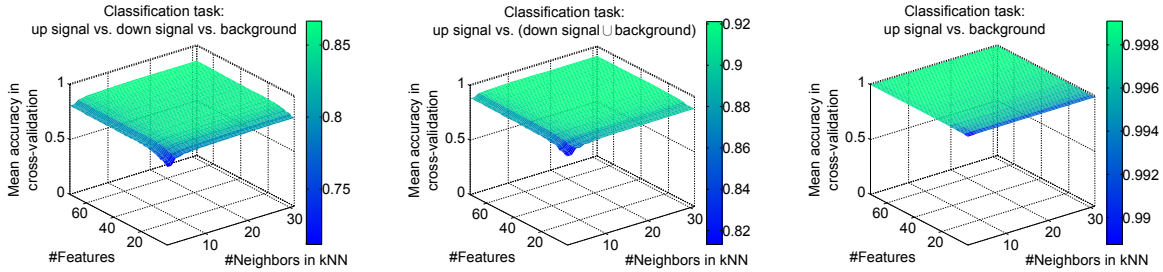
4 Robust Time-Series Classification

A key component in PAMONO object detection is time-series classification, with the task of separating irrelevant background signals and the up and down signals arising around nano-object adhesions (cf. Figure 1, right). A key impediment is the low signal-to-noise ratio in the data. Robust wavelet-based features of time-series were developed to account



(a) (b) (c)

Figure 2: Kuncheva indices (a), single feature accuracies (b) and robustness to noise (c) for the three classification tasks 'up vs. down vs. background' (*short dashes*), 'up vs. (down ∪ background)' (*long dashes*), 'up vs. background' (*solid line*).



(a) (b) (c)

Figure 3: Mean accuracy in 10-fold cross-validation over the number of best features and the number of neighbors to be used in k-NN for the three classification tasks.

for both aspects [7]. These features are translation-invariant (TI), i.e. series with different times of appearance of a nano-object have similar representations in feature space. TI wavelets were developed for denoising [2] but have to the best of our knowledge not been previously used to derive TI features. In collaboration with project A2 a fast condensed k-NN classifier [1] was developed to classify time-series via TI features. Condensation of the training set was carried out by A2's BICO coreset clustering algorithm [4], which can handle the big datasets produced by PAMONO. A feature selection strategy was developed, which proved to be stable in terms of Kuncheva's index [6], cf. Figure 2(a). Feature ranking was shown to be meaningful in terms of balanced classification accuracy by classifying the data via single features, cf. Figure 2(b). Robustness of classification to increasing noise is demonstrated in Figure 2(c): Up to five times larger noise than signal magnitude is tolerated well, and the classifier converges to the limit of random guessing for the respective classification task for noise that exceeds signal magnitude by factor 20. Figure 3 shows that the influence of the two main parameters (number of features and k in k-NN) is small, as long as both are chosen 'large enough'. Figure 3(c) demonstrates that accuracy 0.999 is achieved in separating up signals from background.

5 Future Work

The next step is to publish a final paper, giving an overview of the entire system and to write the dissertation. Further work in B2 is as follows: A generalization to detecting multiple types of objects simultaneously is planned. Object tracking is important because nano-objects like certain influenza viruses may roll over the sensor surface instead binding permanently. Preliminary studies on this have been conducted in [5]. The cooperation of multiple sensors in a network, distributed analysis and error detection/compensation are to be researched. For this scenario, an algorithm paradigm that is sensitive to the execution context (e.g. embedded or server, possibilities for offloading, battery status) is of interest.

References

[1] Fabrizio Angiulli. Fast condensed nearest neighbor rule. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 25–32, 2005.

[2] R. R. Coifman and D. L. Donoho. *Translation-invariant de-noising*. Springer, 1995.

[3] W. K. Moon et al. Computer-aided tumor detection based on multi-scale blob detection algorithm in automated breast ultrasound images. *IEEE Transactions on Medical Imaging*, 32(7):1191–1200, 2013.

[4] Hendrik Fichtenberger, Marc Gillé, Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. BICO: BIRCH meets coresets for k-means clustering. In *Proceedings of the 21st European Symposium on Algorithms (ESA)*, 2013.

[5] T. Heming. Automatisches Tracking von nicht ortsständigen Partikeln in hochverrauschten Daten. BSc. Thesis, TU Dortmund, 2014.

[6] L. I. Kuncheva. A stability index for feature selection. In *Artificial Intelligence and Applications*, 2007.

[7] D. Siedhoff, H. Fichtenberger, P. Libuschewski, F. Weichert, C. Sohler, and H. Müller. *Pattern Recognition (36th German Conference, GCPR 2014)*, chapter Signal/Background Classification of Time Series for Biological Virus Detection, pages 388–398. Springer International Publishing, 2014.

[8] D. Siedhoff, P. Libuschewski, F. Weichert, A. Zybin, P. Marwedel, and H. Müller. *Bildverarbeitung für die Medizin 2014*, chapter Modellierung und Optimierung eines Biosensors zur Detektion viraler Strukturen, pages 108–113. Springer, 2014.

[9] D. Thomann, D. R. Rines, P. K. Sorger, and G. Danuser. Automatic fluorescent tag detection in 3d with super-resolution: application to the analysis of chromosome movement. *Journal of Microscopy*, 208:49–64, 2002.

Subproject B3
Data Mining on Sensor Data of Automated Processes

Jochen Deuse Katharina Morik

Data Stream Appliances for existing Industrial Processes

Hendrik Blom
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
hendrik.blom@tu-dortmund.de

Introduction
An important step towards the vision of the industry 4.0 is the integration of "smart" data based methods into the monitoring and control of processes of already existing and running industrial plants or machines [**?**, 1]. The monitoring and control are based on the analysis of real-time data streams and the learning of models to predict the quality of the process outputs.
The most important insight is, that every industrial machine or plant is different and has to be treated individually. Different security regulations, IT-Systems, operators, sensor placements or just the wear of the machine requires a distinct treatment of every machine or plant. A minimal amount of manual effort is desired. Therefore, the deployment and the operation of the data based methods need to be as much automated and the software and computing hardware infrastructure should be as loosely coupled with the existing infrastructure as possible.

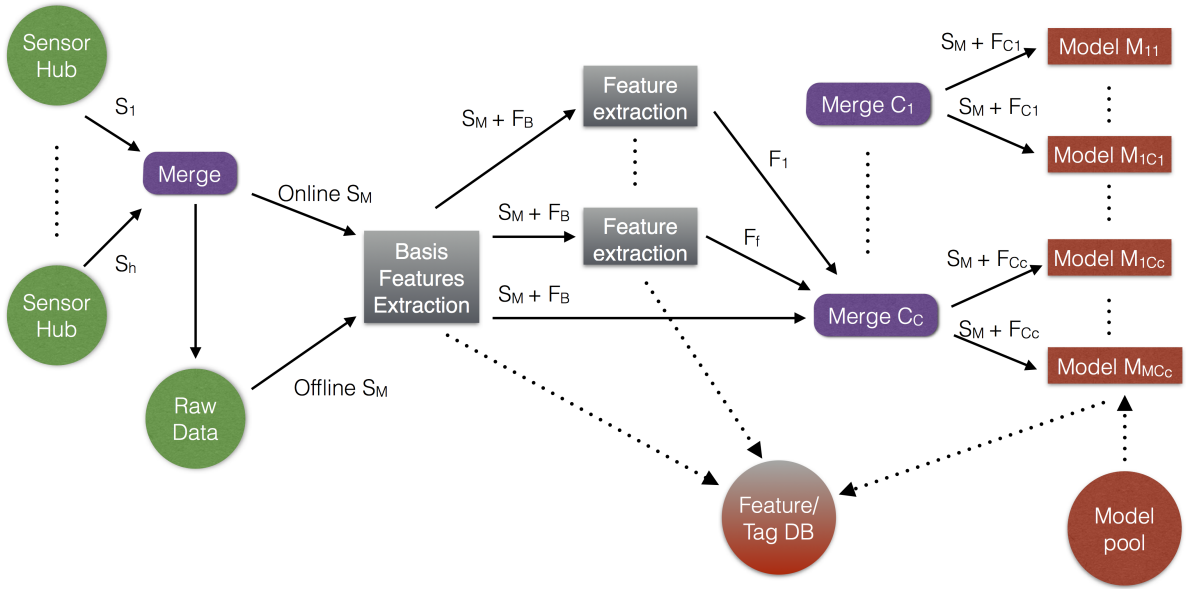
In the following, important definitions and first steps towards a fully autonomous data stream appliance for industrial processes are described. The description of the control and optimization of processes is missing, but it is the focus of my current research.

Plants, processes, models, data and features
Every industrial plant or group of machines A executes a sequence of consecutive processes or batches $p_i \in P_A$. Every process has a set of inputs, e.g. materials, and outputs. The quality of the output of a process should be measurable. The data based models are used to predict the quality of the output of the process. Every processes $p_i \in P_A$ has a defined start $p_i^{(s)}$ and an end $p_i^{(e)}$. The definition and identification of important events can divide the process into more meaningful subprocesses. The data of plant A is given

as a stream of sensor values $S_A = x_1, \ldots, x_t, \ldots$, where t is the actual time. The data of every process $p_i \in P_A$ is a substream $S_A[p_i^{(s)}, p_i^{(e)}]$ of the stream S_A. A feature $f \in F_{p_i}$ is an computed value of the stream or a substream of the process p_i, e.g. the minimum, maximum or the average.

Definition 1 (Model and Model class) *A model $m \in \mathcal{C}_{opt,m}$ is defined as a function $m(x) = y$, with $x \in \mathcal{X}_{opt} \cup \mathcal{X}_m$ and $y \in \mathbb{R}$. A model class $\mathcal{C}_{opt,m}$ is defined by 2 disjunct subsets of the possible features $\mathcal{X}_{opt}, \mathcal{X}_m \subset \mathcal{X}$ and a label Attribute $Y \in \mathcal{Y}$. $\mathcal{X}_{opt}$ defines the set of optional and $\mathcal{X}_m$ the set of mandatory features for the model. If an optional feature does not exists for the process, it will be replaced with 0.*

Definition 2 (Model object and model instance) *A model object $\mathcal{O}_C(R_m, R_p)$ represents 2 set of rules on the features and the label of model class $\mathcal{C}$. The first set selects processes based on metadata of the process and the second set selects processes based on feature values. These two rule sets control the similarity of the used data.*
A model instance is the implementation of a model. It will be learned by a learning algorithm and stored in the model database.



Model pool and model lifecycle
The robustness of the monitoring and control of the process can only be guaranteed, if there is always at least one model instance that can operate on the set of available sensors. Therefore, a pool or set of model instances for every quality label should be maintained. The pool of model instances managed automatically and all the models are evaluated continuously. If a model instance is performing rather poorly over a longer period of time, it will be removed from the model pool. Based on the actual prediction

quality of the model instances and the availability of sensor values a "pivot" model instances should be selected. Only the result of this model instance will be used to control and monitor the quality of the process.
Based on the prediction quality, sensor failures or simply the age new model instances will be created. Another approach is the use of Concept Drift Detection methods [2] to trigger new learning tasks. A learning task is defined by a selected model object. The learning data should be selected either on time (the last n processes) or on the similarity to the last n processes. If the mandatory features have not been extracted yet, they are extracted automatically from the raw data. If multiple learning algorithms are available, every algorithm will be used to learn a model. These models will be added to the model pool.

The number of models and the dynamics of the model pool need further investigation.

Raw data storage
In contrast to the usual approach on data stream analysis, all the data should be stored temporary. The harsh conditions in industrial processes and long maintenance cycles require dynamic adoptions to the set of operational sensors. If sensors fail, their values could be replaced or imputed by other sensor values. Based on the usage of the data for model instance creation and age, raw data should be removed automatically. The maximal volume of raw data depends heavily on the process itself. If only one product is produced and the maximal maintenance cycle is very short, only few processes need to be recorded completely. With more products and longer maintenance cycles more data should be recorded.
To give an example, the volume of the complete production data of one year with a maximal sensor frequency of 1 kHz of 2 Blow Oxygen Furnaces (BOF) [3] is only 350 GB. The raw data volume of one process is around 4 GB. After sparsification and compression the volume can be reduced to 40 MB.

The number of processes and the needed volume of the temporarily stored raw data need further investigation.

Online and offline execution
There are 2 modes of model application. The "online" application is based on the real execution of the process. Therefore, the maximum arrival speed of data is limited by the highest frequency of a sensor. In the "offline" mode, the maximum arrival speed is bounded by the i/o-speed of the hardware or the execution time of the slowest processing route. Except for the data sources, all the following processing steps are identical in the online and the offline processing of the production data. First all basis features and tags are extracted, like the start or end of a process or the static data of the process. After that, arbitrary features can be extracted. All the features and tags are stored in the feature/tag database. Then, all the needed features for the defined model classes $c_i \in C$

are merged and forwarded to the Model instances. The results of the models are also stored in the feature/tag database.

The speed of the execution is key to the adaption and the dynamic of the model pool. If more model instances can be learned and evaluated, the model pool can adapt better to the dynamics of the given process.

References

[1] Thomas Bauernhansl, Michael ten Hompel, and Birgit Vogel-Heuser. *Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung· Technologien· Migration*. Springer-Verlag, 2014.

[2] Henning Kagermann, Wolfgang Wahlster, and Johannes Helbig. Umsetzungsempfehlungen für das zukunftsprojekt industrie 4.0–abschlussbericht des arbeitskreises industrie 4.0. *Forschungsunion im Stifterverband für die Deutsche Wissenschaft. Berlin*, 2012.

[3] Stefan Rötner. Behandlung von concept drift in zyklischen prozessen. Master's thesis, TU Dortmund, 2014.

[4] Jochen Schlüter, Hans-Jürgen Odenthal, Norbert Uebber, Hendrik Blom, Tobias Beckers, and Katharina Morik. Reliable bof endpoint prediction by novel data-driven modeling. In *AISTech Conference Proceedings*. AISTech, 2014.

Inline Quality Prediction for online Process Control

Benedikt Konrad
Institut für Produktionssysteme
Professur für Arbeits- und Produktionssysteme
Technische Universität Dortmund
benedikt.konrad@ips.tu-dortmund.de

This Tech Report is to give a brief summary of some of the results developed in the first period of the collaborative research center 876. Based on the results described the improvement potential of the approach presented is discussed.

1 the IQP System developed

In the first period of the collaborative research center 876 it was the aim of project B3 to design and to develop a methodology to assess intermediate product quality during production on the basis of process parameters recorded. To achieve this aim, data gathering was introduced to the production processes. Figure 1 gives a summary of parameters collected at different processes. For quality prediction those parameters are automatically preprocessed and features derived by means of statistical methods [3]. Product quality is treated on an aggregated level. Instead of labeling each quality derivation revealed by ultra sonic tests at the process chain's end and trying to predict those errors for each steel rod, a quality level ranging 0 to 1 is predicted for the entire steel bar. This approach is necessary as process parameters are only available for steel bars - mapping those parameters on rods which are cut off from a bar is not possible. Consequently, quality levels represent a weighted aggregation of critical or relevant quality derivation for a single steal bar.

The Inline Quality Prediction (IQP) System consolidates all relevant modules to complete data preprocessing and to generate a quality prediction for each product [2]. An overview

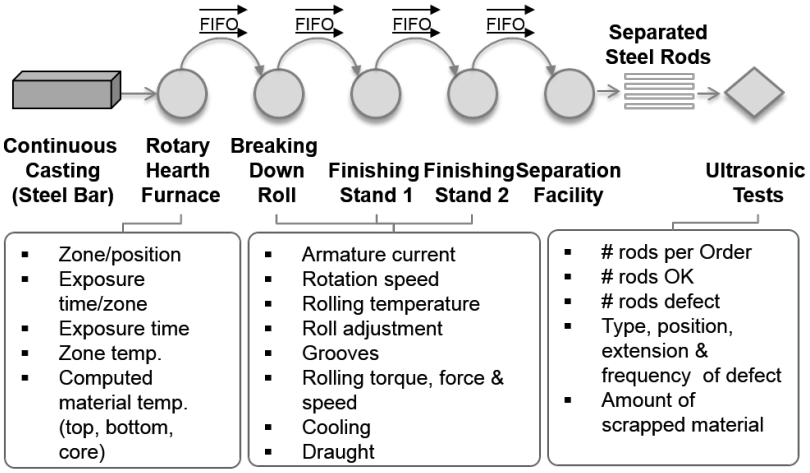


Figure 1: Process Parameters for Quality Prediction

on data preprocessing steps developed and data mining models deployed is given in [3]. On the basis of this estimation together with the progress made in the production process the IQP allows for deriving a decision support on whether the product in question should be processed any further, or not. Decisions are drawn according to a threshold approach [2].

Although the solution methodology developed during the first period solves the problem statement, it has certain limitations: first of all concentrating on aggregated quality levels leaves aside certain information on dimensions of errors. Features computed from process parameters are not specific enough to be linked to certain quality deviations easily. Process Control decisions only differentiating between ejecting the product from the process chain and continuing the production process does not pay attention to possible improvements which can be made by changing the production process itself.

Due to these improvement potentials the concept for the second period of the collaborative research center has been developed. The remainder of the tech report gives an overview on the planned steps for enhancing the IQP system.

2 Enhancement of the IQP System

As lined out in the introductory part, the problem statement for the second period is to extend the IQP concept, such that multiple process control decisions can be realized. Providing decision support for judging on a possible ejection of an intermediate product helps to reduce waste of energy etc. in the production processes but still leads to material waste, as the intermediate products usually have to be scrapped. This insight led to the idea to integrate more sophisticated process control approaches into the IQP concept.

In case of steel production processes two possible control actions were identified. Firstly, process parameter adaptions while continuing to process the product and, secondly, re-assigning intermediate products to customer orders, such that quality expectations are met. [1]

Process parameter adaptation aims at generating feasible parameters based on the quality assessment for a certain intermediate product as well as the next production steps. In order to check the adaption's feasibility, the IQP System has to be expanded beyond methods of data mining. As long as the production process for an intermediate product has not been recorded in the past the influence of the altered parameters on the product given the previous production processes has to be evaluated. To do so, material forming simulation based on finite element methods (FEM) is incorporated. Obviously, FEM simulations require computation times that are longer than the timespan between two consecutive production steps. Consequently, integrating these simulations into the online concept of IQP is not an option as their result has to be known when the IQP system has to draw a dicision on the control action to be chosen. Therefore, simulations have to be completed beforehand, leaving the research question of how to determine process parameter adaptions that have to be known for decision making in the (near) future.

The second extension on process control actions stated above, re-assigning intermediate products to customer orders is focused. The concepts described previously aim at maximizing product quality. Yet, in some cases minor deviations might be tolerable. This is especially the case, when such a product can be given to another customer whose quality requirements are still fulfilled despite of the deviations predicted. In order to allow any re-assignment, several constraints must not be violated. First of all, at least one more customer order has to be identified that fits the basic characteristics (material, shape, dimensions etc.) of the intermediate product assessed. Secondly, process setup has to be checked as fixed setup cycles might impede short term order swaps. Moreover, it may be assumed that - in general - order due dates have to be met. I. e., swapping customer orders in the production process is only feasible if - after re-sequencing all orders in the process - all customers receive their products on time. Besides these three basic constraints, many more process specific ones have to be considered. One possible solution to this approach is given by mathematical programming and applying methods of operations research for solving the model.

Although the methodologies described above can be implemented in stand-alone solutions, an integration into the IQP system is strived for. To accomplish the decision support system implemented in the IQP has to be enhanced as it has to be able to differenciate between four possible control actions to be chosen (see fig. 2).

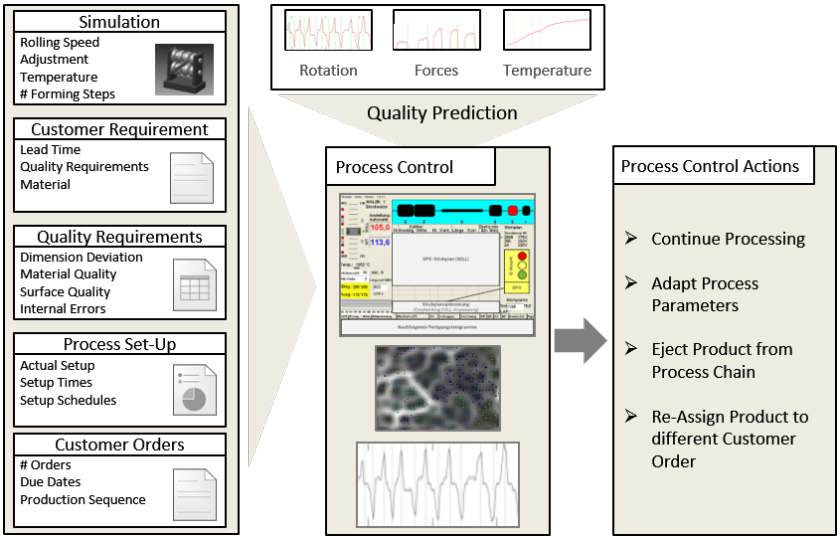


Figure 2: Concept of Process Control Action for second Period [1]

3 Conclusion

The enhanced concept of the IQP system described offers the opportunity to reduce waste of material as well as energy. While it increases the flexibility to react on quality deviations appropriately it demands significant computational effort to derive the control action suited best in each situation. Consequently, developing a methodology that efficiently chooses the control action is a major issue in the second period of research.

References

[1] Jochen Deuse, Benedikt Konrad, and Fabian Bohnen. Reduzierung von variabilität - einsatz von data mining in produktionsystemen. In Axel Freidewald and Hermann Lödding, editors, *Produzieren in Deutschland - Wettbewerbsfähigkeit im 21. Jahrhundert*, pages 299–316. Gito, Berlin, 2013.

[2] Benedikt Konrad, Daniel Lieber, and Jochen Deuse. Striving for zero defect production: Intelligent manufacturing control through data mining. In Katja Windt, editor, *Proceedings of the CIRP Sponsored Conference RoMaC 2012 Robust Manufacturing Control*, pages 215–229, Berlin and Heidelberg, 2012. Springer.

[3] Daniel Lieber, Marco Stolpe, Benedikt Konrad, Jochen Deuse, and Katharina Morik. Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. In *Procedia CIRP - 46th CIRP Conference on Manufacturing Systems*, volume 7, pages 193–198. Elsevier, 2013.

Current Work and Improvements on the Analysis of Distributed Sensor Data

Marco Stolpe
Lehrstuhl für Künstliche Intelligenz
Fakultät für Informatik
Technische Universität Dortmund
marco.stolpe@tu-dortmund.de

In the context of a case study from the steel industry, millions of sensor measurements along the process chain of a hot rolling mill have been gathered and analysed. New techniques have been developed for the segmentation and aggregation of value series data, for the reconstruction of labels from aggregated label information and for distributed anomaly detection in the vertically partitioned scenario. This report presents our current work and improvements on each step of the data analysis process.

1 Introduction

Project B3 of the Collaborative Research Center SFB 876 deals with quality prediction in automated production processes. Based on measurements from distributed sensors that describe how a product is processed, its final quality should be predicted as early as possible and in real-time during the running process. Data analysis can be divided into several substeps, like *business understanding*, *data acquisition*, *preprocessing*, *modeling* as well as the final *deployment* of trained prediction models. Driven by a real-world case study from the steel industry, previous work [2, 3, 7, 8] has already focused on each of these steps. Complex data analysis processes are usually iterative by nature, where knowledge gained in later analysis steps may lead to new ideas and insights for earlier ones. The following sections therefore present recent ideas and improvements that build on experiences gained in our earlier work.

2 Data Acquisition and Management

Although the project's main focus is on the development of decentralized data mining techniques, a central storage of historical data allows for the repeatability of experiments and easier comparisons between already existing and newly developed analysis methods. In the context of the case study's hot rolling mill process, for each steel block two different types of data were stored in a single relational database: *Static meta information*, like the steel block's unique ID, its material, its dimensions or the date and time on which it was processed and *value series data*, consisting of sensor measurements of process parameters over time. All sensor measurements of a single processing step were stored in a single database table, reasoning that the usage of indices and standard SQL could potentially ease the filtering and aggregation of such values in later analysis steps. However, with millions of measurements stored in a single table, filter and aggregation operations began to take several minutes, despite of indices being defined on the relevant fields. Moreover, due to the use of indices, the database grew more and more out of proportions with each sensor measurement inserted. Importing new data became slower and slower, due to the automatic reorganization of indices. As it turned out, filters were mostly applied only on the stored meta data, while aggregation operations could be applied independently on single value series. Therefore, the whole storage scheme has been reorganized. While static meta information is now still stored in a relational database, sensor measurements are getting stored in a single compressed CSV file per steel block. Thereby, the previous storage space of 160 GB (including indices) could be reduced to a total of 7.5 GB, without having to compromise on the needed functionality. Based on the available meta information, it still can be decided which sensor measurements should be exported to RapidMiner for further processing, which then aggregates the information per value series.

3 Data Preprocessing

In previous work [3], the value series of each steel block were first cleansed, segmented and then the values of each segment were aggregated by statistical measures like the mean, standard deviation, minimum and maximum. The resulting values were then further aggregated to a fixed number of statistical measures, thereby becoming independent from the different lengths of the given value series. It has been shown empirically that the aggregation allows for the identification of different operational modes corresponding to the steel blocks' final dimensions, but might not preserve the quality-related differences between value series. The identified patterns further seem to numerically distort and overlay the quality-relevant information. Current work therefore focuses on a different symbolic representation of value series and their proper normalization, based on the previously identified global patterns.

Normalization consists of the calculation of centroids for all value series belonging to a particular operational mode, using a fast averaging technique with the dynamic time warping (DTW) distance [5]. Then, the centroid is subtracted from each corresponding value series, based on the previously calculated warping path. All values are then divided by the centroid's standard deviation. The resulting values now only encode differences to the identified global patterns, making the value series of different operational modes comparable.

In the field of text processing, different numbers of characters and words per document have been successfully handled by the *vector space model* (VSM) which encodes frequencies of words in sparse vectors. A frequency encoding of value series (see e.g. [6]), however, first requires their discretization into symbols and words (i.e. sequences of symbols). For example, a well-known symbolization technique, SAX, divides the value series into fixed length intervals and assigns symbols to each interval according to their mean. However, the performance of SAX decreases if the series' values don't follow a normal distribution, which is the case in the given scenario. Moreover, aggregation by the mean again might not preserve the quality-relevant differences between value series. The performance of SAX currently is therefore compared to a technique which assigns symbols based on previously computed centroids of clusters over the patterns in all intervals. First results suggest that the automatic determination of symbols with the k-Means clustering algorithm leads to a higher classification performance.

Another preprocessing step may consist of reconstructing the labels of individual observations from aggregated label information, e.g. given in the form of label proportions per customer order. Here, a recently published method [4] achieves a significantly higher classification performance than several state-of-the-art methods. However, all experiments were conducted using linear classifiers, while our previously developed LLP algorithm [8] focuses on non-linear classification. It is therefore planned to compare LLP to the new method on several linearly separable and non-linearly separable datasets.

4 Modeling and Prediction

While the aforementioned feature extraction is independent from other processing steps, a steel block's final quality might depend on the processing at different stations. Previous work [7] has focused on the detection of anomalies in this vertically distributed scenario. The related algorithm is based on the 1-class version of the Core Vector Machine (CVM) [9]. This version, however, can only work with normalized kernels, excluding the linear kernel, which makes comparisons with existing distributed linear SVMs difficult. In contrast, the later developed Generalized Core Vector Machine (GCVM) [10] has the desired properties. Current work therefore focuses on the design and implementation of a distributed algorithm for the GCVM. Once implemented, the algorithm's communication

costs and prediction performance can be compared to a vertically distributed SVM [1] which is based on the Alternating Direction Method of Multipliers (ADMM).

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011.

[2] D. Lieber, B. Konrad, J. Deuse, M. Stolpe, and K. Morik. Sustainable interlinked manufacturing processes through real-time quality prediction. In *Leveraging Technology for a Sustainable World*, pages 393–398, Berlin, Heidelberg, 2012. Springer.

[3] D. Lieber, M. Stolpe, B. Konrad, J. Deuse, and K. Morik. Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. In *Procedia CIRP - 46th CIRP Conf. on Manufacturing Systems*, volume 7, pages 193–198. Elsevier, 2013.

[4] G. Patrini, R. Nock, T. Caetano, and P. Rivera. (Almost) no label no cry. In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 190–198. Curran Associates, Inc., 2014.

[5] F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recogn.*, 44(3):678–693, March 2011.

[6] P. Senin and S. Malinchik. SAX-VSM: Interpretable time series classification using SAX and vector space model. In *13th Int. IEEE Conf. on Data Mining (ICDM)*, pages 1175–1180, Dec 2013.

[7] M. Stolpe, K. Bhaduri, K. Das, and K. Morik. Anomaly detection in vertically partitioned data by distributed core vector machines. In *European Conf. on Machine Learning and Knowledge Discovery in Databases - ECML/PKDD 2013*. Springer, 2013.

[8] M. Stolpe and K. Morik. Learning from label proportions by optimizing cluster model selection. In *European Conf. on Machine Learning and Knowledge Discovery in Databases - ECML/PKDD 2011*, volume 6913 of *LNCS*, pages 349–364, Berlin, Heidelberg, 2011. Springer.

[9] I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core vector machines: Fast SVM training on very large data sets. *J. Mach. Learn. Res.*, 6:363–392, December 2005.

[10] I. W.H. Tsang, J. T. Kwok, and J.A. Zurada. Generalized core vector machines. *Trans. Neur. Netw.*, 17(5):1126–1140, September 2006.

Detailed Prediction of Product Quality for Advanced Process Control

Mario Wiegand
Institut für Produktionssysteme
Professur für Arbeits- und Produktionssysteme
Technische Universität Dortmund
mario.wiegand@ips.tu-dortmund.de

This report outlines some of the central elements of the Inline Quality Prediction System developed in the first period of the collaborative research center 876. Particularly, the process control approach with special emphasis on required data for decision support is presented. Furthermore, a brief overview regarding the upcoming enhancement of this approach is given.

1 Introduction

Within the Collaborative Research Center 876 project B3 focuses on the development of algorithms for the time-constrained analysis of sensor data by means of data mining. The analysis of recorded data is conducted using the example of hot rolling processes in the steel industry. The first period aims at predicting the final product quality during the running production process by monitoring and analyzing real-time process data. This method allows for identifying quality deviations at an early stage providing a decision support on whether the processing of the current product should be canceled or continued. The second period targets the extension of this approach to an advanced process control system. In addition to the previous binary control decision sophisticated options of intervening in the running production process shall be developed. While the former approach only needs binary quality data, the enhanced control alternatives require differentiated information about products' quality.

2 Conducted Work

The first period of the Collaborative Research Center 876 was focused on the development of the Inline Quality Prediction (IQP) System [1]. The IQP contains all components necessary for enabling a quality prediction based on recorded process data by using data mining techniques. For that reason a wide variety of process data, including rolling speed, force and temperature, is measured for each steel bar by installed sensors at every process step. Subsequently, the recorded data is automatically preprocessed and useful features are extracted from the entire data set [2]. The relevant features are the input for the developed data mining model assigning each steel bar the quality label 0 respectively 1. Accordingly, the quality prediction is a classification task. The label 0 indicates an insufficient, the label 1 a high quality of the product. In the case of a predicted low product quality the operator of the rolling process has the possibility to eject the product from the process chain. Figure 1 gives an overview on the central steps of the IQP System.

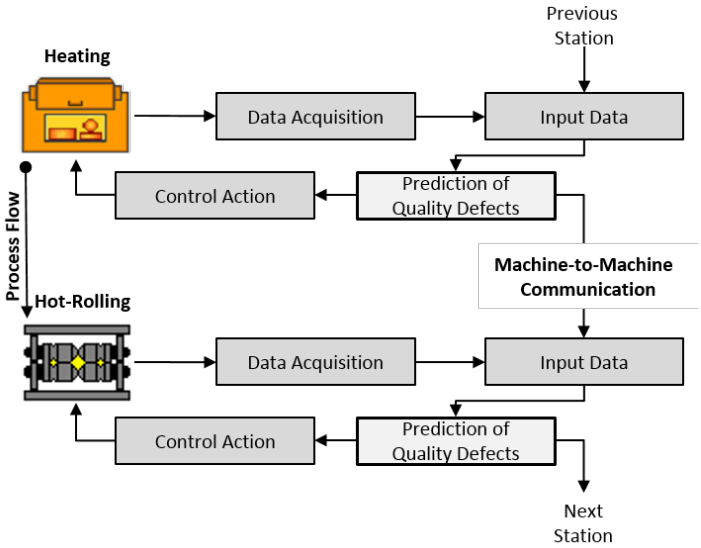


Figure 1: Quality prediction and process control

The model was trained by historical data creating the link between process parameters and quality labels. For the assignment of binary quality labels a quality level was defined. The quality level is an aggregated value between 0 and 1 combining different quality properties obtained by an ultrasonic test at the end of the process chain [1]. As classification methods can only be applied to the prediction of nominal labels, all metric labels are transformed to a binary scale of measurement using a domain-specific threshold. Steel bars with a quality level $\leq 0,75$ are assigned the label 0, whereas bars with a quality level $> 0,75$ obtain the label 1.

Process parameters are measured and analyzed during the whole production process. Thus IQP delivers a real-time decision support empowering the operator to decide if the process of the current product should be canceled or continued.

3 Prediction of detailed quality information for process control

The developed solution approach provides the opportunity to prevent a further processing of inferior products. Therefore production costs and time can be reduced. However, a great amount of the ejected products cannot be reused and has to be scrapped. This leads to a waste of material, whereby more sophisticated control approaches are required.

As a result the second period targets the extension of the IQP system to an advanced process control system. This advanced approach expands the former binary control decision by two additional control alternatives. One of these alternatives consists in adapting process parameters during the running production process. If the predicted quality of the currently produced steel bar does not meet the fixed requirements, process parameters are changed in order that these requirements are fulfilled. For this purpose the influence and the feasible range of possible adaptions have to be considered. The other alternative allows for the reassignment of contemporary processed products to another customer order. This option can be reasonable if customers have different quality requirements. Thus a processed product with a predicted low quality can be allocated to another customer having minor quality requirements compared to the primal one. To this end different restrictions have to be regarded like necessary setup activities and delivery dates related to the customer orders.

The previous binary control decision only demands an aggregated quality indicator for the assignment of a label to every steel bar. The more sophisticated control approaches need more detailed information including position and kind of quality errors. This leads to the assignment of multiple labels to one steel bar. Between those labels dependences could exist. For example errors in early steps of the production process could have an influence on later process steps leading to additional errors in the product. The resulting kind of prediction tasks is called structured output learning. Differentiated information about error types for learning the model can be obtained by detailed results from ultrasonic tests or by existing material forming simulation. Defective products and process operations are rare in proportion to the entire production process. This lead to an unbalanced relation between positive and negative examples. To generate negative examples artificially material forming simulation based on finite element methods can be used.

Possible solutions to the described learning problem with structured output are given by approaches like classifier chains [3] or support vector machines for structured output [4].

4 Conclusion

The presented approach of an Inline Quality Prediction System developed in the first period of the collaborative research center 876 allows for predicting final product quality during the running production process. Thus the operator has the possibility to eject inferior products from the process chain. With the upcoming enhancement of this approach additional control options will be available leading to further improvements of the process. For this purpose more detailed quality information about the product as well as new prediction methods will be necessary.

References

[1] Benedikt Konrad, Daniel Lieber, and Jochen Deuse. Striving for zero defect production: Intelligent manufacturing control through data mining. In Katja Windt, editor, *Proceedings of the CIRP Sponsored Conference RoMaC 2012 Robust Manufacturing Control*, pages 215–229, Berlin, 2012. Springer.

[2] Daniel Lieber, Marco Stolpe, Benedikt Konrad, Jochen Deuse, and Katharina Morik. Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. In *Procedia CIRP - 46th CIRP Conference on Manufacturing Systems*, volume 7, pages 193–198, 2013. Elsevier.

[3] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3): 333–359, 2011.

[4] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6: 1453–1484, 2005.

Subproject B4
Analysis and Communication for dynamic traffic prognosis

Michael Schreckenberg Christian Wietfeld

Microscopic Modelling of Multi-Lane Highway Traffic

Lars Habel
Physik von Transport und Verkehr
Universität Duisburg-Essen
lars.habel@uni-due.de

This report illustrates recent developments in the simulation of multi-lane highway traffic. In many countries, legal regulations distinguish between driving lanes and overtaking lanes. Therefore, asymmetric lane change rules are needed. In this contribution, rules and simulation results from our recent publication [1] are shown.

In the present report, vehicular motion is simulated using the CA traffic model by Lee *et al* [3]. In the following, we very shortly introduce the steps of the model, which are necessary to understand our model additions, which follow thereafter. A detailed analysis of the Lee model can be found in [5]. Lane changing is explained in detail in [1].

In contrast to many other CA traffic models, each vehicle n in the Lee model does not have unlimited braking capabilities, but is only allowed to brake $D = 2\,\text{cells/s}$ in each time step $\Delta t = 1\,\text{s}$. The cell size is $\Delta x = 1.5\,\text{m}$. With respect to D, the calculation of the next velocity $v_n^{(t+1)}$ of vehicle n requires to determine a maximum safe velocity $\tilde{c}_n^{(t+1)}$ first. This is the maximum $c_n^{(t+1)}$ that fulfills

$$x_n^{(t)} + \Delta_n^{(t)} + \sum_{i=0}^{\tau_f(c_n^{(t+1)},D)} \left(c_n^{(t+1)} - Di\right) \leq x_{n+1}^{(t)} + \sum_{i=1}^{\tau_l(v_{n+1}^{(t)},D)} \left(v_{n+1}^{(t)} - Di\right) \quad . \tag{1}$$

For the calculation of $\tau_f(c_n^{(t+1)}, D)$ and $\tau_l(v_{n+1}^{(t)}, D)$, it has to be determined, if the vehicle is "optimistic" or "pessimistic". Pessimistic vehicles await a traffic breakdown and maintain a safe distance to their preceding vehicles. Optimistic vehicles only maintain a small gap, that can only compensate small disturbances. The definitions of τ_f, τ_l and the safety gap $\Delta_n^{(t)}$ between vehicles n and $n+1$ can be found in [3].

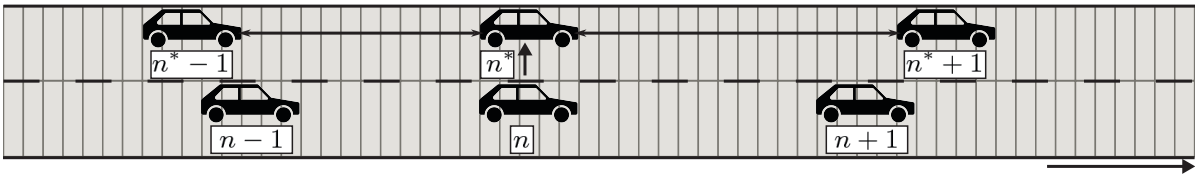


Figure 1: Situation of vehicles during a lane change: Vehicle n attempts to change the lane. On the new lane, it is labelled as n^*.

After $\tilde{c}_n^{(t+1)}$ is calculated, a velocity $\tilde{v}_n^{(t+1)}$ is obtained by cutting $\tilde{c}_n^{(t+1)}$, so that D is respected. Then, $v_n^{(t+1)}$ is obtained by applying the randomisation step on $\tilde{v}_n^{(t+1)}$.

The lane change rule set consists of four components: Rules for guaranteeing safety of a lane change, decision-making rules for lane changes to the right as well as to the left and an optional right lane overtaking ban. The notation of vehicles can be obtained from Fig. 1. In the following, we assume that vehicles drive on the right side of the road and overtake on the left side.

The safety of a lane change is checked using the conditions from the symmetric ruleset presented in [6]. In simulations with more than two lanes, an additional security check has to be performed for lane changes to the central lane(s), because a vehicle from the left and a vehicle from the right could attempt to lane change into the same gap in the same update step.

A vehicle changes to a lane further to the left, when it can drive faster there compared to staying on the present lane. As in the symmetric ruleset, this can be modelled with

$$\tilde{v}_{n^*}^{(t+1)} > \tilde{v}_n^{(t+1)} \quad . \tag{2}$$

A vehicle changes back to a lane further to the right, when its overtaking manoeuvre shall end. This is the case, when staying on the present lane is not advantageous any longer compared to a lane change to the right. On the other hand, the vehicle is not allowed to gain an advantage through a lane change to the right, because it then would offend against the right lane overtaking ban. As a consequence, a lane change to the right is only allowed, if

$$\tilde{v}_{n^*}^{(t+1)} = \tilde{v}_n^{(t+1)} \quad . \tag{3}$$

Generally, this rule conforms with legal regulations, but empirical observations [7] require an additional time gap t_{hlc}, which is ensured through the conditions

$$x_{n+1}^{(t)} - x_n^{(t)} - l_{n+1} \geq v_n^{(t)} t_{\mathsf{hlc}} \tag{4a}$$

$$x_{n^*+1}^{(t)} - x_{n^*}^{(t)} - l_{n^*+1} \geq v_n^{(t)} t_{\mathsf{hlc}} \quad . \tag{4b}$$

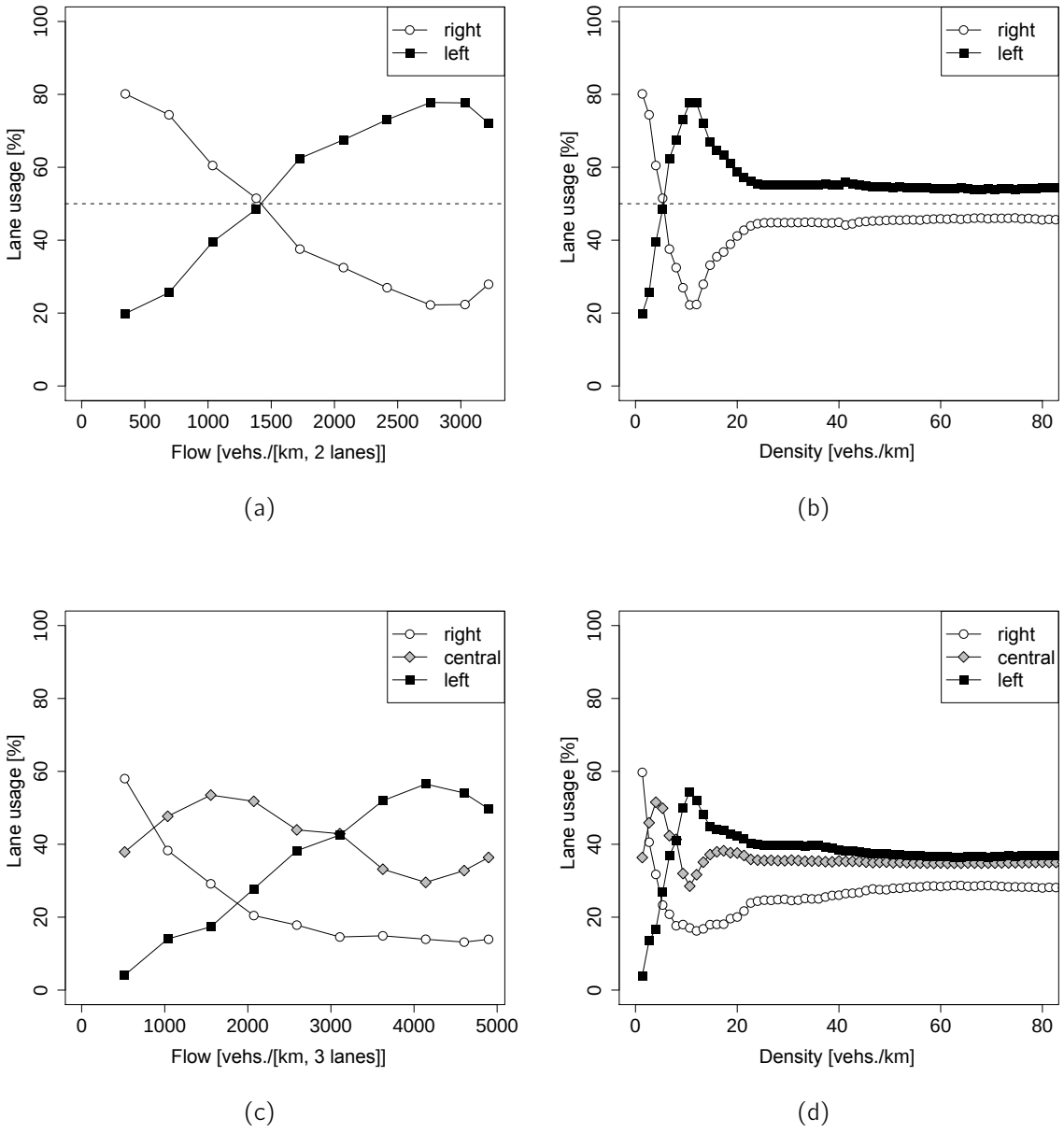


Figure 2: Results of two-lane and three-lane simulations [1]. (a): Distribution of vehicles on the two lanes depending on the cumulative traffic flow. (b): Lane usage depending on the density. (c): Distribution of vehicles on the three lanes depending on the cumulative traffic flow. (d): Lane usage depending on the density.

Below a velocity v_{sym}, we allow lane changing according to the symmetric rules. Overtaking on the right is prevented for higher velocities (see [1] for details).

In Fig. 2, we present simulation results of multil-lane highway traffic with the asymmetric lane change rules, using the default parameter values from [1]. To allow for a comparision with empirical data [2, 4, 7], 10 % of the vehicles were assumed to be trucks driving on the right lane. Fig. 2(a) and Fig. 2(b) show that the presented lane change rules are able to cause an asymmetric distribution of vehicles on the lanes of a two-lane highway. The empirical observations by [7] are reproduced well. In Fig. 2(c) and 2(d), simulation results for three-lane highways are presented. Those have been obtained in the same way as for two-lane highways. As our lane change rules allow for an arbitrary number of lanes, the implementation of this scenario is simple. It can be seen, that the rules are able to reproduce the empirically found location of the lane usage inversion points as well as the qualitative course of the curves [2, 4].

References

[1] Lars Habel and Michael Schreckenberg. Asymmetric Lane Change Rules for a Microscopic Highway Traffic Model. In Jaroslaw Was, Georgios Ch. Sirakoulis, and Stefania Bandini, editors, *Cellular Automata*, volume 8751 of *LNCS*, pages 620–629. Springer, 2014.

[2] Wolfgang Knospe, Ludger Santen, Andreas Schadschneider, and Michael Schreckenberg. Single-vehicle data of highway traffic: Microscopic description of traffic phases. *Phys. Rev. E*, 65(5):056133, 2002.

[3] Hyun Keun Lee, R. Barlovic, Michael Schreckenberg, and D. Kim. Mechanical restriction versus human overreaction triggering congested traffic states. *Phys. Rev. Lett.*, 92(23):238702, 2004.

[4] W. Leutzbach and F. Busch. *Spurwechselvorgänge auf dreispurigen BAB-Richtungsfahrbahnen*. Institut für Verkehrswesen, Universität Karlsruhe, 1984.

[5] Andreas Pottmeier. *Realistic Cellular Automaton Model for Synchronized Two-Lane Traffic - Simulation, Validation, and Applications*. PhD thesis, University of Duisburg-Essen, 2007.

[6] Andreas Pottmeier, C. Thiemann, Andreas Schadschneider, and Michael Schreckenberg. Mechanical Restriction Versus Human Overreaction: Accident Avoidance and Two-Lane Traffic Simulations. In Andreas Schadschneider, Thorsten Pöschel, Reinhart Kühne, Michael Schreckenberg, and Dietrich E. Wolf, editors, *Traffic and Granular Flow '05*, pages 503–508. Springer, 2007.

[7] U. Sparmann. *Spurwechselvorgänge auf zweispurigen BAB-Richtungsfahrbahnen*, volume 263 of *Forschung Straßenbau und Straßenverkehrstechnik*. 1978.

Resource-Efficient LTE Communication based on Connectivity Forecasts

Christoph Ide
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
christoph.ide@tu-dortmund.de

Resource-efficient communication is a key to affordable network services with high Quality of Experience (QoE). In this report, we discuss the predictive Channel-Aware Transmission (pCAT) scheme, which leverages *favorable* channel conditions that require much less spectrum resources than *bad* channel conditions. By the usage of user trajectories and recurring spots with favorable channel conditions, the so-called *LTE connectivity hot spots*, background traffic transmissions can be scheduled by the client according to the expected mobile connectivity and the application data priority. By applying the pCAT scheme, the spectrum consumption of background traffic can be reduced significantly. In addition, the resource-efficient usage of spectrum has also an impact on the power consumption of the mobile devices.

1 Predictive Channel-Aware Transmission (pCAT) Scheme

The pCAT scheme (cf. [1] for more details) bases on time-variant transmit probability $p_T(t)$, which depends on the data priority, the current Signal-to-Noise Ratio ($SNR(t)$) and the average predicted channel quality $\overline{SNR}(t, t+\tau)$ between the current time t and a prediction time window τ. The data priority as well as the strength of the forecast component can be scaled by the two parameters α_n and γ_n (n separates different data priority classes). The formal description of the pCAT transmission probability $p_T(t)$ can be found in [1]. The scheme distinguishes between several cases:

- First of all, it is checked whether the requested transmission falls into a time interval between a minimum time t_{min} and a maximum time t_{max}. If the time duration since the last transmission $t - t'$ is smaller than t_{min}, no data will be transmitted.
- If the time duration since the last transmission exceed t_{max}, the data will be transmitted immediately.
- If the transmission is generally allowed ($t - t'$ is larger than the minimum time t_{min} and smaller than the maximum time t_{max}), the transmission probability is scaled dependent on the channel quality forecast. If a better SNR than the current SNR is predicted, the transmission probability is reduced in contrast to the pure CAT ratio $\left(\frac{SNR(t)}{SNR_{max}}\right)^{\alpha_n}$ by increasing the exponent α_n (that is the global CAT weight) by z_1. The pure CAT also considers the current channel quality, but does not include a channel forecasting component. z_1 depends on the degree of channel improvement $\Delta SNR(t)$, the current SNR and a coefficient γ_n. If a SNR lower than the current one is predicted, the exponent is decreased by a coefficient z_2 to increase the transmission probability, because no better channel conditions are expected and at least the delay can be reduced by no longer buffering the data.

2 Performance Evaluation of pCAT

The subsequent results focus on quantifying the efficiency gains on the spectrum utilization in LTE networks and a proof of concept of CAT in the field is provided.

2.1 Spectrum-efficiency of pCAT

For the performance analysis of CAT regarding spectrum efficiency, a Markovian model that describes the resource utilization of an LTE cell with heterogeneous traffic (Human-to-Human (H2H) and background traffic) and channel conditions, is used. Fig. 1 shows the resulting Physical Resource Block (PRB) utilization and H2H blocking probability if no CAT is used, CAT is applied and for pCAT with one user class. Results regarding more user classes are presented in [2]. The results show that both, the average PRBs utilization and H2H blocking probability can be reduced by CAT, but an additional reduction can be achieved by pCAT. If the channel conditions are moderate and better conditions are predicted (connectivity hot spots) for the near future, pCAT provides a much lower transmission probability than CAT. Furthermore, the larger SNR spread for CAT compared to pCAT leads to some transmissions with very bad channel conditions. These conditions result in an overproportionally longer transmission time due to a very low data rate and therefore a higher PRB utilization as well as a higher H2H blocking probability for CAT in contrast to pCAT (cf. Fig. 1). For pCAT, the maximum waiting time is reached rarely, because the transmission probability is increased if a lower SNR is

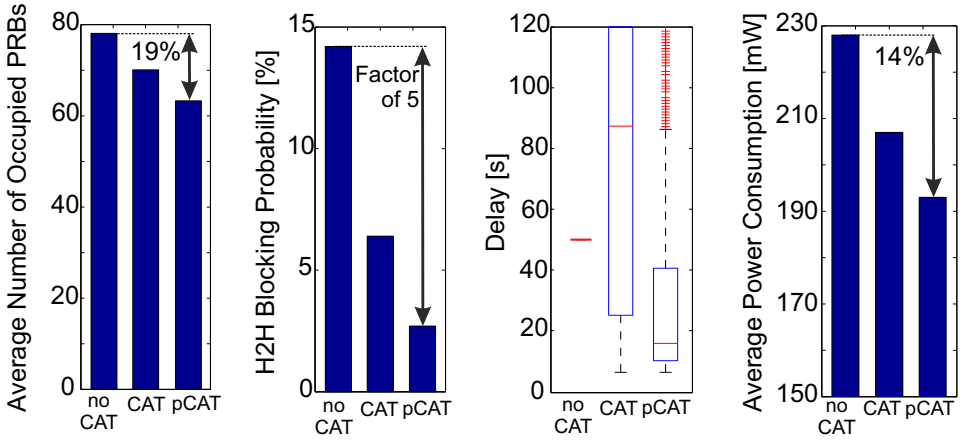


Figure 1: Delay (based on Ray Tracing Analyses and SUMO Trajectories), Resource Utilization as well as H2H Blocking Probability (based on Markovian Model, cf. [3]) and Power Consumption (based on CoPoMo) for Periodical Transmission (50 s), CAT and pCAT.

predicted. The additional delay that has to be inserted in order to benefit from better channel conditions (wait for connectivity hot spots) is illustrated in Fig. 1. These values are gained from SUMO trajectories and an SNR map (cf. [1] for more details). It can be seen that pCAT leads to a significantly smaller delay compared to CAT. This is due to the increased transmission probability if worse conditions are expected. This leads to a transmission just before a region with a low SNR is expected.

For the evaluation of the impact of pCAT on the average power consumption of an LTE UE and therefore its battery lifetime, the Context-aware Power Consumption Model (CoPoMo, collaboration with project A4) has been applied. The results (cf. Fig. 1) show that for the case of no channel aware transmission a long term average power consumption of 228 mW can be observed. Compared to that, the usage of CAT leads to a reduced average power consumption of only 207 mW which corresponds to reduction of 9%. pCAT leads to an even lower power consumption (193 mW) and a reduction of 14%.

2.2 Proof of Concept of CAT in the Field

In this section, a proof of concept of the basic CAT scheme is provided by means of field measurements. The communication with better channel conditions leads to a much faster data transmission. Fig. 2 illustrates the transmission time of 100 kByte uplink data as box plots. It can be seen from the figure that a very wide range of transmission times is given for a transmission without CAT. In contrast to that, for CAT, the maximum transmission time is below 1 s. In addition, the average transmission time for CAT decreases to 390 ms in contrast to 490 ms for a periodical transmission. Fig. 2 also

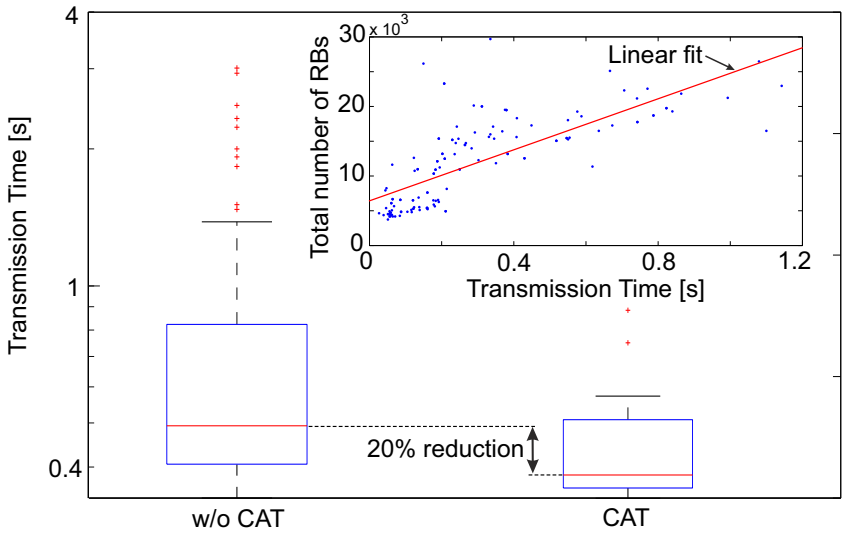


Figure 2: LTE Transmission Time Reduction by CAT as well as Correlation between Transmission Time and Total Resource Allocation (based on Static Field Measurements) for 100 kByte Payload.

provides the relationship between transmission time and required LTE resources in terms of PRBs. These measurements are performed in a static scenario in order to capture the radio resources by means of a real-time spectrum analyzer (cf. [3] for more details). We see that a strong correlation between total number of occupied LTE PRBs and data transmission time is given.

References

[1] C. Wietfeld, C. Ide and B. Dusza, *Resource-efficient Wireless Communication for Mobile Crowd Sensing*, Proc. of the 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Fransisco, USA, Jun. 2014.

[2] C. Ide, L. Habel, T. Knaup, M. Schreckenberg and C. Wietfeld, *Interaction between Machine-Type Communication and H2H LTE Traffic in Vehicular Environments*, Proc. of the IEEE 79th Vehicular Technology Conference (VTC-Spring), Seoul, Korea, May 2014.

[3] C. Ide, B. Dusza and C. Wietfeld *Client-based Control of the Interaction between LTE MTC and Human Traffic in Vehicular Environments*, IEEE Transactions on Vehicular Technology, Jul. 2014, in press.

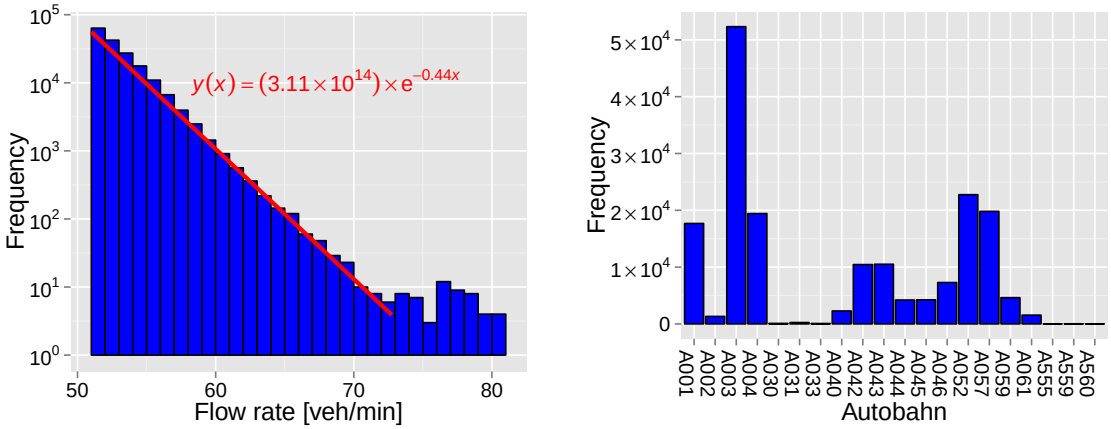
High Flows in Traffic Data

Thomas Zaksek
Physik von Transport und Verkehr
Universität Duisburg-Essen
thomas.zaksek@uni-due.de

High vehicular flow rates of traffic (exceeding around 50 vehicles per minute and lane) are an interesting subject of research for different reasons. At high vehicle flow rates, a transition from free to congested flow is likely to occur—resulting in a considerable decrease of the flow rate and significant changes of other traffic characteristics such as the average velocity. But states of high traffic flow are not only interesting from a physical point of view. At high flow rates, the road is operating close to its optimum. Therefore, it is is also of practical importance to investigate under what conditions these so-called high-flow states occur. In this contribution results from our recent analysis (see [2]) are presented.

1 Introduction

The analysis is based on detector data provided by more than 3000 loop detectors during December 19, 2011 and May 31, 2013 on the motorway network of the state of North Rhine-Westphalia (Germany). Inductive loop detectors still are the most common source of traffic data: for each (1 min)-interval, loop detectors count the number of passing vehicles, they measure the vehicles' velocity distinguished by vehicle type (passenger cars and trucks), and they determine the fraction of time within they are occupied by passing vehicles. The analysis was restricted to 178 800 measurements exhibiting high flow characteristics (flow $>$ 50 veh/min) and to valid values only for each of the just mentioned observables. Figure 1 shows the frequencies of high-flow states depending on the corresponding flow rate (1(a)).



(a) (b)

Figure 1: (a) Frequency of high-flow states and (b) the number of the Autobahn.

It turns out that, up to flow rates of about 73 veh/min, the frequencies of high-flow states of a given flow rate J follow the power law

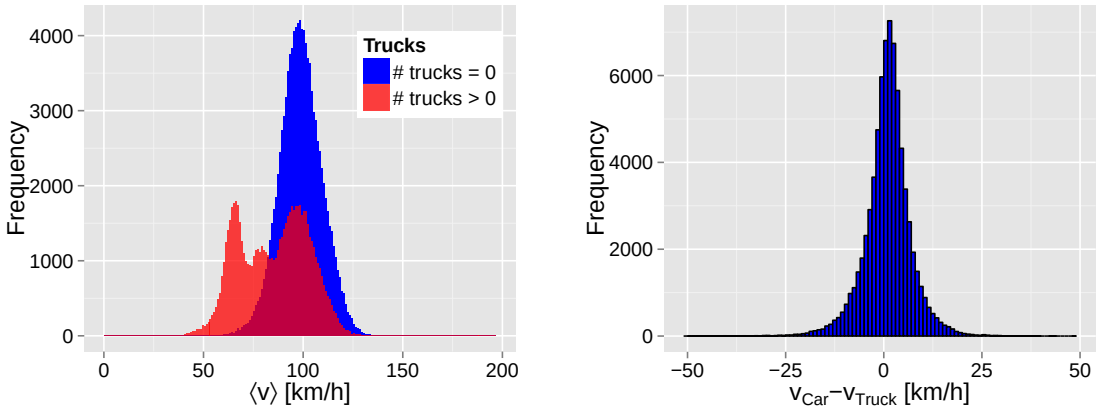
$$\text{frequency} = A \times \exp(-\alpha \times J)$$

with $A \approx 3.11 \times 10^{14}$ and $\alpha \approx 0.44$.

For the small sample of data sets with a higher flow rate (less than 0.03%) it is not clear whether they actually deviate from the power law or whether these data sets indicate erroneous measurements: As we will see, the average velocity of high-flow states varies between 60 km/h and 120 km/h. Therefore, high flow rates with an average time headway of 1 s and less pose an actual risk for drivers. Even though such time headways have already been observed at similar velocities for single vehicles [1], it may be doubted whether this behavior can be observed for a sequence of 70 (and more) vehicles.

2 Velocities of High-Flow States

The average velocities of high-flow states range from approximately 60 km/h to 120 km/h. In figure 2(a), one can see the distribution of the average velocities. For better analysis, these measurements were subdivided into two classes: (i) measurements without trucks (in this case the depicted average velocity is identical to the average velocity of all cars on the road) and (ii) measurements in which at least one vehicle was identified as a truck. Figure 2(b) shows the difference in the average velocities for the observed high flow states. This histogram illustrates the synchronization of average velocities in high traffic flow: the distribution's mean is at 0 km/h (0.52 km/h) with a variance of



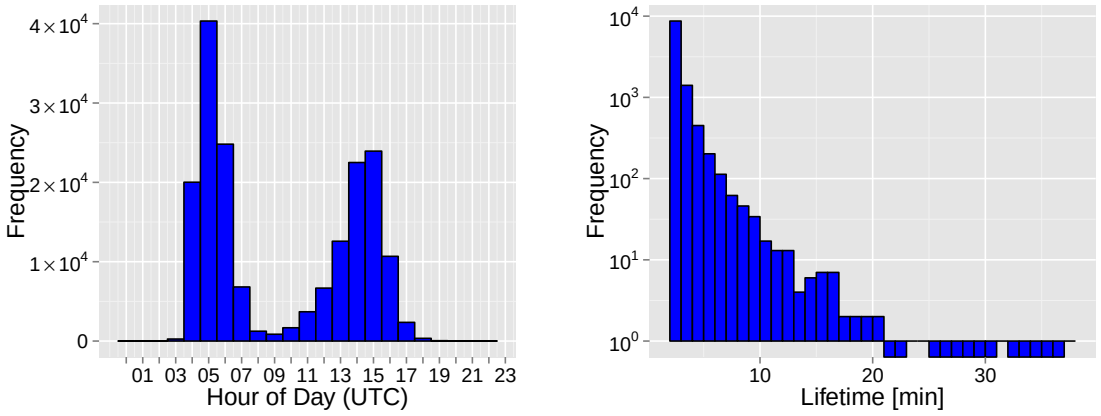
(a) (b)

Figure 2: Average velocities $\langle v \rangle$ of high-flow states and the difference in the average velocities of trucks (v_{Truck}) and cars (v_{Car}).

$\sigma^2 \approx 35\ (\text{km/h})^2$. Moreover, the resulting distribution is strongly peaked around its mean (leptokuric with a sample excess kurtosis of 3.89).

3 Temporal Occurrences & Lifetimes

The histograms given in figure 3 show the temporal occurrence and the duration (i.e., lifetime) of high-flow states. If one considers that a high flow rate indicates a high traffic volume, the results of figure 3(a) are easy to understand. High-flow states occur on work days during peak-hours. At these times, there is a huge demand of commuters (i.e., many passenger cars) traveling to or from work.As the lifetime of a high-flow state we defined the number successive (1 min)-intervals that were classified as "high-flow state". The resulting distribution of lifetimes is given in figure 3(b). One can easily see that such states hardly last longer than a few minutes. This observation only confirms the long-known metastable character of traffic flow: An increased flow rate also increases the probability of a traffic breakdown [3, 4]. Especially at flow rates such as the ones considered in this article traffic flow is very unstable and long-lasting high-flow states could not be expected.



(a) (b)

Figure 3: Temporal distribution of high-flow states depending on (a) the hour of day. (The hour of day is given as UTC. The actual hour of day by adding one or two hours—depending on daylight saving time.) (b) The frequency of successive measurements classifying as high-flow.

References

[1] Cécile Appert-Rolland. Experimental study of short-range interactions in vehicular traffic. *Phys. Rev. E*, 80:036102, 2009.

[2] Florian Knorr, Thomas Zaksek, Johannes Brügmann, and Michael Schreckenberg. Statistical analysis of high-flow traffic states. In *Proc. Traffic and Granular Flow '13*. Springer, 2014.

[3] Bhagwant Persaud, Sam Yagar, and Russell Brownlee. Exploration of the breakdown phenomenon in freeway traffic. *Transp. Res. Rec.*, 1634:64–69, 1998.

[4] Andreas Schadschneider, Debashish Chowdhury, and Katsuhiro Nishinari. *Stochastic Transport in Complex Systems: From Molecules to Vehicles*. Elsevier Science, Oxford, 2010.

Projekt C1
Feature selection in high dimensional data for risk prognosis in oncology

Katharina Morik Alexander Schramm

Evaluation of a chemical JARID1C inhibitor in neuroblastoma cell lines

Kathrin Fielitz
Onkologisches Labor - Universitätskinderklinik Essen
Universität Duisburg-Essen
Kathrin.Fielitz@uk-essen.de

Neuroblastoma ist the most common solid extracranial tumor in childhood. Since genetic alterations are the major cause for this malignancy, we are seeking for genes relevant for neuroblastoma development and progression. These genes could serve as target structures for medical treatment. We were already able to identify JARID1C as a gene important for neuroblastoma biology. We used the small molecule inhibitor Pbit for JARID1C inhibition in SHEP and IMR5 neuroblastoma cells and were able to show morphological changes and an increase in apoptosis. Yet we did not detect a methylational shift in any of the cell lines and concluded that the effect of Pbit in neuroblastoma cell lines must be an off-target effect.

Neuroblastoma is the most common solid extracranial malignancy of childhood [1], showing a large heterogeneity in the clinical course. While unfavorable tumors proceed aggressively, quickly and fatally, tumors with favorable biology even happen to disappear without treatment. In order to treat children suffering of neuroblastoma in the best possible way, prediction markers are needed which can inform us about proceeding and outcome of the illness. We already know that factors like age at diagnosis, tumor stage and MYCN amplification status have an enormous impact on the patient's outcome, yet these markers are not always sufficient. Therefore new marker genes are needed.

We analyzed Affymetrix exon array data from 113 patients [2]. From this data, we identified the histondemethlyase JARID1C as being important for neuroblastoma biology, which we already introduced. JARID1C is upregulated in primary neuroblastoma tumors with bad prognosis (Fig. 1A), yet independent of the MYCN amplification status

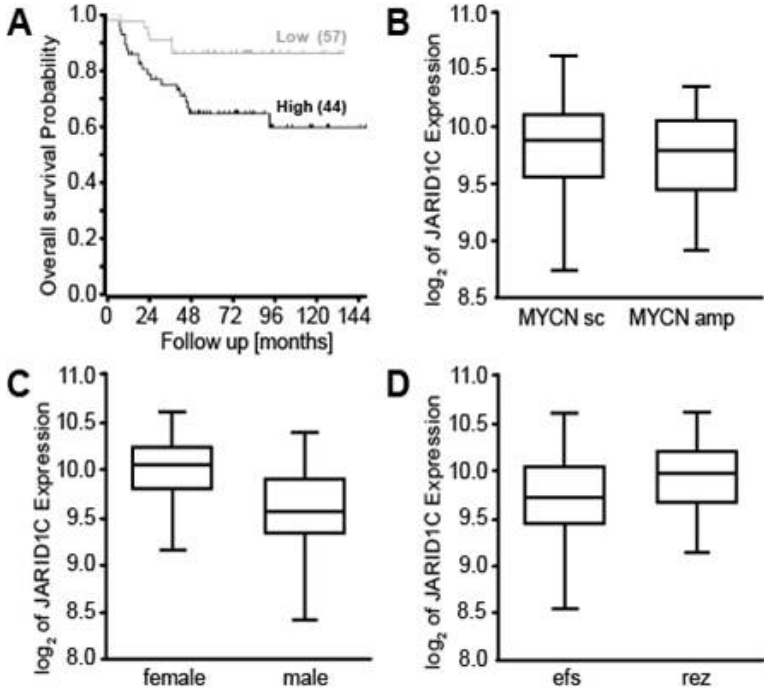


Figure 1: **JARID1C expression in primary neuroblastoma** - High JARID1C expression is correlated with poor survival **(A)** and poor outcome **(D)**. It is independent of MYCN amplification status **(B)** and is expressed higher in females **(C)**.

(Fig. 1B). The JARID1C encoding gene is located on the X-chromosome and was first identified as cause for X-linked mental retardation [3]. We therefore analyzed whether the JARID1C expression differs between genders. Female patients have a slightly higher expression than male patients (Fig. 1C). A high expression is also correlated with poor outcome (Fig.1D).
We were already able to show that siRNA mediated inhibition of JARID1C causes an increase of apoptosis in cells, concluding that JARID1C is essential for neuroblastoma cell survival and could serve as a target structure for medical treatment.

2-4(4-methylphenyl)-1,2-benzisothiazol-3(2H)-one (Pbit) was identified as a JARID1B inhibitor, which also inhibits JARID1C in breast cancer cells [4]. Sayegh et al. were able to demonstrate that Pbit treatment decreased levels of trimethylation at lysin 4 at histone 3 (H3K4me3) and also attenuates cell proliferation in breast cancer cell lines [4]. Since the siRNA mediated inhibition of JARID1C had an effect on cell survival, we tested whether chemical inhibition of JARID1C through Pbit has comparable effects.
Since the siRNA mediated inhibition of JARID1C had an effect on cell survival, we tested whether chemical inhibition of JARID1C through Pbit has comparable effects. We used SHEP and IMR5 cells, as examples for MYCN single copy and MYCN amplified cell lines and determined the half maximal inhibitory concentration (IC50). Thereafter we treated the cells for 72 hours with the Pbit, using the IC50 concentrations (SHEP: 12,5uM;

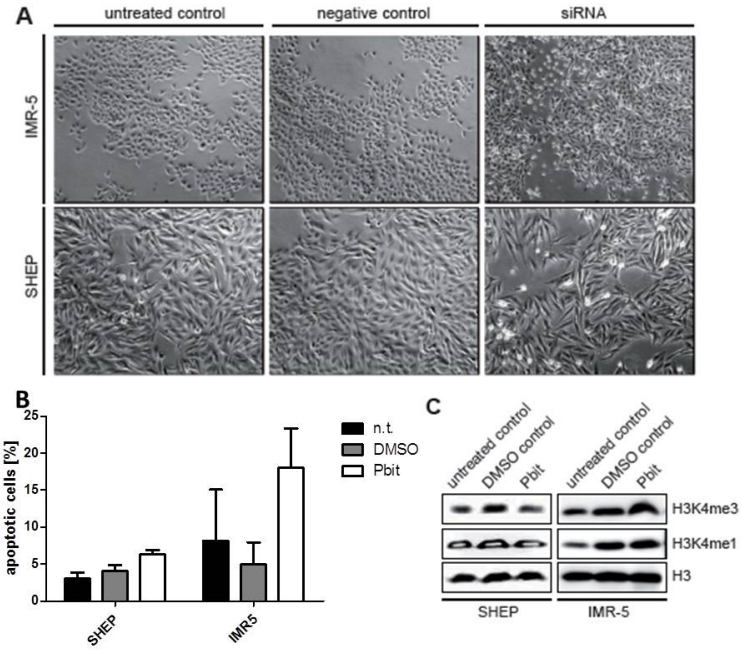


Figure 2: **Pbit treatment on neuroblastoma cells** - **(A)** SHEP and IMR5 cells treated with Pbit show morphological changes and an increase in apoptosis **(B)**. There is no methylational shift decteable **(C)**.

IMR5: 7,5uM). We were able to show that neuroblastoma cells undergo morphologic changes upon Pbit treatment. Comparable to siRNA treatment SHEP cells present a "bubbly" appearance, characteristical fpr apoptosis. Unlike in siRNA transfected cells, we observe this phenotype in IMR5 cells also (Fig. 2A). In order to confirm an increase in apoptosis, we performed FACS (**F**luorescence **a**ctivated **c**ell **s**orting) analysis. In fact we could detect a higher fraction of apoptotic cells compared to controls (Fig. 2B).
Since the JARID1C activity should be inhibited through Pbit,we expected changes in the methylation status at H3K4. We expected to see, as we did after siRNA transfection, a decrease of trimethyl of H3K4 (H3K4me3) and an increase of monomethyl (H3K4me1), as the demethylating activity of the histonedemethylase should be inhibited. Yet we were not able to detect any methylational shift in treated cells and therefore concluded that the effects we see upon Pbit treatment must be off-target effects.

References

[1] Oberthuer, A.; Berthold, F.; Warnat, P.; Hero, B.; Kahler, Y.; Spitz, R.; Ernestus,

K.; König, R.; Haas, S.; Eils, R.; Schwab, M.; Brors, B.; Westermann, F.; Fischer, M.: Customized oligonucleotide microarray gene expression – based classification of neuroblastoma patients outperforms current clinical risk stratification. Journal of Clinical Oncology; 24, 5070 – 5078; 2006

[2] Schramm, A.; Schowe, B.; Fielitz, K.; Heilmann M.; Martin M.; Marschall, T.; Köster, J.; Vandesompele, J.; Vermeulen,J.; de Preter, K.; Koster, J.; Versteeg, R.; Noguera, R.; Speleman, F.;Rahmann, S.; Eggert, A.; Morik, K.; Schulte, J.H.: Exon-level expression analyses identify MYCN and NTRK1 as major determinants of alternative exon usage and robustly predict primary neuroblastoma outcome. In British Journal of Cancer. 2012

[3] Jensen, L.R.; Amende, M.; Gurok, U.; Moser, B.; Gimmel, V.; Tzschach, A.; Janecke, A.R.; Tariverdian, G.; Chelly, J.; Fryns, J.P.; Turner, G.; Reinhardt, R.; Kalscheuer, V.M.; Ropers, H.H.; Lenzner, S.: Mutations in the Jarid 1c gene, which is involved in transcriptional regulation and chromatin remodelling causes X-linked mental retardation. In The American Journal of Human Genetics, 76, 227 – 236; 2005

[4] Sayegh, J.; Cao, J.; Zou, M.R.; Morales, A.; Blair, L.P.; Norcia, M.; Hoyer, D.; Tackett, A.J.; Merkel, J.S.; Yan, Q.: Identification of Small Molecule Inhibitors of Jumonji AT-Rich Interactive Domain 1B (JARID1B) Histone Demethylase by a sensitive High-throughput Screen. In Journal of Biological Chemistry, 13, 9408 – 9417, 2013

PLXNA4 is an indicator for favorable outcome in neuroblastoma

Melanie Schwermer
Oncology Lab-Children's Hospital Essen
Universität Duisburg-Essen
Melanie.Heilmann@stud.uni-due.de

Neuroblastoma is an embryonal cancer of the sympathetic nervous system and diagnosed in early childhood. The tumor originates from precursor cells of the peripheral nervous system and arises in a paraspinal localisation in the abdomen or chest. The clinical presentation of this tumor can be very heterogeneous. Thus, it is very important to understand neuroblastoma relevant pathways and find new neuroblastoma markers to enable a better outcome prediction. The search for such genes was here based on "Exon Array" analysis interrogating all mRNAs and their variants. We identified Plexin A4 as a gene correlated with a good prognosis and expression of TrkA, a marker for favorable outcome. PlexinA4 is a semaphorin receptor and plays a role in axon guidance in the developing nervous system but its function in tumorgenesis is almost unknown.

Neuroblastoma (NB) is the most common and deadly solid tumor in childhood. It accounts for 7-10 % of all childhood cancers. NB derives from the neural crest and usually arises in a paraspinal localisation in the abdomen or chest [1, 2]. The median age at diagnosis is 17 months and the incidence of neuroblastoma is 10.2 cases per million children under 15 years [3, 4]. This kind of cancer exhibits diverse and often dramatic clinical behavior. To enable a better outcome prediction, it is essential to extend the knowledge of the molecular mechanisms in NB. In the last years many genetic features correlating with clinical outcome could be identified. It is known that an increased gene copy number of MYCN is associated with a poor outcome, whereas a high expression of the neurotrophin receptor TrkA is a favorable indicator [1]. To contribute to a better risk assessment, this project deals with the identification of NB relevant genes. Gene expression profiling identified Plexin A4 (PLXNA4) as a marker for excellent overall survival

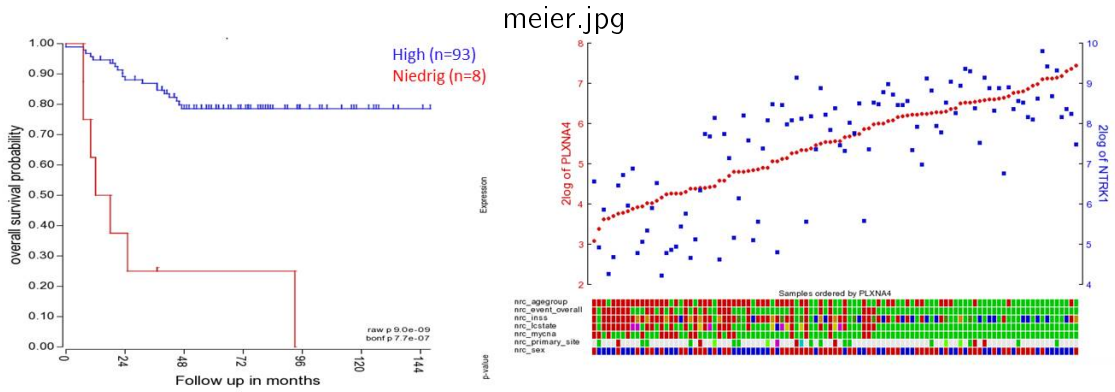


Figure 1: **PLXNA4 expression is an indicator for favourable outcome.** Kaplan-Meier analyses of 101 primary NBs reveals that high PLXNA4 expression is correlated with a good prognosis (p=7.7*10-7). PLXNA4 mRNA expression is also linked to the expression of the neurotrophin receptor TrkA (p=1.3*10-18).

probability. PLXNA4 is a semaphorin receptor with high affinity for SEMA 3A, 6A and SEMA 6B. Semaphorine signalling has an impact on cancer progression and metastasis and it is already known that PLXNA4 can inhibit tumor cell migration [5]. Up to now there are three known isoforms of PLXNA4 and Exon array expression data reveal that isoform 1 is the isoform which causes the effect on survival. This suggestion is based on the facts that the highest correlation for survival and expression was observed for exon 4 and 5 whereas the correlation of the other analysed exons (1, 2, 6 and 7) was moderate. The only isoform containing exon 4 and 5 but not 1, 2, 6 as well as 7 is isoform 1.

So far we found one PLXNA4 inducible clone showing upregulated mRNA- and protein levels after induction with tetracycline (24 h, 48 h and 72 h). For this clone we investigated the impact of PLXNA4 on cell viability and on cell cycle progression (Fig.3). PLXNA4 overexpression had no effects on these cell functions. In the next steps we want to analyse the effects on migration as well as adhesion and try to clarify the signalling pathway of PLXNA4. We are also interested in finding interaction partners of PLXNA4. Interactions between integrins and Plexins have been described [6]. Additionally our Exon array analyses reveal correlation between PLXNA4 and three integrin subunits (ITGA8, ITGB5 and ITG8B). Thus, integrins could be potential interaction partners of PLXNA4. Integrins are cell surface proteins connecting the cell with other cells or with the extracellular matrix. It will be interesting to analyse the effect of PLXNA4 expression on the heterotypic interaction with NB associated cells such as immune cells and schwann cells.

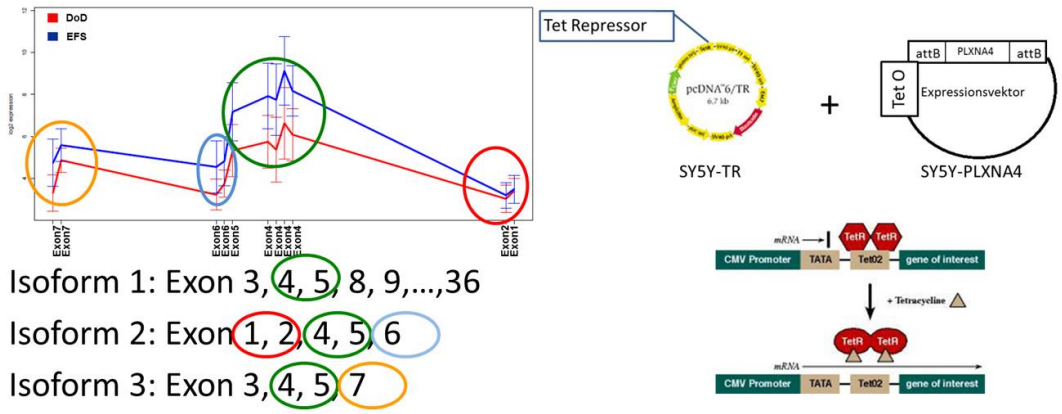


Figure 2: **Establishing a PLXNA4 (isoform1) tetracycline inducible cell line.** The curves show the expression of the individually exons in patients without events (blue) and of patients who died of disease (red). The expression pattern reveals that high expression of exon 4 and 5 but not 1, 2, 6 and 7 is linked with a good prognosis. This indicates that isoform 1 is correlated with a favourable outcome. Therefore isoform 1 was cloned in a tet-inducible vector system and inducibly expressed in a NB cell line.

References

[1] Brodeur G.M. Neuroblastoma biological insight into a clinical enigma. Nature Reviews 2003, 3:203-216

[2] Hoehner JC, Gestblom C., Hedborg F., Sandstedt B., Olsen L., Pahlman S. A developmental model of neuroblastoma: differentiating stroma-poor tumors' progress along an extra-adrenal chromaffin lineage. Lab Invest 1996, 75:659-75

[3] London W.B., Castleberry R.P., Matthay K.K., Look A.T., Seeger R.C., Shimada H., Thorner P., Brodeur G., Maris J.M., Reynolds C.P., Cohn S.L. Evidence for an age cutoff greater than 365 days for neuroblastoma risk group stratification in the Children's Oncology Group. J Clin Oncol 2005, 23:6459-65

[4] Maris J.M. Recent advances in neuroblastoma. N ENG J MED 2010, 362:2201-11

[5] Balakrishnan A, Penachioni J., Lamba S., Bleeker F., Zanon C, Rodolfo M., Vallacchi V., Scarpa A., Felicioni L., Buck M., Marchetti A., Comoglio P., Bardelli A., Tamagnone L. Molecular Profiling of the 'Plexinome' in melanoma and pancreatic cancer Hum Mutat 2009 August ; 30(8): 1167–1174

[6] Lorena Capparuccia and Luca TamagnoneSemaphorin signaling in cancer cells and in cells of the tumor microenvironment – two sides of a coin. Hum Mutat. 2009 August ; 30(8): 1167–1174

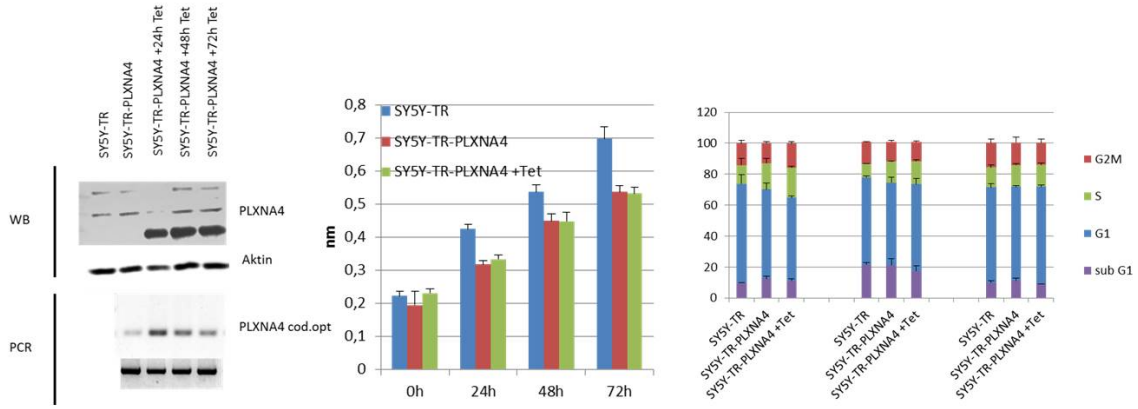


Figure 3: **PLXNA4 overexpression does not alter the cell viability or cell cycle progression.** After induction of the PLXNA4 inducible NB cell line (SY5Y-TR-PLXNA4) PLXNA4 was upregulated on mRNA- and protein-level. The cell viability and the cell cycle progression were not affected by the induction of PLXNA4.

Projekt C3
Multi-level statistical analysis of high-frequency spatio-temporal process data

Roland Fried Wolfgang Rhode

Detection of Abrupt Signal Changes in Time Series

Sermad Abbas
Statistik in den Biowissenschaften
Technische Universität Dortmund
sermad.abbas@tu-dortmund.de

A simple method to detect abrupt signal changes in time series is the application of a two-sample test in a moving window. It is splitted up into two subwindows which are then compared by the test statistic. This report gives some results on the relationship between the Average Run Length (ARL) and a significance level α of a test, when there is no change in the signal. They indicate that the relationship depends on the underlying distribution if the test is not distribution-free. Therefore a desired ARL will not always be attained under all distributions.

The aim of this work is to detect abrupt jumps in the signal of a time series, which is allowed to have a slow nonmonotonic trend. Let $(Y_t)_{t\in\mathbb{N}}$ be a time series. The additive component model is used to describe the random variables:

$$Y_t = \mu_t + \varepsilon_t + \eta_t,\ \ t \in \mathbb{N}.$$

Here, $(\mu_t)_{t\in\mathbb{N}}$ is the true but unknown underlying signal of the time series. It is assumed to exhibit only a few abrupt jumps. The random variables $(\varepsilon_t)_{t\in\mathbb{N}}$ are additive random noise with expectation $\mathrm{E}(\varepsilon_t) = 0$ and variance $\mathrm{Var}(\varepsilon_t) = \sigma^2 > 0$. The outlier-generating process $(\eta_t)_{t\in\mathbb{N}}$ is zero for the majority of the time but leads sporadically to large absolute values.
Two-sample tests in a moving time window of width $n = h + k,\ n, h, k \in \mathbb{N}$, are used to identify jumps in the signal [3]. The time window at time $t \geq h$ is

$$\boldsymbol{Y}_t = (\underbrace{Y_{t-h+1}, \ldots, Y_t}_{=:\boldsymbol{Y}_{t-}}, \underbrace{Y_{t+1}, \ldots, Y_{t+k}}_{=:\boldsymbol{Y}_{t+}})'.$$

The test window $\boldsymbol{Y}_{t+}$ is compared with the reference window $\boldsymbol{Y}_{t-}$ to test for a change between t and $t+1$. The signal is assumed to be locally constant in both subwindows, i. e. $\mu_{t-i+1} = \mu_{t-},\ i = 1, \ldots, h$, and $\mu_{t+j} = \mu_{t+},\ j = 1, \ldots, k$. Then, $\mu_{t+} = \mu_{t-} + \Delta_t$ where $\Delta_t \in \mathbb{R}$ is the unknown jump magnitude between t and $t+1$. Several two-sample tests for the location problem are considered to test the null hypothesis $H_{0,t} : \Delta_t = 0$ at time t. They are constructed by standardizing an estimator for the location difference with a scale estimator [2] and are based on

- The difference of the arithmetic means (t-test).
- The difference of the one-sample Hodges-Lehmann estimators (HL1-test).
- The two-sample Hodges-Lehmann estimator (HL2-test).
- The difference of the sample medians (MD-test).

They are compared with the two-sample median test [7], the two-sample Wilcoxon rank-sum test [5] and the Shewhart control chart [6]. The latter subtracts a prespecified reference value μ_0 from the arithmetic mean of $\boldsymbol{Y}_{t+}$. The difference is standardized with a given standard deviation for the random variables. If it is unknown, it can be estimated by the empirical standard deviation of the test window.
The tests are applied in small time windows to justify the assumption of a locally constant signal even in trend periods. The distributions of the tests based on the HL1, the HL2 and the MD estimators are unknown in finite samples [2]. A possible solution to derive a distribution for the test statistics is the permutation principle. This leads to distribution-free tests but the procedure is very time consuming if it has to be applied in each time window. Two alternatives to get to a test decision are considered:

1. A reference distribution is calculated by using the permutation principle only in the first time window with n observations. The critical values are used for each of the following windows. This implies the assumptions that all observations in the time series come from the same distributional class and the critical values are a good approximation for the true critical values.
2. The distribution is approximated by simulation under the assumption of a standard normal distribution. This has the disadvantage of a distributional assumption.

To compare the tests, the *Average Run Length* (ARL) is used. It is the expected duration to reject the null hypothesis for the first time [1], which is given by the number of subsequent tests. If there is no jump in the signal, it is in control and the ARL is called the in-control ARL (ARL_0). If there is a jump, the signal is out of control. The corresponding out-of-control ARL (ARL_1) depends on the jump height. An underlying assumption is that the shift occurs at the beginning of the surveillance [4]. To compare methods by the ARL, the ARL_0 is fixed and the ARL_1 is evaluated for different jump heights.

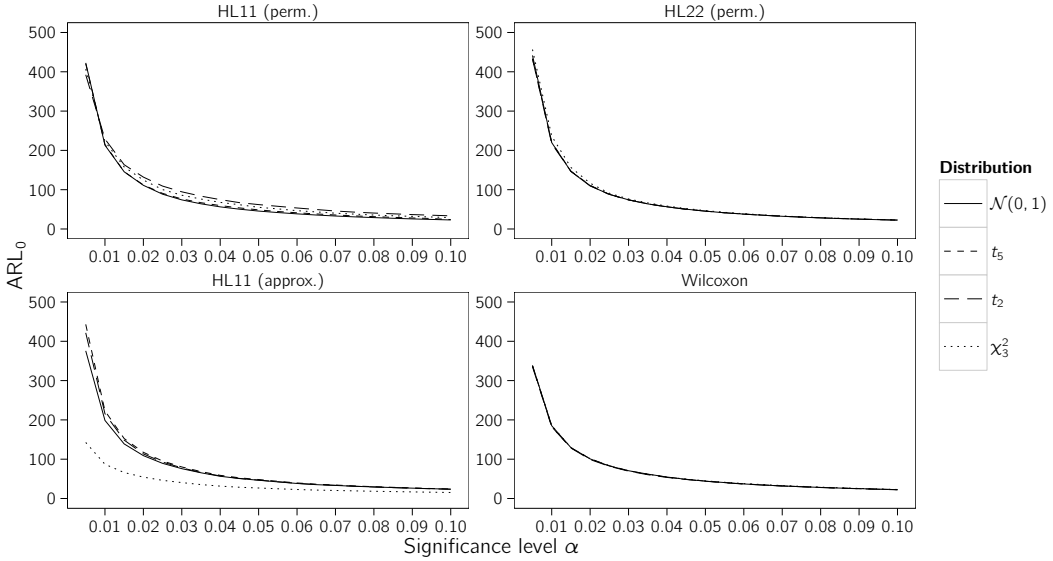


Figure 1: ARL_0 against the significance level α for a test based on the one-sample Hodges-Lehmann estimator and the two-sample Hodges-Lehmann estimator using a permutation principle, a test based on the one-sample Hodges-Lehmann estimator which uses simulated critical values under the $\mathcal{N}(0,1)$ assumption and the Wilcoxon rank-sum test (from top left to bottom right).

A simulation study is conducted to analyze the tests with respect to their ARL_0 behaviour. Normally, the significance level α is prespecified, but in this context the ARL_0 should be fixed. Hence, the relationship between the ARL_0 and the significance level will be described for different distributions. In the simulation, the subwindow widths $h = k = 10$ are used. For the random noise the standard normal distribution ($\mathcal{N}(0,1)$), the t_5-distribution, the t_2-distribution and the χ^2_3-distribution are chosen. In total, 21 tests have to be examined.

Figure 1 shows the simulation results for some of the tests. The significance level is plotted against the estimated ARL_0 separated by the different distributions. The tests in the upper row are based on the HL1 and the HL2 estimator and use the aforementioned permutation principle. The test in the bottom left corner is based on the HL1 estimator and uses the simulated distribution under the $\mathcal{N}(0,1)$ assumption. The results in the bottom right corner belong to the Wilcoxon rank-sum test.
In total, only the nonparametric tests have a distribution-free ARL_0 function. The applied permutation principle does not lead to a distribution-free ARL_0 as can be seen for the HL11-test. In some cases the results are nearly distribution-free for the ARL_0, like for the HL22-test. The normal approximation leads in none of the considered cases to a distribution-free ARL_0.

The simulation results indicate the following relationship between the ARL_0 and the

significance level α:

$$\begin{aligned} \mathrm{ARL}_0 = \beta_0 \cdot \alpha^{\beta_1} &\Leftrightarrow \log \mathrm{ARL}_0 = \log \beta_0 + \beta_1 \cdot \log \alpha,\ \beta_0 > 0, \beta_1 < 0 \\ &\Leftrightarrow \alpha = \exp\left(\frac{\log \mathrm{ARL}_0 - \log \beta_0}{\beta_1}\right). \end{aligned}$$

The unknown parameters $\log \beta_0$ and β_1 are estimated with the ordinary least-squares estimator. As the ARL_0 depends on the distribution of the data in some cases, this is done under the $\mathcal{N}(0,1)$ assumption. The coefficient of determination has a value greater than 0.99 in all cases.

Only the nonparametric tests seem to be able to attain a given ARL_0 under any arbitrary distribution. For the remaining tests this depends on the similarity of the ARL_0 curve under the $\mathcal{N}(0,1)$ distribution to the curves for the other distributions. In all cases the curves get more similar with an increasing significance level which means smaller ARL_0 values. For a small ARL_0 like 100, the MD-tests and the HL2-test seem to reach the given value under the considered distributions. For the larger $\mathrm{ARL}_0 = 250$ the MD-tests attain it.

The next research step is to compare the tests in out-of-control situations with different jump magnitudes. It should be assessed how much slower they are in detecting the jumps in comparison to the Shewhart control charts.

References

[1] M. Basseville and I. V. Nikiforov. *Detection of abrupt changes: Theory and application*. Prentice Hall, New Jersey, 1993.

[2] R. Fried and H. Dehling. Robust nonparametric tests for the two-sample location problem. *Statistical Methods & Applications*, 20(4):409–422, 2011.

[3] R. Fried and U. Gather. On rank tests for shift detection in time series. *Computational Statistics & Data Analysis*, 52(1):221–233, 2007.

[4] M. Frisén. Statistical Surveillance. Optimality and methods. *International Statistical Review*, 71(2):403–434, 2003.

[5] M. Hollander and D. A. Wolfe. *Nonparametric statistical methods*. Wiley, New York, 2nd edition, 1999.

[6] H-J. Mittag. *Qualitätsregelkarten*. Hanser, Munich and Vienna, 1993.

[7] D. W. Wayne. *Applied nonparametric statistics*. Houghton Mifflin, Boston, 1978.

The FACT Camera's Photon Response: Calibration & Simulation

Jens Björn Buß
Experimentelle Physik 5
Technische Universität Dortmund
jens.buss@tu-dortmund.de

The FACT experiment (First G-APD Cherenkov Telescope) is the first instrument of its kind that is using silicon-based photo sensors to detect messenger particles from far distant sources (e.g. Galaxies) entering the Earth's atmosphere [1]. An incoming particle induces an air shower of secondary particles that altogether produce blue light flashes.
A camera with fast electronics and sensitive photon detectors is needed, in order to acquire images of those short faint light flashes. By a sophisticated analysis of the collected images information of the primary particles and along with this information of the observed source are gathered. The major aim is to determine the energy spectrum and the flux of particles coming from an observed source.
Annotated data cannot be provided for the observed messenger particles, thus Monte Carlo simulations are needed. A deeper understanding of the camera's response to incoming photons is demanding for the low-level analysis of the collected data as well as for a correct simulation. The following report sketches some of the methods implemented in the low-level analysis software *fact-tools* [3] and the camera simulation software *Ceres* [4].

1 Introduction

The FACT camera features 1440 pixels, each is equipped with a silicon photo multiplier (SiPM) which is accountable for the transformation of Cherenkov light into an electronic signal. The typical waveform of a single photon is displayed in figure 1. Its shape is a

result of the SiPM and the camera electronics. If several coincident photons are detected, the resulting signal is a superposition of single photon pulses.

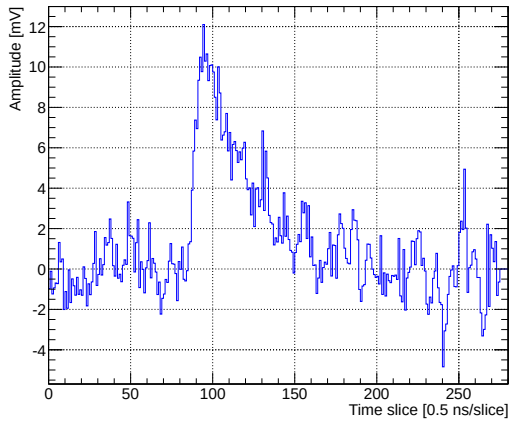


Figure 1: Typical waveform of single photon measured in a single pixel of the FACT camera.

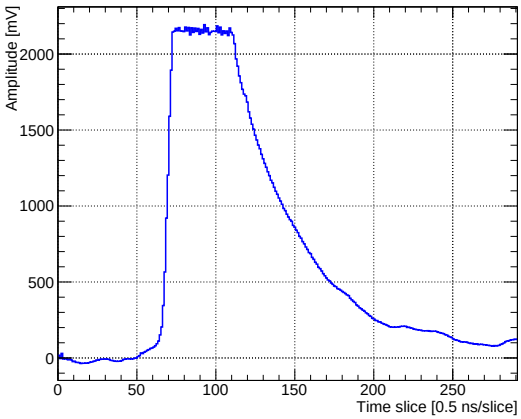


Figure 2: Waveform of saturated signal in a single pixel of the FACT camera.

The first analysis step for the data taken with FACT (or a telescope alike) is the reconstruction of air showers from the photon charge (number of photons) and arrival time of Cherenkov photons in the cameras pixel [6]. A measure for the photon charge is determined from the ratio of the fixed width integrals of the signal waveform and a single photon pulse. This extraction step requires knowledge of the integral of a single photon pulse, hereafter called *gain*. A appropriate method for this factor is the generation of a so called single p.e. spectrum and is explained in the next section.

Going up to higher energies, the photon quantity of very bright air showers may exceed the dynamic range of the readout electronics. The dynamic range of the used analog-to-digital converter reaches up to $\sim 2\,V$. Signal amplitudes above this limit are simply cut off, as depicted in figure 2. These saturation effects have to be taken into account in the low-level analysis. An alternative approach to the extraction of saturated signals is sketched in the course of this report.

2 Calibration

The gain of an individual pixel is determined from the so called single p.e. spectrum, which contains the distribution of extracted pulse integrals for different photon quantities. A detailed explanation of the generation of such a spectrum and its properties is given in [2] and [5]. The resulting spectrum for a single pixel, see figure 3, shows the integral distribution for signals of one up to five photons. Basically the gain can be estimated by the distance of two neighboring peaks. A more accurate extraction of the gain is

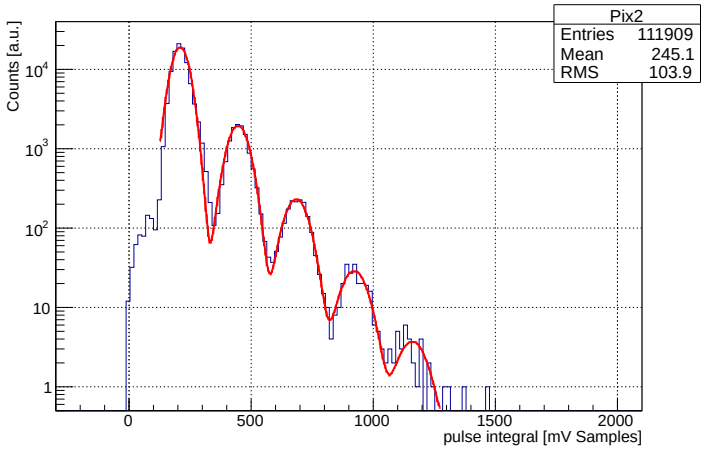


Figure 3: Example of a single p.e. spectrum of single pixel with a fitted model.

performed by fitting a more sophisticated model (red curve) to the spectrum. Details about the fitted model can be found in [5]. Since the model covers all peaks, the extracted gain was computed on more statistics than only the peak distance. On a side note, the fitted model yields additional information about some detector effects e.g. crosstalk and afterpulses, which can be used to tune those effects in the Monte Carlo simulation.

In the low-level analysis of the FACT data the gain is used to calibrate the photon charge. The depicted spectrum was compiled with the data analysis framework MARS [4]. An ongoing implementation of this method into the *fact-tools* [3] yields the opportunity to monitor the gain over time and correct for possible changes.

This gain analysis was so far only performed on data in absence of any background light. Effects of background light to the single p.e. spectrum could not yet be verified, but are currently under investigation in a comparable analysis, which as well is being implemented in the *fact-tools*.

3 Saturation

In the case of saturated pixels the regular photon extraction algorithm [6] underestimates the photon charge systematically, due to the waveform cut off, as shown in figure 2. In order to extract the correct charge for this kind of events an alternative algorithm was implemented into the *fact-tools* [3]. By knowledge of the regular waveform of the signal, one is able to compute the charge. The determination of the standard waveform is explained in [5]. Instead of integrating the signal and normalizing it to the gain, one can also use the width of the signal at a certain threshold to determine the photon charge. The correlation of pulse width and photon charge was determined for different thresholds in the range of 500 mV to 1500 mV and shows an exponential function for

each threshold. Since this method is unbiased to saturation effects, it allows to improve the feature extraction for high-energy showers in the *fact-tools* and gives the opportunity to expand the determination of the energy spectrum in the FACT analysis chain to higher energies.

4 Simulation

In the course of improvement studies of the implementation and settings of *Ceres*, the simulation of the SiPMs was tuned to the results of the gain analysis. For that to happen, random trigger events comparable to the data used in the gain analysis (no background light) were simulated and evaluated with the above described algorithm. The SiPM properties in the simulation were successfully tuned in order to bring the single p.e. spectra of simulation and measurement in agreement.

The handling of saturated pixels can possibly be improved by comparing the entire shape of the well-known waveform to the signal. For usage in *fact-tools* [3] it is necessary to optimize this method to perform with a short runtime, this is one of the main goals of low-level analysis.

References

[1] H. Anderhub et al. Design and operation of FACT - the first G-APD Cherenkov telescope. *Journal of Instrumentation*, 8(06):P06008–P06008, June 2013.

[2] A. Biland et al. Calibration and performance of the photon sensor response of FACT - the first G-APD Cherenkov telescope. *Journal of Instrumentation*, 9:12P, October 2014.

[3] Christian Bockermann and Hendrik Blom. The streams Framework. Technical Report 5, TU Dortmund University, 2012.

[4] T. Bretz and D. Dorner. MARS - CheObs ed. – A flexible Software Framework for future Cherenkov Telescopes. In *Astroparticle, Particle and Space Physics, Detectors and Medical Physics Applications*, pages 681–687, April 2010.

[5] Jens Björn Buß. *FACT - Signal Calibration: Gain Calibration and Development of a Single Photon Pulse Template for the FACT Camera*. Diploma thesis, TU Dortmund, 2013.

[6] Fabian Temme. *FACT - Data Analysis: Analysis of Crab Nebula Data using PARFACT a newly Developed Analysis Software for the First G-APD Cherenkov Telescope*. Diploma thesis, TU Dortmund, 2012.

The MAGIC Monte Carlo production chain and first results of the Mrk 421 data analysis

Katharina Frantzen
Experimentelle Physik 5
Technische Universität Dortmund
katharina.frantzen@tu-dortmund.de

In this report a short summary of the automatic Monte Carlo production chain at the TU Dortmund of the MAGIC telescopes is given. Particular attention is paid to the application of the Monte Carlo files in the analysis of the source Markarian 421 (Mrk421) data of 2012.

1 Introduction

Primary particles emitted by galactic or extragalactic sources produce air showers of secondary particles in the atmosphere. These secondary particles in turn emit Cherenkov light travelling through the atmosphere. Imaging Air Cherenkov Telescopes (IACTs) like the Major Atmospheric Gamma-ray Imaging Cherenkov Telescopes (MAGIC), located on the Canary Island La Palma, can detect this Cherenkov light. The intention of the observations with these telescopes is to discover new sources and to determine their energy spectra. Simulated Monte Carlo data (MC) are essential to reach this goal - they are necessary to distinguish between the wanted gamma showers and unwanted hadron showers and to determine energy, particle type and origin of the shower inducing particle.

2 MC production chain

In order to produce the MCs, an automatic production chain using bash scripts and a MySQL database has been implemented at the available clusters. These are the LiDO cluster (Linux Cluster Dortmund) with ∼3000 CPUs and ∼215TB storage and the PhiDo (Physics cluster Dortmund) with ∼1200 CPUs and ∼200TB storage.
If there is a MC request, the requirements of the request (number of particles to simulate, energy of primary particle, ...) are stored as parameters in the database and passed to the programs (see Fig. 1). After writing the parameters to the database, the jobmanager script, which is permanently running, has a deeper look to the database and starts the whole production chain using the parameters in the database.

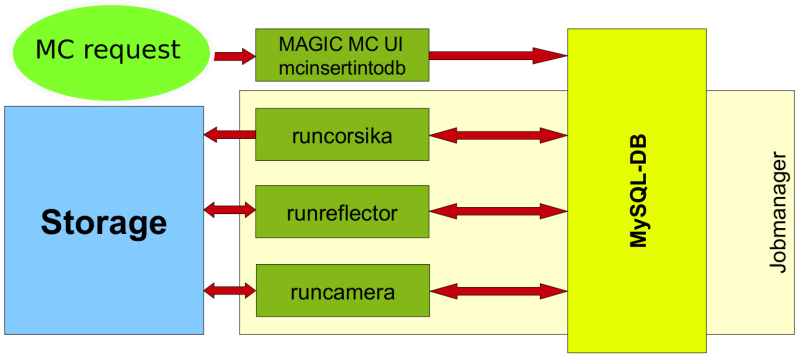


Figure 1: Schematic view of the MC production using runscripts and a MySQL database.

The first program in this chain is called *CORSIKA* [2].It simulates the interaction of the primary particle with molecules and atoms in the atmosphere and generates the Cherenkov light which is simulated the whole way to the telescopes.
The second part of the simulation chain is the simulation of the mirrors, which is done in the program *Reflector* [3]. Reflectivities and arrival times of the Cherenkov photons are simulated for example.
The next step is to simulate the camera of the telescope. The program *Camera* [1] simulates the electronics of the camera and the trigger.
After this simulation chain, the calibration of the data takes place. So the next steps have to be done in the same way for real data and MC data. All these programs used in the calibration chain are programs of the Magic Analysis and Reconstruction Software (*MARS*) [4].
Beginning with *Sorcerer* the calibration and extraction of the signal is performed. Then *Star* contains the image cleaning and calculation of the Hillas parameters. The stereoscopic reconstruction of the showers is done in *Superstar*.

3 Analysis

The data that was analysed is data of Mrk 421 recorded in 2012. It can be separated in two major parts: There are periods in which both telescopes worked fine and stereoscopic observations were performed and there are periods in which only MAGIC-2 was observing while MAGIC-1 was out of order. The analysis of these periods is performed similar, but different programs were used, which are specifically for mono and stereo observations.
After downloading the data, suitable background data and MC files from the Grid, the analysis can start.
At first the data has to be checked to sort out bad data with bad weather, much moonlight, car flashes, etc. This step is performed for background data a Mrk 412 data.
Afterwards Random Forests (RF) are trained on one part of the simulated signal MC files and the background. This step is done within *Coach*.
In *Melibea* the RF are applied to the data and the other part of the MCs and a signal/background separation is performed, having in mind that the signal to background ratio is about 1:1000. Furthermore the energy of the showers in the data is reconstructed.
A possible next step is the calculation of a light curve, which is done by *Flute*.
The light curve shows the flux of the source on one specific analysis period with stereo observations (see Fig.2)

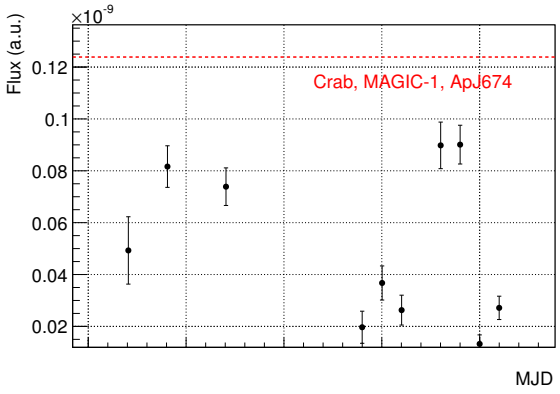


Figure 2: Light curve of Mrk421 stereo data of February 2012.

CombUnfold and *TRUEE* are the programs to unfold the spectrum of the source.
The discribed tasks have to be performed on all parts of the data. An example of the spectrum can be seen in Fig.3.

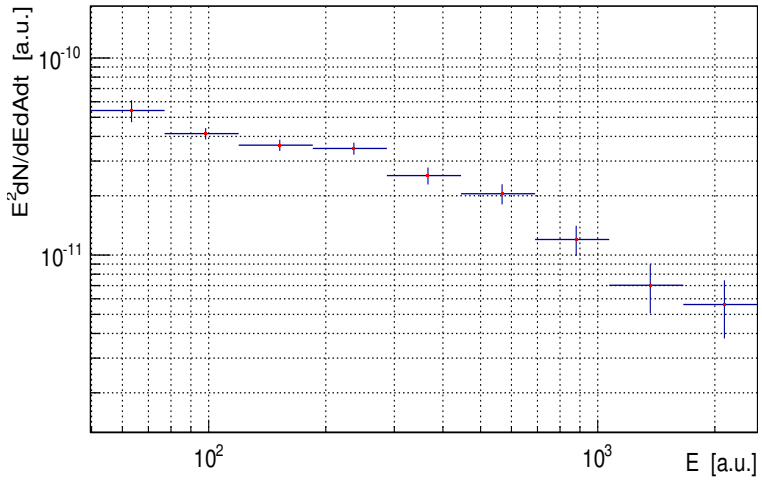


Figure 3: Spectrum of Mrk421 stereo data of February 2012.

4 Conclusion and outlook

This TechReport gave a short résumé about the MC production chain for the MAGIC experiment at the TU Dortmund and pointed out the importance of MC data for the analysis of real sources. Furthermore it showed the first results of the analysis of the Mrk 421 data. The mono parts of the data have to be analysed in the future and a comparison of the unfolded spectra with DSEA, developed in the SFB876, will be made.

References

[1] O. Blanch. How to use the Camera simulation program 0.7. *TDAS notes (internal notes of the MAGIC collab.)*, September 2004.

[2] D. Heck, J. Knapp, J. N. Capdevielle, G. Schatz, and T. Thouw. *CORSIKA: a Monte Carlo code to simulate extensive air showers.* February 1998.

[3] A. Moralejo. The Reflector Simulation Program v.0.6. *TDAS notes (internal notes of the MAGIC collab.)*, November 2003.

[4] R. A. Moralejo, M. Gaug, E. Carmona, P. Colin, C. Delgado, S. Lombardi, D. Mazin, V. Scalzotto, J. Sitarek, and D. Tescaro. MARS: The MAGIC Analysis and Reconstruction Software, November 2010. Astrophysics Source Code Library.

Analysis of high energetic muons in IceCube

Tomasz Fuchs
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
Tomasz.Fuchs@tu-dortmund.de

By analyzing high energetic muons in IceCube a measurement of the prompt muon flux is possible. To do this muon bundles are analyzed which consist mostly of just one muon which carries most of the energy of the bundle. These events are selected using the feature selection mRMR and a random forest. The selection of these events is possible using a random forest. When reconstructing the muon spectrum a flattening in the spectrum for higher energies is expected.

1 Introduction

In the field of astro particle physics galactic or extragalactic particles reaching the earth are analysed. This can be done by satelites, balloons or ground based experiments. IceCube is a ground based particle detector at the geographical south pole and is able to detect high energetic particles which are created in the atmosphere or are extraterrestrial.

When high energetic protons or other elements enter the atmosphere multiple particles are produced. Most of these particles have a long lifetime and lose energy through propagation in the atmosphere. Due to the energy losses their energy spectrum is steeper than the spectrum of the primary particles. At higher energys particles with a short lifetime are created. These particles are called prompt because they decay faster compared to the conventional particles [1]. Because of the quick decay they have the same spectral index as the primary particles.

The cross section to create these particles is not covered by accelerator experiments since the most probable angle is in forward direction. Reconstructing the flux of these prompt muons can provide knowledge about the production cross sections. Also parton distribution functions can be calculated from these cross sections.

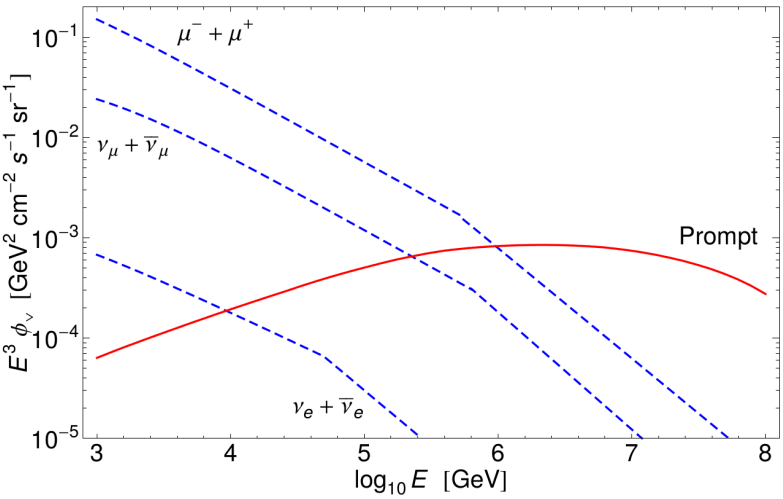


Figure 1: Flux of conventional (blue) and prompt leptons (red).

2 Analysis

When muons enter the IceCube detector they arrive in a bundle. The number of muons in a single event is in a range from 1 to over 1000 muons. To detect the most energetic muon in these bundles the topology of these events is important. Two of these events are shown in figure 2. The left event of figure 2 shows a high energetic muon which carries more than 50 % of the total energy of the bundle. These events will be called leading-muon-events because they have a muon which has most of the energy of the event. The right event of this figure shows a typical muon bundle which does not contain such high energetic muon but the energy of the bundle is distrubuted between all muons in the bundle. These events will be called background-events. The leading-muon-events produce a high energetic stochastic loss in the detector. Using this high energetic stochastic loss to select such events is a good approach since only high energetic muons are able to produce these signatures.

To select the leading-muon-events an advanced data-mining approach is needed since the topologies of these events can be very similar to the background-events. To apply state of the art data-mining procedures to the data and use sophisticated algorithms the software Rapidminer [2] is used.

For the selection over 400 attributes are available which describe an event reaching the detector. The first step of the analysis was to get a set of attributes which is capable to distinguish between leading-muon-events and background-events. To get this set of attributes the mRMR (minimum Redundancy Maximum Relevance) [3] feature selection is used. This algorithm selects a set of features which are highly correlated to identify the leading-muon-events and are least correlated to each other. With this it was possible to reduce the number of attributes to the most important 30 attributes.

With the selected 30 attributes one needs an algorithm to separate the leading-muon-events from the background-events. One of the best algorithms to seperate signal from background events is the random forest. In this analysis a random forest implementation by WEKA is used. This algorithm builds multiple decision trees and selects a random subset of attributes on each node which is considered for the seperation. The best of the random choosen attributes is then selected. To classify the leading-muon-events a random forest with 200 trees, 30 attributes and 5 attributes per node was trained. Also the training of the random forest was validated using a five fold cross validation.

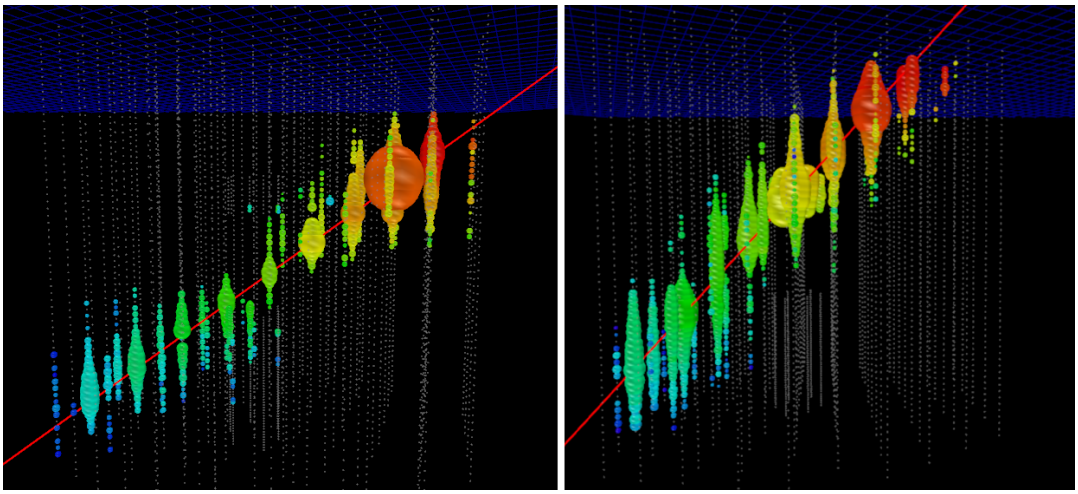


Figure 2: Event with an leading-muon-event (left) and a background-event (right).

3 Results and Outlook

The separation of leading-muon-events and background-muon-events is possible when using the random forest with the settings as mentioned in the section before. This can be seen in figure 3 where the leading-muon-events (blue) and the background-events (red) are shown for different confidences of the random forest. For a confidence level greater than 0.7 the number of leading-muon events is dominating over the background-events.

A similar distribution is derived when applying the trained random forest to the IceCube 2011 data set. This means that it is not only possible to select this leading-muon-events from the simulation but also from the data.

In the future a proper confidence level has to be chosen to have enough events to build a spectrum of high energetic single muons. To do this one has to minimize the number of background-events but also keep as many leading-muon-events as possible. The spectrum of high energetic single muons will be derived using the software TRUEE. With this a first measurement of the prompt component can be done.

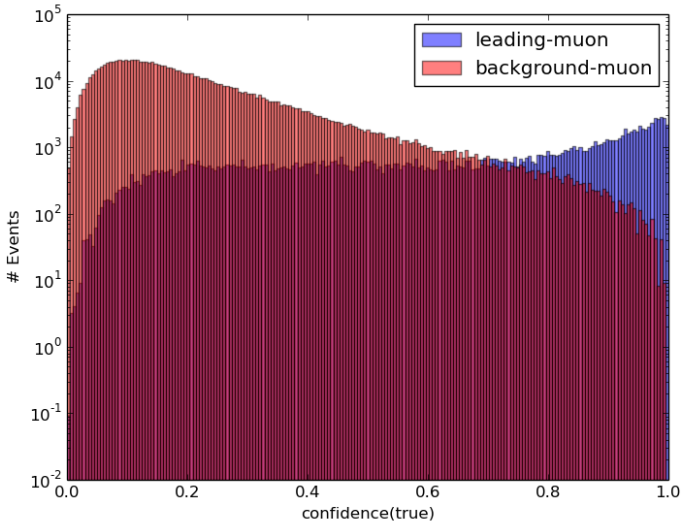


Figure 3: Event classification of leading-muon-events (blue) and background-muon-events (red).

References

[1] Rikard Enberg, Mary Reno, and Ina Sarcevic. Prompt neutrino fluxes from atmospheric charm. *Physical Review D*, 78(4):043005, August 2008.

[2] Ingo Mierswa. A flexible platform for knowledge discovery experiments: Yale–yet another learning environment. 2003.

[3] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1226–38, August 2005.

Periodicity search in a gamma-ray light curve of Mrk421 recorded by FACT

Ann-Kristin Overkemping
Experimentelle Physik 5
Technische Universität Dortmund
ann-kristin.overkemping@tu-dortmund.de

One goal in astroparticle physics is the examination and understanding of the long time behaviour of astrophysical sources like Active Galactic Nuclei. Markarian 421 (Mrk421) is a source of this type and regularly monitored by FACT, the First G-APD Cherenkov Telescope. In this report the recording of a light curve, the temporal evolution of the source's flux, shall be discussed as well as first results of periodicity studies on this light curve will be shown.

FACT observations of Mrk421

FACT is an Imaging Air Cherenkov Telescope (IACT), dedicated to the observations of γ-ray emitting sources. The detection technique is based on the indirect detection of the γ-rays. When these enter the Earth's atmosphere, they induce air showers of secondary particles. When the charged particles of these cascades travel faster than light in the atmosphere, they emit Cherenkov light which can subsequently be detected by IACTs like FACT during the night. [3]

FACT is situated on the Canary Island of La Palma at a height of 2.2 km and it started observations in October 2011 [2]. It is the first IACT which is using a G-APD camera instead of the classic Photomultiplier Tubes. FACT is monitoring several galactic and extragalactic sources on a regular basis. Examples of sources are the Crab Nebula, a supernova remnant, as galactic source and the blazars Mrk421 and Mrk501 as extragalactic sources. [1]

Mrk421 is observed by FACT whenever it is possible. Unfortunately observation times are limited. First of all, Mrk421 is only visible from November until June every year from the telescope site on La Palma due to astronomical circumstances. Secondly, bad weather conditions like rain, snow and clouds can make observations impossible. For this study,

data taken between December 2012 and June 2014 is considered. An additional cut has been applied to the accepted zenith distance range, so that only data with less than 35° distance from the zenith was accepted. A secondary cut was set for the minimum set threshold for the trigger criterion to only include data taken under dark sky conditions. These cuts were applied because the automated analysis used for the Quick Look Analysis (QLA) was optimized for those conditions. All QLA results can be found at the following link: `fact-project.org/monitoring`, where they are presented in form of an excess rate curve. [1]

The resulting light curve of Mrk421 in the considered time range can be seen in Figure 1.

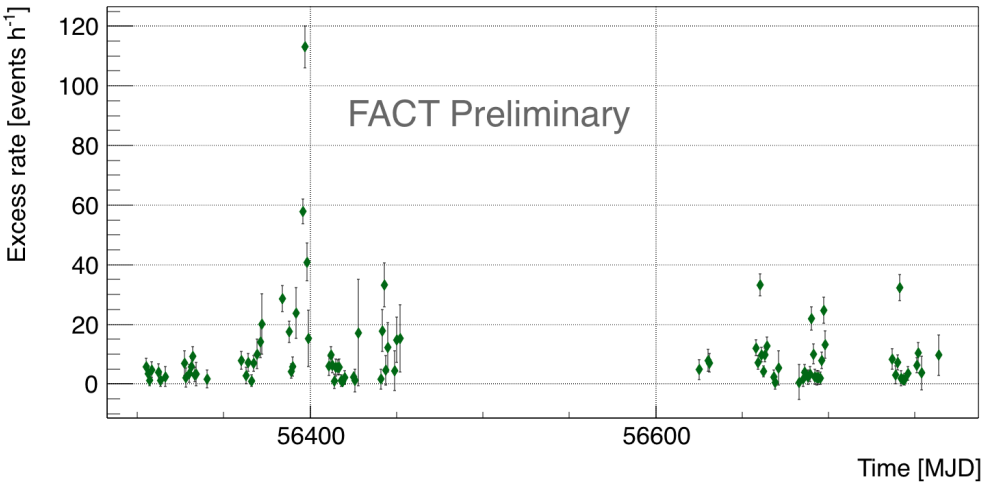


Figure 1: Light curve of Mrk421 from 11th January 2013 (MJD 56305) to 1st June 2014 (MJD 56764) recorded by FACT. The daily light curves of all observed sources by FACT, including the one shown here, can be found on `fact-project.org/monitoring`.

Periodicity search in light curves

For the search of periodicities in light curves, the R tool `RobPer` was developed by [6]. It can handle data consisting of a date, a flux value and a corresponding flux error and it was developed to also be able to handle unevenly sampled data and take the measurement errors by weighting into account. [6]

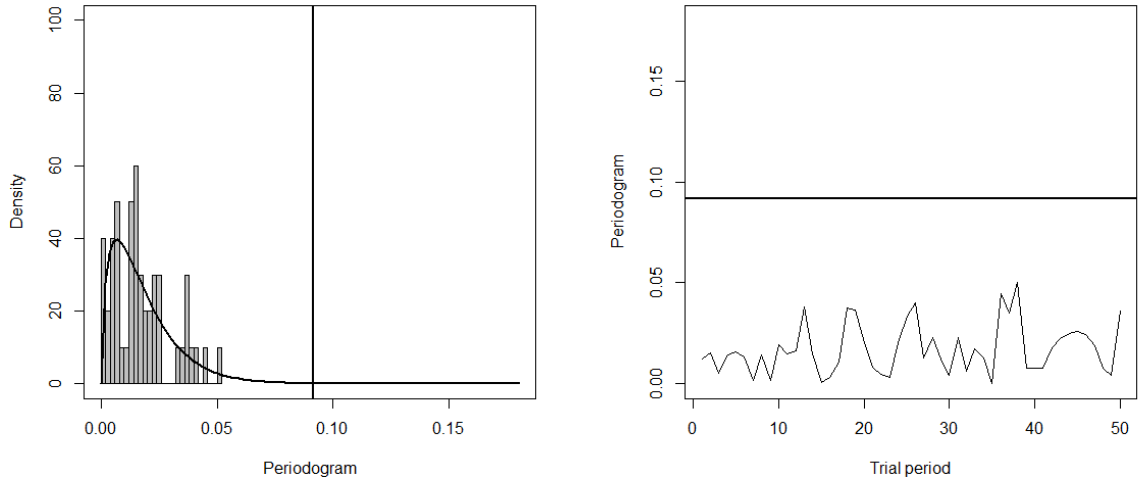
A function, which is repeated with a chosen period, is modelled to the light curve points. In `RobPer`, one of six different functions can be chosen. For the modelling, seven different regression methods are available. Then the periodogram is determined for all trial periods in the selected range. The periodogram describes the goodness of the model for the existent light curve. The goal is to find an outstanding trial period. Then an underlying periodicity would be found. [6]

Here a step and a cubic spline function were studied as periodic functions. As regression method, the Huber regression was chosen from the available methods. It is a robust

regression technique which can handle also outliers like high peaks in the light curve, which is achieved by down-weighting these outliers during the modelling [5]. Additionally, the effect of the weighting of the data points by their errors was tested.

With `RobPer` the chosen periodic function is modelled with a periodic length, here between 1 and 50, using the chosen regression method. Then for each trial period the periodogram is calculated. To find out if a period is outstanding and a periodicity is present in the data, the distribution of the periodograms is examined. In Figure 2(a) this distribution can be seen exemplary for the spline function and the Huber regression, not yet taking the data point errors into account. A Beta function is then fitted to this distribution under the assumption of no periodic fluctuations present. Then a valid period can be found if it is above the 95% quantile of the Beta function. The Beta function and the significance threshold are displayed by the black lines. [5]

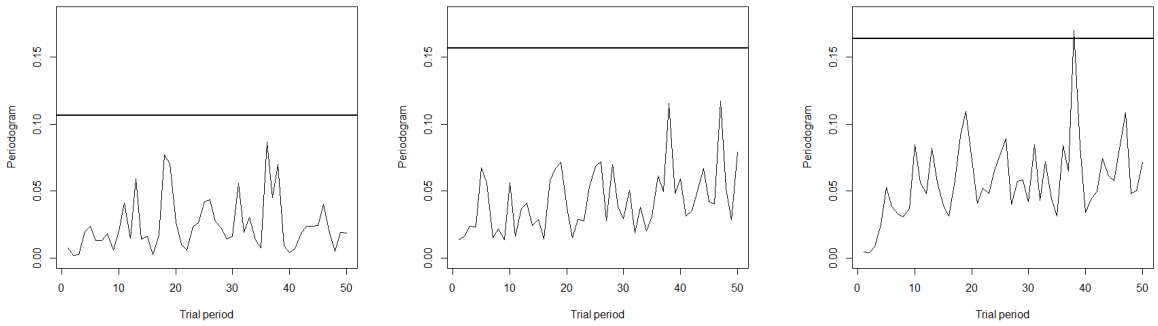
In Figure 2(b) the periodogram for each trial period between 1 and 50 is shown. In this plot the outstanding trial period could be identified if present, but this is not the case here.



(a) (b)

Figure 2: Periodogram results for the spline function and the Huber regression. The weighting according to the measurement errors is not taken into account in this case.

Taking the measurement errors into account when modelling the periodic function to the data points, improves the profile of all periodograms, so that periodograms of the trial periods 19 and 38 days are more pronounced, but not yet significant. This can be seen in Figure 3(a). Then the periodic function is replaced by a step function. In the case the errors are not taken into account, the profile of the periodograms does not improve (see Figure 3(b)). When the errors are included into the calculation, 38 days is found to be a significant trial period (see Figure 3(c)).



(a) (b) (c)

Figure 3: Periodogram results for two different functions using the Huber regression. In (a) the spline function was used and the measurement errors are taken into account. (b) and (c) both use the step function, whereas (b) does not take the errors into account and (c) does.

It can be concluded that the modelling with both periodic functions shows the highest periodogram bars for 19 and 38 days. Taking the measurement errors into account as weighting factor for the modelling, improves the results. In the case of the spline function no significant trial period was found, but for the step function a period of 38 days seems to be a significant trial period.

References

[1] Anderhub, H. et al. (FACT Collaboration). Design and operation of FACT - the first G-APD Cherenkov telescope. *Journal of Instrumentation*, 8:P06008, June 2013.

[2] CERN Courier. Innovative camera records cosmic rays during full moon. November 2011.

[3] Grupen, C. *Astroparticle Physics*. Springer Verlag, Berlin, Heidelberg, 2005.

[4] Punch, M. et al. Detection of TeV photons from the active galaxy Markarian 421. *Nature*, 358:477–478, August 1992.

[5] Thieler, A.M., Backes, M., Fried, R., Rhode, W. Periodicity detection in irregularly sampled light curves by robust regression and outlier detection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6:73–89, January 2013.

[6] Thieler, A.M., Fried, R., Rathjens, J. RobPer: An R Package to Calculate Periodograms for Light Curves Based on Robust Regression. *Technical Report SFB 876 TU Dortmund University*, February 2013.

Development of a datamining based analysis chain for FACT

Fabian Temme
Experimentelle Physik 5
Technische Universität Dortmund
fabian.temme@tu-dortmund.de

FACT (**F**irst G-**A**PD **C**herenkov **T**elescope) is a groundbased imaging air Cherenkov telescope. It detects the Cherenkov light emitted by the secondary particles of a cosmic ray air shower. FACT is interested in detecting the flux and energy spectrum of high energy gamma rays sources, but is overwhelmed by the 1000 times more numerous charged cosmic ray background. Therefore an analysis chain is developed which performs a separation using modern datamining methods to distinguish between gamma ray shower and charged cosmic ray shower. Also new approaches for reconstructing the energy of the primary particle and therefore improving the unfolding of the energy spectrum of the gamma ray source are developed.

1 Introduction

FACT [2] is located on the Canary Island of La Palma on the mountain Roque de los Muchachos at an altitude of about 2200 m. To analyse the raw data the preprocessing software facttools is used. Different analysis steps are performed to calculate a parameter set describing the properties of the triggered event. In the datamining framework RapidMiner [4] a separation between the gamma induced air showers and the charged cosmic ray (mostly protons) induced air showers is applied to the events. The resulting gamma set is unfolded, using the software TRUEE [5]. Figure 1 illustrates the principle design of this datamining based analysis chain.

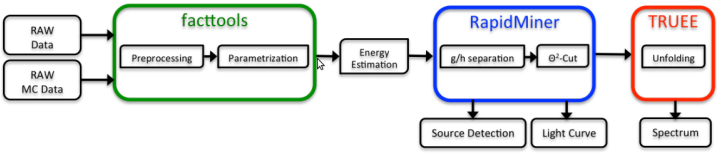


Figure 1: Principle design of the datamining based analysis chain in FACT

2 facttools

The preprocessing software facttools is written within the Java based streams [3] framework. The eventwise analysis of the data is perfect suitable for the usage of streaming application. The easy to use interface with xml control files afford a good approximation level for designing the analysis process, without the need to control the underlying streaming processes.
The processors implemented in facttools perform a low level analysis of the raw data of FACT. Therefore the data is calibrated, to compensate hardware dependent differences between pixels. Than the number of registered photons and the arrival time of this photons are extracted from the timeline. To identify the pixels containing the air shower a cleaning algorithm is applied to the data. The last step calculates different parameters describing the properties of the image of the air shower (The different analysis steps are described in [6]).

3 Energy Estimation

To estimate the energy of the primary particle a random forest regressor and a k nearest neighbor regressor are trained, using simulated monte carlo events. The correlation between the simulated energy (E_{MC}) and the reconstructed Energy (E_{REC}) is shown in figure 2. The random forest regressor clearly shows a better correlation, although it is not umabiguous. Therefore an unfolding of the energy spectrum is necessary (see section 5). The correlation between E_{REC} and E_{MC} is also used to estimate the energy resolution of the telescope as it is displayed in figure 2 (right).

4 Separation

In the datamining framework RapidMiner [4] a separation process to distinguish between gamma induced air shower and proton induced air showers is designed. At first a feature selection is performed on the parameter set calculated by facttools. The minimum redundancy maximum relevance algorithm determines a set of parameters with a high

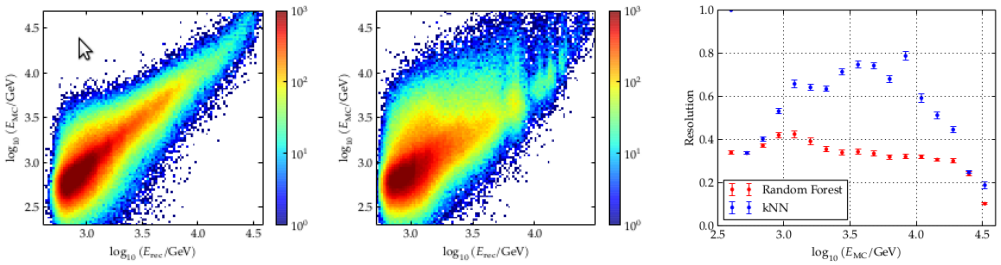


Figure 2: Dependance between E_{REC} and E_{MC} for the Random Forest regressor (left), the kNN regressor (middle)and the resulting estimation of the energy resolution (right).

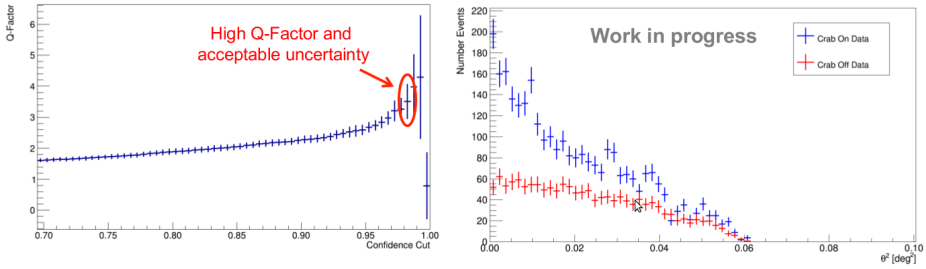


Figure 3: Performance of the random forest classificator (left) and application of the model to crab nebula data (right).

correlation to the label and a low correlation to each other. Using this parameter set a random forest classificator is trained. The resulting model is tested in a k-fold cross validation and a performance value, the so-called Q-Factor ($Q = \frac{E_{\mathrm{G}}}{\sqrt{E_{\mathrm{P}}}}$, E_{G}: Efficiency Gamma ; E_{P}: Efficiency Proton) is calculated for different cuts in confidence. Figure 3 shows the resulting q-factor. A confidence cut with high Q-Factor and acceptable uncertainty is choosen.
The trained model is than applied to a real data set of the crab nebula, the so called "standard candle" in high energy gamma ray astronomy. Figure 3 shows an excess in events from the desired source position in comparison to a OFF position.

5 Unfolding

The determination of the spectrum of the detected gamma flux from measured observables is an inverse problem. The software TRUEE [5] resolve this inverse problem. It applies a likelihood fit, with a tikhonov regularization using the 2nd derivative, on simulated events to determine the entries of the response matrix. TRUEE supports the

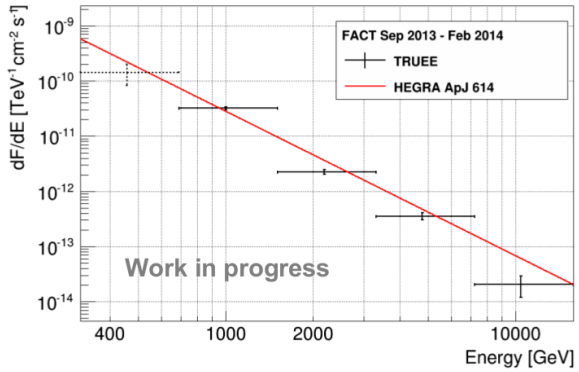


Figure 4: Unfolded energy spectrum of the crab nebula without systematic uncertainties. For comparison the flux of the crab nebula, observed by HEGRA [1] is displayed.

usage of up to three observables and has different modes for validation of the unfolding settings.
Figure 4 shows the unfolded spectrum of the separated data of the crab nebula. Due to uncertainties of the monte carlo simulation, no systematic uncertainties are included.

References

[1] F. Aharonian et al. The crab nebula and pulsar between 500 gev and 80 tev: Observations with the hegra stereoscopic air cerenkov telescopes. *The Astrophysical Journal*, 614(2):897, 2004.

[2] Anderhub et al. Design and operation of FACT - the first G-APD Cherenkov telescope. *Journal of Instrumentation*, 8, June 2013.

[3] Christian Bockermann and Hendrik Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012. Project SFB 876-C1.

[4] Mierswa et al. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.

[5] N. Milke et al. Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2012.

[6] Fabian Temme. FACT - Data Analysis: Analysis of Crab Nebula Data using PARFACT a newly Developed Analysis Software for the First G-APD Cherenkov Telescope. Diploma thesis, Technische Universität Dortmund, December 2012.

Signal-Background Separation Study of FACT Data

Julia Thaele
Experimentelle Physik 5
Technische Universität Dortmund
julia.thaele@tu-dortmund.de

An important aspect in astroparticle physics is the separation of signal events from background events. The First G-APD Cherenkov Telescope (FACT) detects air showers induced by gamma and hadronic particles coming from distant astrophysical sources. In order to separate the wanted gamma showers from the unwanted hadronic showers a Random Forest algorithm is trained with a set of Monte Carlo Simulations and is applied to real data recorded with FACT using the data mining environment RapidMiner. In this report the results of the training and testing of the built models are presented.

The so-called Imaging Air Cherenkov Telescopes (IACTs) are able to detect very high energy gamma-rays of galactic or extragalactic objects like supernovae or Active Galactic Nuclei (AGN). Due to the neutral electric charge gamma-rays are not influenced and deflected by intergalactic magnetic fields. Thus the direction they are coming from points directly to the astrophysical source. When very high-energetic gamma or hadronic particles are hitting the upper atmosphere layers of Earth, they induce an extensive air shower which emits a blueish light, the so-called Cherenkov light [7].
FACT is the first IACT which uses Geiger-mode Avalanche PhotoDiodes (G-APDs) instead of photomultipliers as photosensors to detect this light [4]. Due to a signal to background ratio of 1:1000 the separation of gamma showers from hadronic showers is very important to increase the sensitivity of the telescope and thus the effective observation time. The building and testing of the separation model is done with a Random Forest (RF) algorithm [5], which is implemented in the RapidMiner analytics platform [2]. In particular the RF of the implemented WEKA [8] package was used for this analysis. The models are trained and tested on gamma- and proton Monte Carlo (MC) simulations for FACT, which were further processed by the analysis software Modular Analysis and

Reconstruction Software (MARS) [3] and as well by the analysis software FACTTools (FT) developed within the SFB876. [1]. After data processing quality cuts were applied to each data set to filter out nonphysical events and to cut already away a large amount of background events. The RF was trained with parameters which describe the shower

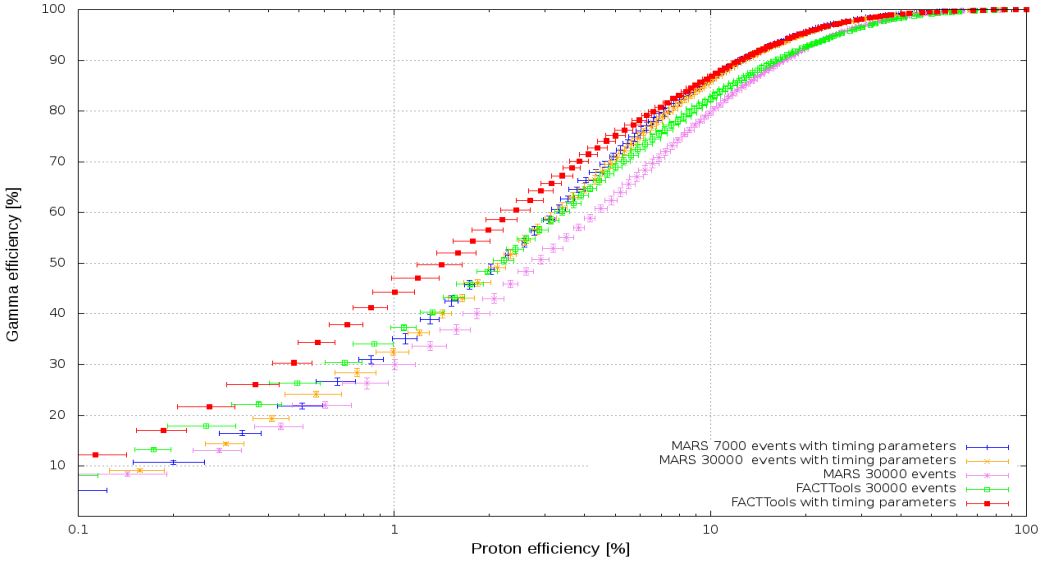


Figure 1: Displayed are the ROC curves of different trained Random Forest models on Monte Carlo simulation processed by the analysis software MARS and FT respectively. The x-axis is displayed in a logarithmic scale to show the interesting region in small proton efficiencies.

images and thus allow to distinguish between gamma showers and hadronic showers. The features for this analysis were selected with the MRMR algorithm [6]. For the RF 100 trees were built and five randomly chosen features taken out of a total amount of 25 parameters. To ensure and estimate the stability of the model a five fold cross validation was applied to the RF models. In this way statistical means and error values can be determined. Both software programs contain a certain intersecting-set of features for the RF training. The software FT provides also a certain amount of additional features. In Fig.1 the ROC curves of different models are shown. The violet and green datapoints show the results testing on MC data processed by MARS respectively by FT, trained with 30000 events for each class and the same features provided by both programs. By adding timing features to both training sets of the RF and additional specific features in FT, represented by the blue resp. red datapoints, an improvement of the ROC curves for both programs can be seen, as for a given proton efficiency the gamma efficiency is higher. The additional features in FT yield the best performance of all models. The challenge is to find a minimum confidence cut at which not too much gamma events are cut away while the purity of the dataset is still high. One possibility to decide which minimum confidence cuts offers the best results is to determine a so-called quality factor Q, which describes basically the gain of a simple statistical significance of a signal. Thus,

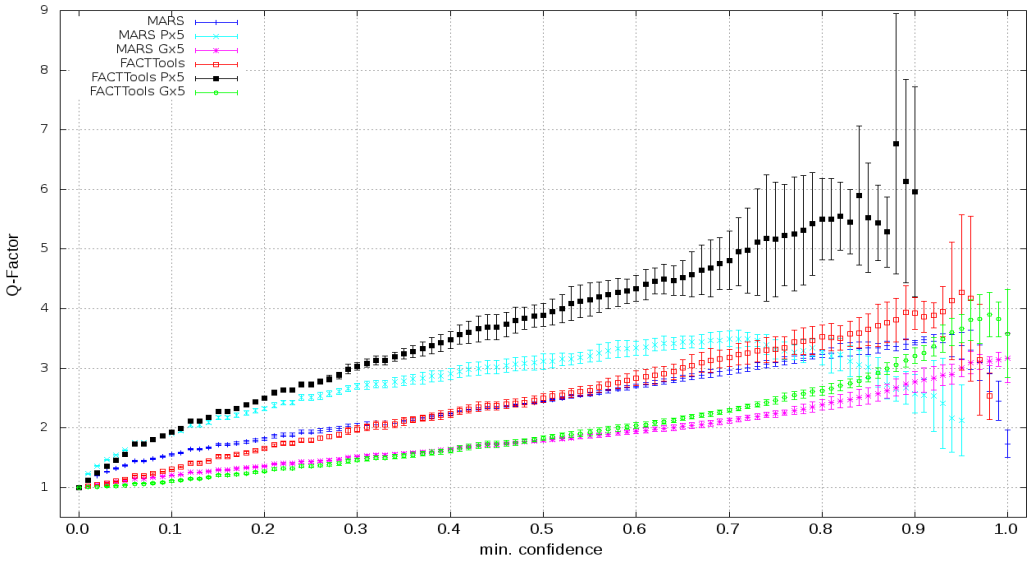
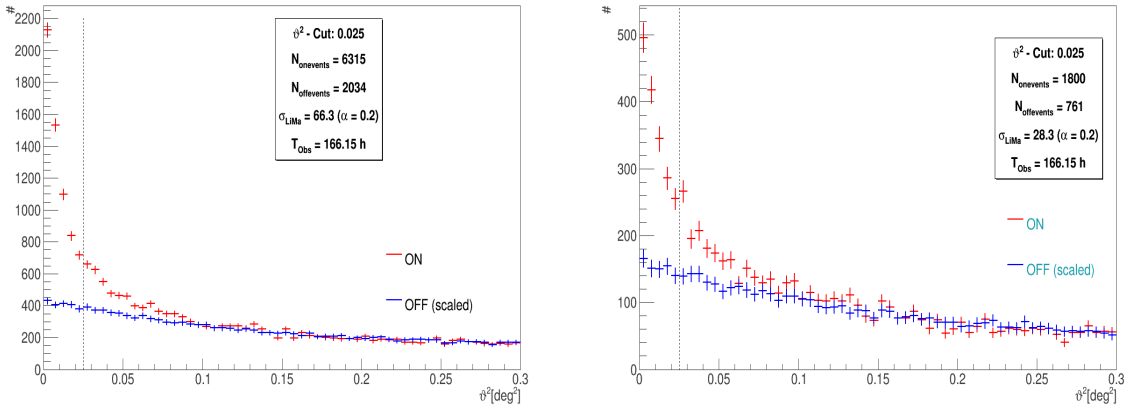


Figure 2: Displayed are the Q-Factors against the minimum confidence for RF models trained with different training ratios and with different analysis software programs. The blue, cyan and magenta datapoints represent the different training ratios of data sets processed by the analysis software MARS and red, black and green datapoints those processed by FT.

the minimum confidence with the maximum Q-Factor can be selected. It describes the ratio of the efficiency for gammas to the efficiency of hadronic showers and is

$$Q = \frac{E_G}{\sqrt{E_P}},$$

whereas E_G describes the gamma efficiency and E_P the proton efficiency. The statistics of the MC data is limited, as only 35000 proton events are left for training purposes. Thus, to investigate the influence of the training ratio on the performance, the absolute amount of events had to be scaled down. In Fig.2 the Q-Factors of the different models against the minimum confidence cut are shown. Here are the Q-Factors displayed for a traning ratio of 1:1 with 7000 Events for each class, five times more gamma events than proton events (35000 gamma events, 7000 proton events) and five times more proton events vice versa. In both analysis software programs a higher amount of protons show a slightly higher maximum of Q-Factors than in the other ratios. The maximum is also shifted towards smaller minimum confidence cuts the more proton events relatively to gamma events are contained in the training datasets. The Q-Factor values for FT are higher than for MARS, but have larger errorbars. Note that the performance of FT is not optimized. The resulting models were applied to real data of the Crab Nebula processed by MARS and FT. The minimum confidence cuts were choosen to gain high Q-Factors with acceptable errorbars. In Fig.3 the distributions of the distances of the reconstructed to the real source position are shown. In both cases a significant detection can be seen.



(a) Source Detection with MARS

(b) Source Detection with FT

Figure 3: Significant source detection of the Crab Nebula with MARS and FT.

References

[1] FACT Tools program
http://sfb876.tu-dortmund.de/FACT/.

[2] Rapidminer Homepage
https://rapidminer.com/products/studio/.

[3] T. Bretz and D.Dorner: MARS - CheObs ed. - A flexible Software Framework for future Cherenkov Telescopes. *WSPC Proceedings*, Nov 2009.

[4] Innovative camera records cosmic rays during full moon. *International Journal of High-Energy Physics*, Nov 2011.

[5] Leo Breiman. Random Forests. *Machine Learning*, 45:pp. 5–32, 2001.

[6] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Proceedings of the Computational Systems Bioinformatics*, pages 523–528, 2003.

[7] Claus Grupen. *Astroteilchenphysik: Das Universum im Licht der kosmischen Strahlung*. Vieweg, 2000.

[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, pages 10–18, 2009.

Correcting simulated data and a data mining approach to unfolding

Max Wornowizki
Statistics in Biosciences
TU Dortmund University
wornowiz@tu-dortmund.de

1 Solving inverse problems via random forests

In physics the distribution of an unobservable quantity X is often of interest. Typical examples for such quantities are energy or a certain angle. Since those cannot be measured directly, a reconstruction from other observables is necessary. This problem is called unfolding. In this paper a novel data mining based approach towards unfolding is given, treating the inverse problem as a multinominal classification task. We also propose unbiased estimators for the variance of our method allowing to construct confidence regions.

It is a common challenge in particle and astroparticle physics, that a distribution of interest $f(x)$ cannot be accessed directly and a second distribution $g(y)$ is measured instead. Due to smearing effects and a limited acceptance of the detector, $g(y)$ cannot be directly converted into $f(x)$. Instead both distributions are connected by the Fredholm integral of first kind:

$$g(y) = \int_a^b A(x, y) f(x) dx$$

where $A(x, y)$ represents the response function of the detector. This is commonly referred to as an inverse or ill-posed problem. Several alorithms for the solution of inverse problems exist, with regularised unfolding as implemented in RUN and TRUEE probably being the most widely known. These methods discretize the problem and estimate A, a matrix after discretisation, on Monte Carlo data. This estimate is then used to solve the problem

approximately in a second step. The downside of this approach is the estimation of only one matrix for the whole procedure. This approximation is fully valid for small and homogeneous detectors but might become problematic for large detectors utilising natural media, e.g. large scale neutrino telescopes. Particles of the same energy will cause significantly different event patterns depending on where they enter the detector. Thus, geometrical information on the track of the individual particles, will provide useful information, that can be utilised in order to improve the unfolding. Unfortunately the number of input variables is limited in many unfolding procedures because the number of observations required to get a reasonable fit grows exponentially in the number of input variables. Also unfolding for different values of a specific attribute like time can only be carried out by applying the procedures multiple times.

We propose a machine learning based approach to unfolding, which accounts for all of the drawbacks discussed above. Along the lines of the matrix methods, the data mining approach operates on the discretized data. Hence, instead of trying to reconstruct the direct value of the target variable, the probabilities for X to lie in certain intervals have to be estimated. From the data mining point of view this is a multinomial classification problem where each class corresponds to a certain bin. There are various algorithms for the classification. We propose a modified version of the Random Forest classifier given in [**?**]. The classifical Random Forest performs quite good in many real life problems and its extension led to a significant improvement of the resutls in this particular application. In addition, the variance of the resulting method can be estimated in a straightforward way allowing to construct confidence regions. In principle, the proposed technique can be applied also with other learners like Boosted Decision Trees or Support Vector Machines as well, which could be the subject of future work.

Our approach is the following: Using a Random Forest, the probability for the i-th event to lie in bin k is estimated by:

$$\hat{p}_{ik} = \frac{n_{i,k}}{n_{trees}} \ . \tag{1}$$

Here $n_{i,k}$ denotes the number of trees voting for the target variable to be in bin k for the i-th event while n_{trees} denotes the total number of trees in the forest. The classical Random Forest algorithm assigns the i-th event to the class with the highest probability and therefore looses the information on the other probabilities. However, we are actually not interested in the classification of an event to a unique class. We rather want to learn something about the distribution of the target variable. Therefore, all probabilities for the event are of great value in our context. Hence, we propose to use estimate the number of events with target variable lying in bin k within a time period of length T by the sum of all individual probabilities:

$$\hat{N_k} = \sum_{i=1}^{N} \hat{p}_{ik} \tag{2}$$

where N is the number of observations.

The variability of this estimator is caused by four sources:

- The unknown observations in the training data
- The random tree growth during training
- The random size of the test data
- The unknown observations in the test data

We take into account all of them in the derivation of a suitable variance estimator. In particuar, we asses the randomness in the training data by using several training data sets and averaging the results appropriately. To capture the effect of the random tree growth we compare pairs of trees. The random size of the test data is assumed to follow a Poisson distribution allowing to estimate the third variance source. We average the results of the test data assumed to stem from independent identical distributed random variables to solve the fourth point.
The results for this method when applied on Monte Carlo data seem quite promising and competative in comparison to the RUN and TRUEE. We hope to even outperform these methods significantly after some refinement on the current procedure, because classification algorithms can incorporate the information of many attributes much more easily than the methods currently used. One main point of possible improvement here is the regularisation of the method, i.e. a suitable processing of the probability vectors $\hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{ik})$ to remove implausible artifacts in this step of the algorithm.
More general details on this topic can be found in [3]. The estimator for the variance of the method as well as extensions and regularisations of the procedure will be discussed in the forthcomin paper by M. Boerner, M. Schmitz, T. Ruhe, T. Voigt and M. Wornowizki.

2 Demixing empirical distribution functions

In applications, the typical approach of testing whether given simulated data fits observed real data well is not completely satisfying. In case of rejection of a good agreement the way to improve the simulation remains unclear. We attempt to close this gap using a fast and intuitive algorithm based on the classical Kolmogorov-Smirnov test. The method is nonparametrical and therefore applicable without any assumptions on the data structure. It computes a correction function describing which values are over- respectively underrepresented in the current simulation in case of rejection. The main idea here is to consider a mixture of the correction with the simulation leading to new simulated data. The correction is fitted such that this corrected data not is not rejected by the test any more when compared to the real data. We show how this concept can be captured analytically, motivate resulting constraint optimisation problems and solve them. The correctness as well as linear run time of our algorithm are proved. More information on the approach are given in [2].
The method is applied to IceCube data [1] in order to identify insufficient simulated

variables and illustrate the problematic regions. As an example the following figure shows the situation for the MPE_TT1_HighNoise_Zd attribute:

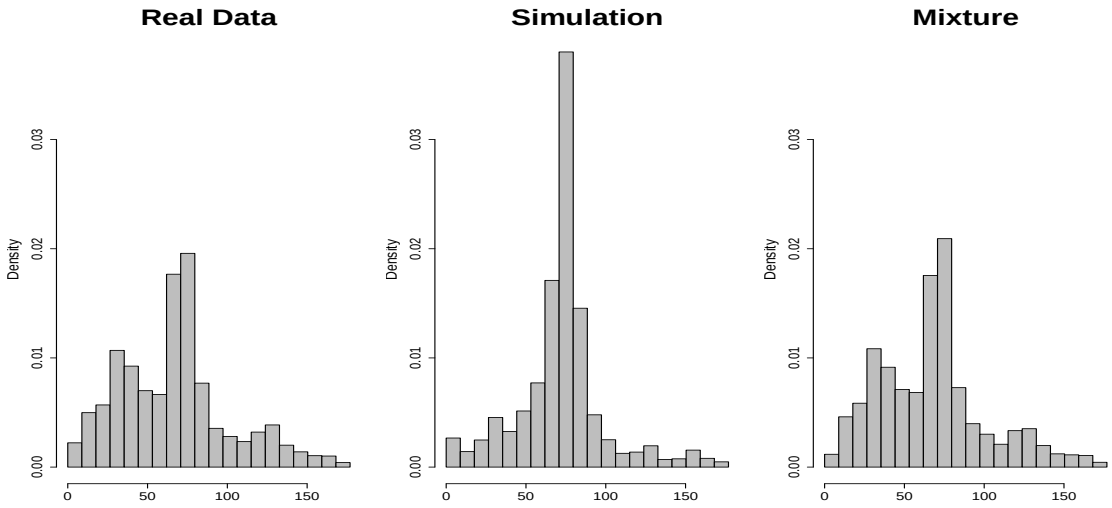


Figure 1: Histrograms of the observed data, the simulated data and the proposed mixture for the variable MPE_TT1_HighNoise_Zd

In comparison to the observed data the simulated ones are overrepresented in the middle between 50 and 100 and at the same time underrepresented for values below 50. The data of the mixture after correction resembles the simulation quite well.

References

[1] The AMANDA Collaboration: J. Ahrens et al. (2004)
Sensitivity of the IceCube detector to astrophysical sources of high energy moun neutrinos.
Astropart. Phys. 20

[2] A. Munteanu and M. Wornowizki (2014)
Demixing empirical distribution functions.
Technical Report
Collaborative Research Center SFB 876

[3] T. Ruhe, M. Schmitz, T. Voigt and M. Wornowizki (2013)
DSEA: A Data Mining Approach to Unfolding
33nd International cosmic ray conference, Rio de Janeiro 2013

Subproject C4
Regression approaches for large-scale high-dimensional data

Katja Ickstadt Christian Sohler

Random projections for Bayesian regression

Leo Geppert
Lehrstuhl Mathematische Statistik und biometrische Anwendungen
Fakultät Statistik, TU Dortmund
geppert@statistik.uni-dortmund.de

Bayesian linear regression is a computationally demanding task, especially in a Big Data setting. To reduce both running time and required memory, we propose using random projections and show that they offer a good approximation of the original linear regression task. The size of the sketched data set does not depend on the number of observations.

Bayesian linear regression

Regression techniques are widely used for statistical analyses. Bayesian linear regression can be used in cases where a classical linear regression is adequate, but prior information about the parameters might be available as well. A Bayesian linear regression model is based on the same model as a classical linear regression model, confer equation (1),

$$Y = X\beta + u, \tag{1}$$

where $X \in \mathbb{R}^{n \times d}$ is a matrix of observations, $Y \in \mathbb{R}^d$ is a vector with the random dependent variables and $u \in \mathbb{R}^n$ is a random vector with independent measurement error, $u \sim N(0, \sigma_u^2 I_n)$. In a Bayesian setting, $\beta \in \mathbb{R}^d$ follows a distribution. Prior information can be incorporated in the so-called prior distribution $p(\beta)$. The aim of the analysis is to obtain the posterior distribution $p(\beta|X)$. It is usually not possible to calculate the results analytically. There are different alternatives. Markov Chain Monte Carlo (MCMC) methods are the current standard. They are slow, but very reliable.

Random projections

In Bayesian analyses, the repeated evaluation of the likelihood is a bottleneck and makes MCMC-methods infeasible for Big Data. Different approximations have been proposed as a remedy. Our approach is to analyse a random projection of the original data set. After the random projection, the new sketched data set is of dimension $k \times d$ with $d < k \ll n$. This speeds up the subsequent analysis and reduces the required memory. The results are provably close to the original results, measured using the Wasserstein distance. The Wasserstein distance is especially useful for models with normally distributed likelihood or prior distributions as it is composed of a mean term and a variance-covariance term, which can be easily interpreted for normal distributions. Our theoretical results show that the distance between the two posterior distributions is an ε-fraction of these two parameters of the original distribution, where ε controls the approximation error.

For the empirical evaluations based on simulated data, we consider three random projection methods: the Rademacher Matrix (BCH) [5], the Subsampled Randomized Hadamard Transform (SRHT) [1] and the Clarkson Woodruff Sketch (CW) [2]. These methods differ in their running time and in the dimensions of the resulting sketch. CW is by far the fastest of the three random projection methods, followed by SRHT and BCH. The values of k only depend on d and ε, but importantly not on n, which makes the approach especially useful for Big Data settings. In our implementation, BCH and SRHT result in similar values of k. CW leads to larger values of k for small d, but smaller k for larger d. For more details on these methods and our implementation, please confer [3].

Empirical evaluations

The simulations were done using the statistical software R [4] and the R-package `rstan` [6]. Some results for $d = 50$ are presented here. We choose $\varepsilon = 0.1$ for all three random projection methods and additionally $\varepsilon = 0.2$ for CW. Table 1 contains the sums of absolute differences between the posterior mean values and the true mean. The posterior means of the CW-method in particular are close to the true values. For all models based sketched data sets as well as the model on the original data set, there is an increase of the sum of absolute deviations as the variance of the error term σ_u^2 increases and the goodness-of-fit of the model decreases. The sum of absolute deviations generally increases with growing number of observations, however, there are some small values for the analysis on the data sets obtained by the CW projection method and $n = 500\,000$.

The means and medians of the original posterior distributions are well-recovered by the analyses on the sketched data sets. There is no indication of systematic biases. In all posterior distributions based on sketched data sets, some additional variation is present compared to the posterior distribution on the respective original data set. The amount of

n	σ_u	original	BCH	SRHT	CW	CW, $\varepsilon = 0.2$
50 000	1	0.085	0.892	0.717	0.319	0.417
50 000	2	0.208	1.565	1.514	0.491	0.992
50 000	5	0.674	3.860	3.660	0.936	1.656
50 000	10	3.202		9.131	4.209	4.043
100 000	2		2.749	3.036	0.808	0.930
100 000	5		7.956	6.754	0.875	3.200
100 000	10				3.614	4.920
500 000	1				0.316	0.517
500 000	2				0.614	0.585
500 000	5				1.997	1.965

Table 1: Sum of absolute deviations of posterior mean values from true mean

the extra variation is highest for the CW-method and $\varepsilon = 0.2$. This is due to the higher value of ε, which controls the approximation error. For $\varepsilon = 0.1$, the amount additional variation is distinctly smaller and roughly the same for all three methods. However, the increase in the spread of the posterior distribution does not lead to differing decisions on the importance of variables for the model.

Table 2 contains the total running times for the analyses on the original data sets and the sketched data sets. For the original data sets, the table gives the running time for the MCMC-algorithm only. For the sketched data sets, it also includes the time required for obtaining the random projection. However, this takes between 30 seconds and a couple of minutes and thus does not constitute a major factor. The running time generally decreases as the standard deviation σ_u increases. It can also be seen that a doubling of the number of observations does not lead to increased running times for all of the random projection methods. For the CW-method, even $n = 500\,000$ does not go along with a clear increase in the running times compared to $n = 50\,000$ and $n = 100\,000$.

For $\varepsilon = 0.1$, the running time is comparable for all three projection methods, but is smallest for CW. Even though CW-based sketches generally result in a larger k, for $d = 50$, the CW-based sketch is smaller than the sketches based on the other two methods, with $k_{CW} = 16\,384$, $k_{BCH} = 20\,546$, and $k_{SRHT} = 20\,547$.

The results for CW and $\varepsilon = 0.2$ illustrate the trade-off between the goodness of the approximation and the required running time and memory. The number of observations is considerably small, $k_{CW,\varepsilon=0.2} = 4\,096$, resulting in a markedly smaller running time. However, while the location of the posterior distribution is well-recovered, the dispersion becomes larger.

n	projection method	$\sigma_u = 1$	$\sigma_u = 2$	$\sigma_u = 5$	$\sigma_u = 10$
50 000	original data set	101.34	87.05	55.98	48.52
50 000	BCH	26.37	24.77	14.47	
50 000	SRHT	28.89	31.45	20.49	15.66
50 000	CW	22.63	21.10	11.05	12.76
50 000	CW, $\varepsilon = 0.2$	3.09	3.12	2.39	2.58
100 000	BCH		22.40	22.72	
100 000	SRHT		26.40	17.41	
100 000	CW		20.69	15.78	15.37
100 000	CW, $\varepsilon = 0.2$		2.97	2.28	3.14
500 000	CW	21.91	23.31	13.36	
500 000	CW, $\varepsilon = 0.2$	2.99	3.08	2.33	

Table 2: Running times (in hours) for data sets with $d = 50$

References

[1] Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual bch codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

[2] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13*, pages 81–90, 2013.

[3] Leo N. Geppert, Katja Ickstadt, Alexander Munteanu, and Christian Sohler. Random projections for Bayesian regression. Technical report, Technische Universität Dortmund, 2014.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[5] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th annual IEEE Symposium on Foundations of Computer Science, FOCS 2006*, pages 143–152, 2006.

[6] Stan Development Team. Stan: A C++ library for probability and sampling, Version 2.3.0, 2013.

An improving streaming algorithm for the least squares regression problem

Alexander Munteanu
Efficient algorithms and complexity theory
Technische Universität Dortmund
alexander.munteanu@tu-dortmund.de

Our work addresses a new computational model for data streams: improving streaming algorithms. Improving streaming algorithms have an approximation ratio that tends to one as the number of items in the stream goes to infinity. In other words, the output of the algorithm is optimal in the limit. The memory used by the algorithm is restricted to polylog(n) size. The least squares regression problem is an important statistical problem that has been discussed extensively in the streaming literature. Under mild assumptions on the input stream, we can give an improving streaming algorithm for that problem. The algorithm is based on sketching techniques that yield a $(1+\varepsilon)$-approximation in the regular streaming setting.

Introduction

Streaming algorithms aim at solving problems in a setting where the input is given as a stream of items like numerical values, points in Euclidean space or edges of a graph and moreover it is not possible to store the entire input in the main memory. Usually a streaming algorithm is allowed only one pass over the data and its working memory is restricted to polylogarithmic size in the length of the stream [6]. For most non-trivial problems, it is not possible to get an exact solution with these restrictions. However, we can focus on the design of efficient approximation algorithms. The seemingly best we can hope for in this situation is a $(1\pm\varepsilon)$-approximation. Such approximation algorithms have been developed for many interesting and important problems. Known results in this area cover a broad variety of computational problems, including $(1\pm\varepsilon)$-approximation

algorithms for estimating the frequency moments of a stream of items [1], least squares regression, low-rank approximation [2] and clustering [3]. These have many applications in machine learning, classification, data mining and other fields of research.

From an information theoretic as well as statistical perspective it seems natural to say that the more data is used in a learning task, the more precise our result will be. In this context one might think of the law of large numbers or central limit theorems. However, these arguments require the observed data to follow some fixed distribution. In our attempt to develop improving streaming algorithms, we don't want to impose such strict assumptions but still have to put mild restrictions on the input stream. This is due to the fact that if some part of the result won't get enough data or only redundant data, i.e., there is no new information on that part, then we cannot hope to improve an error that we have already made.

Our model Our work deals with a new model for the design of streaming algorithms. The above discussion raises the question whether we can develop streaming algorithms, which have a guarantee on the error that approaches zero as the length n of the stream tends to infinity, i.e., that have an approximation ratio of $(1 \pm \varepsilon)$ for some $\varepsilon = o(1)$, while the memory is still bounded by polylog(n). As the space complexity of many problems in the streaming model is polynomial in $1/\varepsilon$ we might think of choosing $\varepsilon = \Theta(\frac{1}{\log n})$. Given an algorithm in the usual streaming model, we could just fix ε to such a value in advance. This means that we already have non-uniform approximation algorithms in the above sense. But this requires to know the length of the stream in advance. If otherwise, the length of the stream exceeds its pre-defined limit, the algorithm will fail to satisfy the desired approximation guarantee. Our intention is therefore to develop algorithms that are *uniform* in terms of n and can deal with potentially infinite input streams. We will call such algorithms *improving streaming algorithms* according to the following definition.

Definition 1. A problem P with objective function $V : S \to \mathbb{R}$ has an *improving streaming algorithm* if there exists a one-pass streaming algorithm that for an infinite input stream I and every $n \in \mathbb{N}$ maintains a solution $s^{(n)} \in S$ that with probability at least $1 - \delta$ satisfies

$$\frac{V(s^{(n)})}{V(s_{opt}^{(n)})} \xrightarrow{n \to \infty} 1$$

where $s_{opt}^{(n)}$ is the optimal resp. exact solution to the substream of length n that has been read. The space complexity of the algorithm is bounded by $\log^{O(1)} n$.

The least squares regression problem

The least squares regression problem is a statistical problem that has been discussed extensively in the streaming literature.

Problem 1 (*least squares regression*)**.** Let A be an $n \times d$ matrix and b be a column vector of size n. Find a solution $\tilde{x}$ such that

$$\|A\tilde{x} - b\|_2 \leq (1+\varepsilon)\,\|Ax^* - b\|_2\,,$$

where $x^* = \operatorname{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$ is an optimal solution.

Regression is a very important problem used in machine learning and statistics to study the dependency between variables. The most efficient algorithms for solving regression problems with little time and space are due to the early works of Sarlós [7] as well as Clarkson and Woodruff [2] and have been further optimized and generalized in the last years. Their approach is to apply space and time efficient versions of the well known Johnson-Lindenstrauss transform [5] as a dimensionality reduction technique to reduce the space and time bounds of their algorithms.

Let A and b the input matrix and target vector obtained by processing the first n items from the input stream. One way to solve this problem for a fixed precision using random sign matrices for sketching is as follows. Sketch both, the matrix A and the vector b with an appropriately rescaled sign matrix S and solve the regression problem in the sketch space. That is, let $\tilde{x}$ be the optimal solution to $\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$. It has been shown in [2] that it is sufficient to have $\Theta(\frac{d}{\varepsilon}\log(\frac{1}{\delta}))$ as the target dimension of the sketching matrix, such that with probability $1-\delta$ we have $\|A\tilde{x} - b\|_2 \leq (1+\varepsilon)\min_{x \in \mathbb{R}^d} \|Ax - b\|_2$.

Now we want to compute a sketch of the matrix A and the vector b such that the approximate solution will have a $(1+\varepsilon)$ relative error, with ε decreasing to zero as the size of the input goes to infinity. Algorithm 1 splits the input matrix into blocks of $n_i = 2^i n_0$ rows at step i. Then it computes a sketch of size m_i for the block i using rescaled sign matrices for sketching. The resulting sketch will simply be the concatenation of the single sketches at each step.

Algorithm 1: Improving sketching algorithm for regression

Input: Matrix $A \in \mathbb{R}^{n \times d}$ given row-by-row
Output: A sketch SA of A
$i \leftarrow 0$
while *not* End of Stream **do**
    Sketch the next $n_i = 2^i n_0$ rows of A with a sketching matrix S_i with m_i rows
    $i \leftarrow i + 1$
return concatenation of the sketches at each step

The procedure described by the algorithm is equivalent to multiplying the input matrix A and the vector b by the block diagonal matrix $S = \operatorname{diag}(S_1, \ldots, S_l)$ where S_i is the matrix used at step i for sketching the input block. In the following, we will choose $\varepsilon_i = \frac{1}{\log n_i} = \frac{\varepsilon_0}{i}$ and $\delta_i = \frac{\delta}{ci^2}$, where c is a constant chosen to be large enough, such that

$\sum \delta_i \leq \delta$ holds. Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be the input to the regression problem after the first n rows have been read from the stream. Let $\tilde{x} = \operatorname{argmin}_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$ and $x^* = \operatorname{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$ be the optimal solution to the sketched problem and to the original problem at that point. We have the following theorem.

Theorem 2. *Assume that the smallest singular value σ_d of A satisfies $\sigma_d^2 \geq f(n)$ for a positive monotonous function f with $f(n) \to \infty$. If the blocks S_i of the sketching matrix S consist of $m_i \geq \frac{Cd}{\varepsilon_i} \log(\frac{1}{\delta_i})$ rows for some absolute constant C, then with probability at least $1 - \delta$ it holds that $\|A\tilde{x} - b\|_2 \leq (1 + \varepsilon) \|Ax^* - b\|_2$, where $\varepsilon = O(\frac{1}{\log f(n)})$.*

Conclusion

Our new model turns out to be quite restrictive, which makes it even more interesting. There are problems for which improving algorithms cannot exist, while they can be approximated within $(1 \pm \varepsilon)$ in the usual streaming model. Even for our positive results we have to put mild restrictions on the nature of the data. These are provably necessary. However, note that the assumptions are far less restrictive than the usual assumptions that underly regression analysis and other learning tasks as we have briefly discussed in the introduction. The full details of our work are available in [4].

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[2] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In Michael Mitzenmacher, editor, *STOC*, pages 205–214. ACM, 2009.

[3] Sariel Har-Peled. Clustering motion. *Discrete & Computational Geometry*, 31(4):545–565, 2004.

[4] Marc Heinrich, Alexander Munteanu, and Christian Sohler. Asymptotically exact streaming algorithms. *CoRR*, abs/1408.1847, 2014.

[5] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Conference in Modern Analysis and Probability*, pages 189–206, 1984.

[6] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

[7] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152. IEEE Computer Society, 2006.