technische universität
dortmund

SFB 876

Technical report for Collaborative Research Center SFB 876

Providing Information by Resource-Constrained Data Analysis

Oktober 2013

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

Contents

1 Subproject A1 **1**

1.1 Nico Piatkowski 2

1.1 Jochen Streicher 6

2 Subproject A2 **14**

2.1 Amer Krivosija 12

2.2 Chris Schwiegelshohn 16

3 Subproject A3 **21**

3.1 Helena Kotthaus 22

3.2 Michel Lang 26

3.3 Eugen Rempel 32

4 Subproject A4 **37**

4.1 Christoph Borchert 38

4.2 Markus Buschhoff 40

4.3 Björn Dusza 44

4.4 Markus Putzke 48

5 Subproject A5 **53**

5.1 Patrick Krümpelmann 54

5.2 Marcel Preuß 58

5.3 Cornelia Tadros 62

6 Subproject B1 **67**

6.1 Marianna D'Addario 68

6.2 Johannes Köster 72

6.3 Dominik Kopczynski 76

7 Subproject B2 **81**

7.1 Pascal Libuschewski 82

7.2 Olaf Neugebauer 86

7.3 Dominic Siedhoff 90

8 Subproject B3 **95**

8.1 Hendrik Blom 96

8.2 Benedikt Konrad 100

8.3 Marco Stolpe 104

9 Subproject B4 **109**

9.1 Lars Habel 110

9.2 Christoph Ide 114

9.3 Brian Niehöfer 118

10 Subproject C1 **123**

10.1 Kathrin Fielitz 124

10.2 Melanie Schwermer 128

11 Subproject C3 **133**

11.1 Fabian Clevermann 134

11.2 Katharina Frantzen 138

11.3 Jan-Hendrik Köhne 142

11.4 Ann-Kristin Overkemping 146

11.5 Florian Scheriau 150

11.6 Martin Schmitz 154

11.7 Fabian Temme 158

11.8 Julia Thaele 162

11.9 Anita Monika Thieler 166

11.10 Tobias Voigt 170

11.11 Max Wornowizki 174

12 Subproject C4 **179**

12.1 Leo Geppert 180

12.2 Alexander Munteanu 184

SFB 876 Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Olaf Spinczyk

The Integer Approximation of Undirected Graphical Models

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

Machine learning on resource constrained ubiquitous devices suffers from high energy consumption and slow execution time. In this paper an integer approximation to the class of undirected graphical models with discrete state spaces is proposed. In numerical evaluation on synthetic and real world data, the performance of the model is investigated. In addition, the runtime on resource constrained is regarded. The overall speedup of the new algorithms is at least $2\times$ while the overall loss in accuracy is low.

Running machine learning methods in resource constrained computational environments endows challenges in terms of the execution time and the energy consumption at the same time. Fortunately, optimizations which reduce the number of cycles in which the CPU is busy also reduce the energy consumption. When reviewing the specifications of processing units, one finds that integer arithmetic is usually cheaper in terms of instruction latency, i.e. it needs a smaller number of clock cycles until the result of an arithmetic instruction is ready. Table 1 shows the latencies of arithmetic instructions measured in terms of clock cycles for CPUs with Sandy Bridge and ARM11 architecture and for GPUs with Kepler architecture. Note that transcendental functions are composed out of multiple instructions and therefore may take substantially more cycles than the ones reported in Table 1. This motivates reducing the number of cycles in which code is executed when designing a new, resource-aware learning algorithm.

The joint prediction of many unknowns based on multiple observed inputs is a ubiquitous subtask of real world problems in various domains, including computational biology, computer vision, and natural language processing. Probabilistic graphical models are well suited for such tasks, but they suffer from the high complexity of probabilistic inference. Many approximate approaches to probabilistic inference were proposed in the last decade,

	Sandy Bridge		ARM11		Kepler	
	FP	INT	FP	INT_{32}	FP	INT_{32}
Addition	3	1	8	1	3/7	3
Multiplication	5	3	8/9	4-5	3/7	14
Division	14/22	13-25	19/33	-	7/-	-
Bit shift	-	3	-	2	-	7
Square root	14/22	-	19/33	-	14/-	-

Table 1: Instruction latencies (in clock cycles) of floating point (FP) and integer (INT) scalar arithmetic operations for three processing architectures [3, 1, 4]. x/y means that latency is x for 32 Bit and y for 64 Bit operands, a single value indicates that both latencies are the same or, in case of ARM11 and Kepler, that 64 Bit integer arithmetic is not supported. For Kepler, the values are based on the operation throughput. Cycles of Sandy Bridge integer division and ARM11 integer multiplication depend on the lengths of their operands.

but nearly all of them try to reduce the asymptotic complexity. In contrast, here, the goal is to reduce runtime and save energy through saving clock cycles, while keeping a good performance. Asymptotically, the new approach has the same complexity as the vanilla BP, but it uses cheaper operations.

Estimation in discrete parameter models was recently investigated by Chaorat and Seri [2]. They discuss consistency, asymptotic distribution theory, information inequalities and their relations with efficiency and super-efficiency for a general class of m-estimators. Unfortunately, the authors do not consider the case when the true estimator is not included in the search space and therefore, their analysis cannot be used to estimate the error in a situation when the optimizer has to be approximated. Bayesian network classifiers with reduced precision parameters were presented by Tschiatschek et al. [5]. The authors evaluate empirically the classification performance when reducing the precision of Bayesian networks probability parameters. After learning the parameters as usual in $\mathbb{R}$ (represented as 64 bit double precision floating point numbers), they varied the bit-width of mantissa and exponent, and reported the prediction accuracy in terms of the normalized number of correctly classified test instances. They found that after learning, the parameters may be multiplied by a sufficiently large integer constant (10^9) to convert the probabilities into integer numbers. However, Tschiatschek et al. missed an important point, namely that real valued probability parameters are necessary only for Bayesian networks. For undirected graphical models, this is not the case. As a result, the general framework of undirected graphical models [6] may be mapped to the integer domain. This allows the learning of integer parameters without the need for any floating point computation

and opens up the opportunity of running machine learning tasks on resource-constrained devices. To be more precise, based only on integers, it is possible to compute approximations to the marginal probabilities, the maximum-a-posteriori (MAP) assignment of the model and the maximum likelihood estimate either via an approximate closed form solution or an integer variant of stochastic gradient descent.

In this paper, new algorithms for integer models are derived. It turns out that the integer approximations do deliver a reasonable quality and are around twice as fast as their floating point counterparts. To the best of the authors knowledge, there is nothing like an integer undirected model so far.

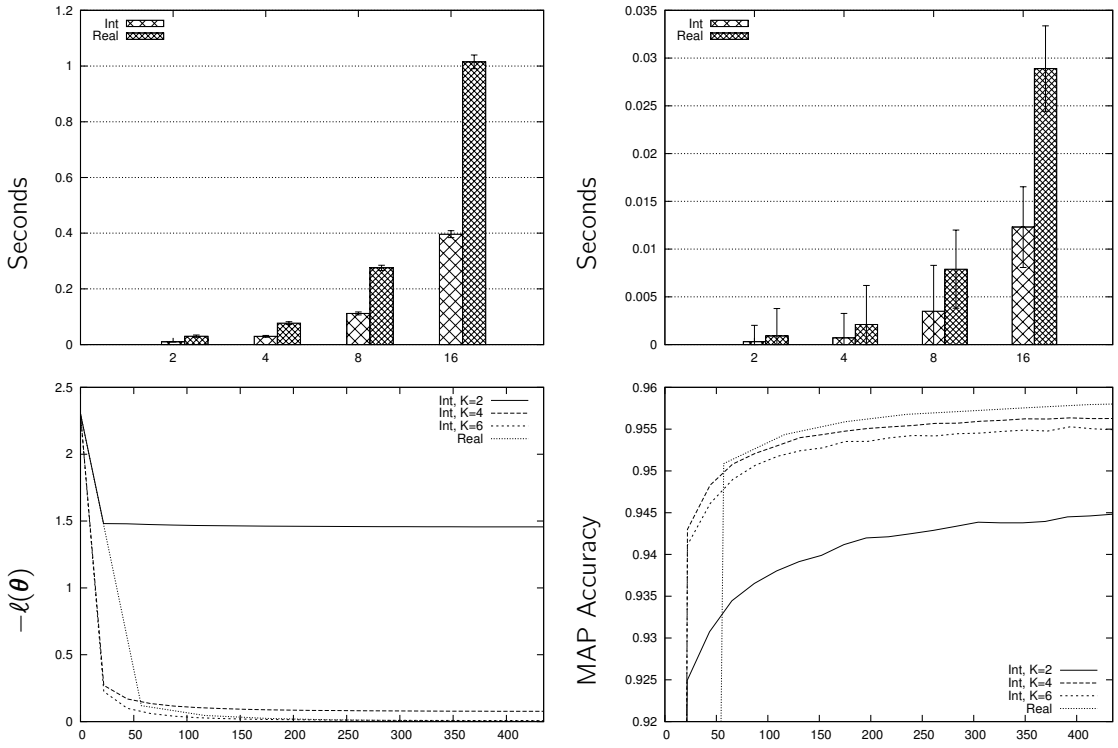


Figure 1: Top: runtime comparison of integer and floating point MRF on two architectures for a varying number of states. Left: Raspberry PI @ 700MHz (ARM11). Right: Intel Core i7-2600K @ 3.4GHz (Sandy Bridge). Bottom: progress of stochastic gradient training in terms of training error and test accuracy of the CRFs over running seconds.

Now, I present some preliminary results that are carried out on an Intel Core i7-2600K 3.4GHz (Sandy Bridge architecture) with 16GB 1333MHz DDR3 main memory. For comparison, some results on synthetic data show the performance of Integer MRF on the Raspberry Pi architecture, that can be considered resource constrained. The implementations of the methods are equally efficient, e.g. the message computation (and therefore the probability computation) executes exactly the same code for all methods,

except for the arithmetic instructions.

The motivation for the integer model was to save resources in terms of clock cycles. I can now demonstrate that the impact of this reduction is larger, if the underlying architecture is weaker, i.e. has slower floating point arithmetic. The two bar charts in Figure 1 show a runtime comparison of the integer MRF on two different CPU architectures. One is Sandy Bridge and the second one is a Raspberry Pi device with ARM11 architecture. As expected, the integer model actually speeds up the execution on the Pi device more than on the other architecture, i.e. the Pi gains a speedup of 2.56× and Sandy Bridge a speedup of 2.34×. In terms of standard deviation, the ARM11 architecture is more stable than the Sandy Bridge, which might be a result of a more sophisticated out-of-order instruction execution in the latter architecture.

In the second evaluation, stochastic gradient training of discriminative models is investigated on the well known CoNLL-2000 phrase-chunking data. An integer linear-chain CRF is constructed and trained by a stochastic gradient descent algorithm. In case of the integer CRF, the parameter updates are computed by means of the scaled integer gradient. Both algorithms perform 20 passes over the training data, each pass looping through the training instances in random order. This was repeated 50 times in order to compute an estimate of the expected quality of the randomized training procedure. The parameter update for the floating point CRF is computed with stepsize $\eta = 10^{-1}$. The ratio of quality per runtime is presented in Figure 1, where the negative log-likelihood is averaged over all training instances and the accuracy is computed w.r.t. the chunk tags. One can see, that for increasing size of the parameter space (K), the models overall performance increases.

[1] ARM Ltd. *ARM1176JZF-S Technical Reference Manual*. 2009. Rev. r0p7.

[2] Christine Choirat and Raffaello Seri. Estimation in discrete parameter models. *Statistical Science*, 27(2):278–293, 2012.

[3] Intel Corp. *Intel 64 and IA-32 Architectures Optimization Reference Manual*. April 2012. Order Number 248966-026.

[4] NVIDIA Corp. *CUDA Toolkit Documentation v5.0, CUDA C Programming Guide*. 2012.

[5] Sebastian Tschiatschek, Peter Reinprecht, Manfred Mücke, and Franz Pernkopf. Bayesian network classifiers with reduced precision parameters. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2012.

[6] 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, 2007.

An Open Smartphone Utilization and Mobility Data Set

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

Prior to our summer school in 2012, we started the distributed collection of utilization and mobility data from Android-based smartphones with the help of several CRC members. While mainly targeted for the smartphone data mining challenge during the summer school, the dataset has many more uses and is now publicly released. This report describes its contents, the modification prior to its publication as well as its use beyond the data challenge.

1 The Data

Using our *MobiDAC* infrastructure, data was collected from various hardware and software subsystems of the participating phones, namely communication (Wi-Fi, Bluetooth and mobile), sensors, power supply, the Linux kernel and Android's application framework. When the device was awake, most of the data was collected high-frequently (temporal resolution of 2s). Other data was received as a callback event or recorded less frequently due its amount (e.g., running processes) or battery-draining hardware (Wi-Fi scans or sensors). To reduce the amount of data that had to be transmitted from the device, we only recorded changes to data values. The device was explicitly awoken every 60 seconds and performed a full acquisition of all data values with the respective intervals. The collected data may be categorized into high-level user context, external sensing, and system internals.

High-Level User Context is utilization data that contains direct hints to the participant's current activity and context. This includes the state of the display (on/off,

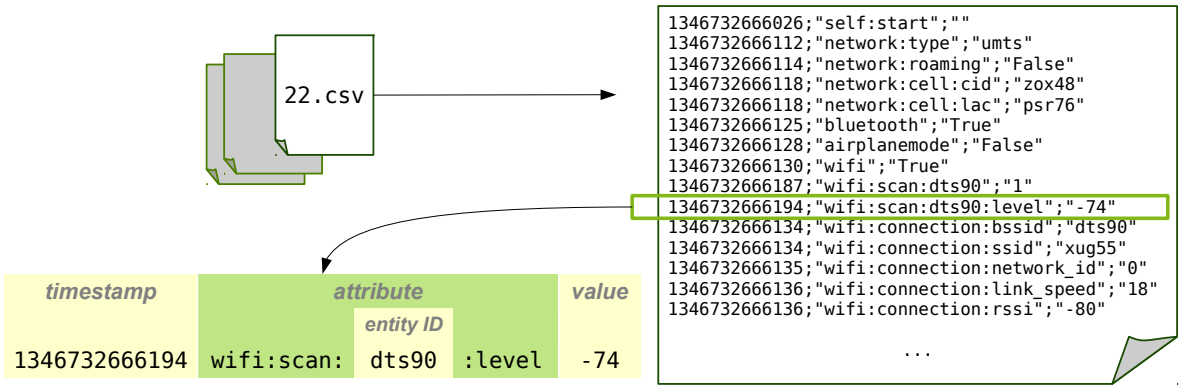


Figure 1: Structure of the public dataset.

brightness) and the phone (idle, ringing, or off the hook). Also the currently running packages belong to this category. Settings can also indirectly tell about the participant's context. For example, turning the phone to silent mode, when it was set to play a ring tone before, is a hint that the situation changed to one that prohibits phone noise, like a meeting or a cinema. Besides audio settings, also the communication settings, whether Bluetooth or Wi-Fi is enabled, or whether the device is in airplane mode, belong to this category.

Sensing data was obtained from various physical sensors as well as positioning and communication hardware. The physical sensors measured acceleration, magnetic field strength, orientation and light intensity. Every two minutes, they were sampled for two seconds with the highest possible frequency. When the participant allowed it, also position data was obtained via network and GPS hardware. Also, communication devices were used to sense the presence or even the signal strength of (potential) peers.

System Internal data mainly describes the overall usage of the system's resources like the CPU, the battery, the main memory and the network interfaces. The use of Android *wakelocks* also belongs to this category. Since wakelocks are used to prevent the device from suspending, (application) bugs regarding their handling can severely increase energy consumption.

For the public version of the data set, the following privacy-preserving measures were performed:

- Everything that uniquely identifies a participant is *globally consistently* replaced with a random value. This is also true for all identifiers from interaction with other entities (e.g., MAC-addresses and SSIDs) as well as for the names of installed and running application packages and processes.

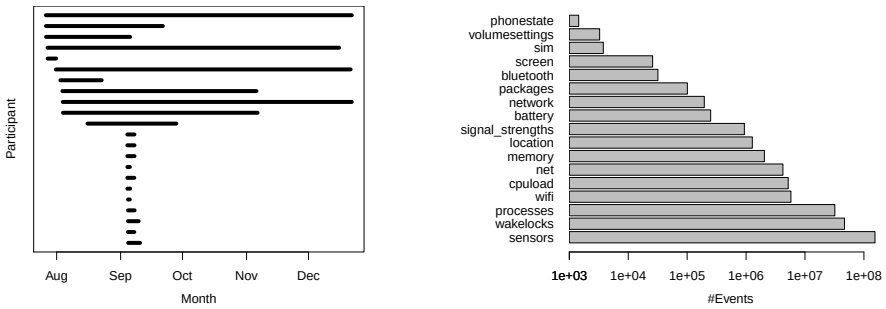


Figure 2: Left: Period of participation for every participant. Right: Total number of events for every event type (changes only, log scale).

- Mobile network cell information was replaced by *locally consistent* random values for each participant. This means that the mapping of cell identification (CID) and location area code (LAC) is different for every participant.
- Direct positions from GPS or network, phone numbers and the country for the SIM card were *completely removed* from the dataset.

Additionally, some cleansing was performed, namely removal (attributes that did not contain any information), reconstruction, as well as fusion of attributes (different names for mobile network devices) and deobfuscation.

For every device, the dataset contains a stream of events. Every event is composed of a timestamp, an attribute name, and the new value for this attribute. Figure 1 shows an excerpt of such a stream. For entity types with multiple instances, like Wi-Fi access points, attribute names contain a unique entity identifier for this resource (e.g., the BSSID for Wi-Fi). The appearance and disappearance of such entities is denoted with "1" or "0" respectively. For example, at time 1346732666194, the WiFi access point "dts90" was discovered. The complete dataset consists of 280 million of these events. Figure 2 illustrates their distribution regarding event type and participant.

2 Utility, Limititations, and Future Work

Besides the uses mentioned in [4], the dataset is currently used within project A1 together with MobiSIM [1] to examine and improve data transmission strategies for *Collective Apps* with energy-delay tradeoffs, like SALSA [2], as it contains long-term quality parameters of surrounding communication networks.

While the data is certainly useful, it also has limitations concerning applications that would be interesting for project A1. There are mainly two reasons:

Lack of detail Network device transmission statistics where sampled every 60 seconds. While this is sufficient to analyze utilization patterns, e.g., via item set mining, a higher resolution or even accurate traces of network packets would have been a valuable long-term input for studying the impact of packet deferral on everyday data communication. During data collection, a Wi-Fi scan was triggered every 5 minutes, while SALSA was originally evaluated with a 20s scan interval. Also, the dataset contains no information about the transmission speed of mobile networks, which would have required frequent active transmission and measurement.

Battery data While some of the data mentioned can be obtained without noticeably decreasing battery life, this is not the case if external sensing hardware is *actively* used. Since the devices were regularly woken up for that purpose, the battery levels in the dataset differ from normal utilization. Nevertheless, energy-intensive utilization, e.g., involving display or mobile data transmission, materialized in faster battery drain, which also supported by the linear model described in [4].

Thus, the dataset is a means for exploration, but not necessarily for final evaluation. Future use of this dataset will therefore be centered around utilization pattern discovery, e.g., via item set mining.

Trading off battery life against the range, detail and resolution of collected data certainly demands flexible data collection infrastructures. Although MobiDAC already provides the necessary flexibility and, the data collection was based on informed consent. This prohibited utilizing the flexibility to an extent more than just correcting small errors during the data collection.

With these experiences in mind, an architecture for flexible and privacy-preserving data collection was proposed that would alleviate the need for repeated written agreements. [3]

References

[1] Markus Buschhoff, Jochen Streicher, Björn Dusza, Christian Wietfeld, and Olaf Spinczyk. Mobisim: A simulation library for resource prediction of smartphones and wireless sensor networks. In *Proceedings of the 46th Annual Simulation Symposium*, ANSS '13, Society for Computer Simulation International, 2013. Society for Computer Simulation International.

[2] Moo-Ryong Ra, Jeongyeup Paek, Abhishek B Sharma, Ramesh Govindan, Martin H Krieger, and Michael J Neely. Energy-delay tradeoffs in smartphone applications. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 255–270. ACM, 2010.

[3] Jochen Streicher, Orwa Nassour, and Olaf Spinczyk. System support for privacy-preserving and energy-efficient data gathering in the global smartphone network – opportunities and challenges. In *Proceedings of the 3rd International Conference on Pervasive and Embedded Computing and Communication Systems (PECCS '13)*, pages 80–85. SciTePress, February 2013.

[4] Jochen Streicher, Nico Piatkowski, Katharina Morik, and Olaf Spinczyk. Open smartphone data for mobility and utilization analysis in ubiquitous environments. In *Mining Ubiquitous and Social Environments (MUSE) within ECML PKDD*, 2013.

Subproject A2
Algorithmic aspects of learning methods in embedded systems

Christian Sohler Jan Vahrenhold

Coresets in the Streaming Learning Processes

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

The streaming scenario is seen often in real time problems and the coresets are one possible approach to keep the $(1+\varepsilon)$-approximation of some set descriptor in the streaming setting. We familiarize with the current results for k-median und k-means problem, as well as with tools of learning theory for the further research.

Introduction

The question of handling large sets of gathered data appears everyday in many aspects of the human life, e.g. finances, security, marketing, politics, telecommunications, etc. However, large data amounts imply difficulties to store and process them and it is often not possible to keep the whole set in memory.

Input data sets are asked to be read only once, as information (that we call *points*) arrive. We assume that points arrive in a queue. Such a scenario is known as a streaming. As it would not be convenient to store all these points, a smaller set of points is used as a sketch of the whole original set instead. These sets are known as coresets.

There are various descriptors of the extent of a set of n points P from the d-dimensional domain for given distance function. They are called extent measures by Agarwal, Har-Peled and Varadarajan in [1]. The points in the set P can be positively weighted or unweighted, i.e. with weights equal to 1. We are particularly interested in the k-median (where we sum the distances to the clustering centers) and k-means (where distances are squared and summed) clustering problems.

For these extent measures and the parameter $\varepsilon > 0$, the notion of (k, ε)-coreset is defined as a (weighted) subset S of P that ε-approximates the cost of distances to any set of k points from P, i.e. it holds that

$$(1 - \varepsilon)\nu_C(P) \leq \nu_C(S) \leq (1 + \varepsilon)\nu_C(P)$$

where $\nu_C(P)$ is the (weighted) sum of (squared) distances of the points from P to the points from C that are the centers for the clustering problem.

The papers of Arora, Raghavan and Rao [4] and Kolliopoulos and Rao [11] produce $(1 + \varepsilon)$-approximation algorithms for k-median problem in euclidean metrics. The paper of Har-Peled and Mazumdar [9] provide $(1 + \varepsilon)$-approximation algorithms in streaming setting using merge and reduce technique. Their idea was modified by Chen in [6] by random sampling.

Further achievements are the results of Feldman and Langberg [7], who introduced two other types of coresets beside existing, newly called strong coresets. The weak coreset is the set D, such that $(1 + \varepsilon)$-approximation of the optimal solution (for chosen clustering problem) on D yields a $(1 + \varepsilon)$-approximation for the optimal solution on the full set P. We can say that through weak coresets we have a reduction of the original problem on P to the problem on smaller set D. The streaming coresets are the weak ones being updated during one pass through the data set and using only limited space.

The result we start from in our the further research was given in the paper by Feldman, Schmidt and Sohler [8], where using principal component analysis (PCA) were found constant-sized coresets for k-means problem. Earlier coresets had size at least linearly dependent on dimension d, that for d near n made them not practical. This result unfortunately does not hold for k-median problem, what will be aim of our further research.

Other tools

To keep the clustering coresets good the `k-means++` tool from the work of Arthur and Vassilvitskii [5] could be used. Their idea of adaptive sampling was further developed by Aggarwal, Deshpande and Kannan in [3]. Again, these tools were developed for k-means problem only.

Ideas from other set descriptors approaches were also interesting to us, as one from Agarwal and Yu [2]. They tend to keep so called ε-kernels as a coreset variant in the streaming settings in the euclidean plane. Their algorithm is space-optimal, that is interesting for us from the point of sizes of the coresets.

The work with ε-approximations was used in already cited [7], that leans to the theory of Vapnik-Chervonenkis (VC) dimensions and ε-nets, where as introductory texts the articles by Vapnik and Chervonenkis [13] and Haussler [10] were used. The improved bounds on

the sample complexity of learning were given by Li, Long und Srinivasan in [12]. It is to be seen if and how the ideas from [7] could be connected with the work of [3, 5].

Our contribution

This report is an overview of the literature that was read, in order to gather enough tools for further research. Therefore our contribution currently cannot be presented.

References

[1] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, July 2004.

[2] P. K. Agarwal and H. Yu. A space-optimal data-stream algorithm for coresets in the plane. In *Symposium on Computational Geometry*, pages 1–10, 2007.

[3] A. Aggarwal, A. Deshpande, and R. Kannan. Adaptive sampling for *k*-means clustering. In *I. Dinur et al. (Eds.): Approximation and Randomization and Combinatorial Optimization. Algorithms and Techniques Lecture Notes in Computer Science*, volume 5687, pages 15–28, 2009.

[4] S. Arora, P. Raghavan, and S. Rao. Approximation schemes for euclidean k-medians and related problems. *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, STOC 1998:106–113, 1998.

[5] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[6] K. Chen. On coresets fo *k*-median and *k*-means clustering in metric and euclidean spaces and their applications. *SIAM Journal of Computing*, 39:923–947, 2009.

[7] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 569–578, 2011.

[8] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the 24th SODA*, pages 1434–1453, 2013.

[9] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 291–300, 2004.

[10] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[11] S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. *SIAM Journal on Computing*, 37:757–782, 2007.

[12] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.

[13] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

Linear Classification in the Streaming Model

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

Learning from data streams is a well researched task both in theory and practice. As remarked by Clarkson, Hazan and Woodruff [6], many classification problems cannot be very well solved in a streaming setting. For previous model assumptions, there exist simple, yet highly artificial lower bounds prohibiting space efficient one-pass algorithms. At the same time, several classification algorithms are often successfully used in practice. To overcome this gap, we give a model relaxing the constraints that previously made classification impossible from a theoretical point of view and under these model assumptions provide the first $(1+\epsilon)$-approximate algorithms for training logistic regression and perceptron classifiers in data streams.

1 Introduction

Given n points in d-dimensional Euclidean space with labels (classes) $y_i = \{-1, 1\}$ (or 0 and 1, depending on model and mathematical convenience), a binary linear classifier is a hyperplane w separating the points into two sets according to the labels. For each candidate hyperplane, we measure the regret as the summation of loss terms of every point with regards to the hyperplane. Designing an appropriate function for the loss incurred by a point is not an easy task, as it entails such requirements as interpretability, mathematical tractability and applicability. We focus on the perceptron loss function, where correctly classified points, that is points on the "correct" side of the hyperplane incur a loss of 0, and misclassified points incur a loss of 1, or formally for a point $x_i \in R^d$ the loss is defined as $L(x_i, w) = \begin{cases} 0 & \text{if } y_i \cdot \langle x_i, w \rangle > 0 \\ 1 & \text{if } y_i \cdot \langle x_i, w \rangle \leq 0 \end{cases}$.

The goal therefore is to find a hyperplane with a minimum number of misclassified points, or formally:

Problem 1 (Minimum Perceptron Classification) *Let X be a d-dimensional set of points, where each point x_i is labeled by $y_i \in \{1, -1\}$. Then the Minimum Perceptron Classification problem is to find a hyperplane w_{opt} minimizing $t := L(w_{opt}) = \sum_{x \in X} L(x, w_{opt})$.*

Note that while this problem has the identical optimal solution as maximizing the number of correctly classified points, the approximate version is much more difficult in our case. There exist multiple algorithms that solve this problem if there exists a separating hyperplane with no misclassifications (SVMs, Perceptron, general linear programming algorithms). By evaluating all possible hyperplanes induced by n points in d-dimensional space, the general problem can be solved optimally in time $O(n^{d+1})$. No polynomial time (with respect to d) constant factor approximation algorithm is possible unless $NP \subset DTIME(d^{\log \log d})$, see Amaldi and Kann [2].

If not only the sign of a point is important but also the distance by which a point is misclassified, we can consider the following loss function $F(x_i, w) = \begin{cases} 0 & \text{if } y_i \cdot \langle x_i, w \rangle > 0 \\ -\langle x_i, w \rangle & \text{if } y_i \cdot \langle x_i, w \rangle \leq 0 \end{cases}$. Finding an optimal hyperplane for the sum of all these losses is no easier than in the case of the Minimum Perceptron Classification problem, but a few modifications enable very efficient algorithms. The first is to use unnormalized hyperplanes w (i.e. hyperplanes with arbitrary $||w||$), making the objective function convex. Both due to the fact that $w = 0$ would now be a trivial optimal solution and that $F(x_i, w)$ is not differentiable, the function can be smoothed over $\langle x_i, w \rangle = 0$, leading to $E(x_i, w) = \ln\left(1 + e^{-y_i \cdot \langle x_i, w \rangle}\right)$, which is also known as logistic regression. Note that for large absolute values of $\langle x_i, w \rangle$ with $||w|| = 1$, $E(x_i, w)$ approaches $F(x_i, w)$.

Problem 2 (Logistic Regression) *Let X be a d-dimensional set of points, where each point x_i is labeled by $y_i \in \{1, -1\}$. Then logistic regression aims to find a hyperplane $w \in \mathbb{R}^d$ minimizing $E(w) = \sum_{x \in X} \ln(1 + e^{-y_i \cdot \langle x_i, w \rangle})$.*

$E(w)$ is also known as the *cross-entropy error function* in literature. Often, to prevent overfitting, especially in the case of linearly separable data where $E(w)$ approaches 0 and $||w||$ approaches ∞, a regularization term $\lambda ||w||_p$ is added to the error function, where $p \in \{1, 2\}$ and $\lambda > 0$ is a penalization parameter.

On a related note, many learning tasks are also specifically studied when the available memory is severely limited or the data set to be processed is very large. In these cases, algorithms typically focus on summarizing the most relevant parts of the data, which is processed in an online fashion. If an algorithm is able to compute a sufficiently good summary with respect to the objective function with a single pass over the data, it is

referred to as a streaming algorithm [9]. Our goal is to design such an algorithm for the perceptron classifier and logistic regression.

Unfortunately, binary classification in the streaming model is usually a very difficult task. To see this for logistic regression, consider a point set X with $n-1$ points labeled 1 and a single point p labeled -1, where we seek a subset of X of size $n-1$ to approximate the optimal hyperplane. Let all 1-class points be on the border of the convex hull C induced by these points and let no three points be collinear. The adversary first submits all the 1-class points. Before reading p, we are forced to discard one and let x_1 be that point. Let C' be the convex hull of the remaining $n-2$ points. We have $C \setminus C' \neq \emptyset$, therefore the adversary submits p in $C \setminus C'$. Then the -1 and 1 classes are separable in the reduced point set and non-separable in the original point set. Let w' be any separating hyperplane for the reduced point set. Then $\langle x_1, w \rangle < 0$. Since the optimal hyperplane of the reduced point set has infinite norm, the contribution of $\ln(1 + e^{-\langle x_1, w \rangle})$ of x_1 to the error function of the original point set is ∞.

This simple example (as well as similar examples for the perceptron loss function) can be extended to account for regularization, random order streams and a randomized selection of points, with slight modifications.

This poses the question whether we are able to find relaxations to the problem such that a space efficient one-pass algorithm exists while at the same time the solution can interpreted in a similar way as before. To this end, we note that the objective function is defined in purely combinatorial terms and does not consider geometric properties of the point set, other than the relative position of the points with respect to a candidate hyperplane. Instead of minimizing the number of all misclassified points, we might try to minimize the number of strongly misclassified points for some suitable and, ideally, scalable notion of "strongly misclassified". Even if we are not able to find small summaries of the point set X, we might be able to summarize a point set W efficiently, where W is obtained by moving the each point of X by a small distance Δ, making points with $y_i \cdot \langle x_i, w \rangle \leq -\Delta$ strongly misclassified. Unfortunately, the parameter Δ cannot be chosen independently of the input point set, as we can scale the input to make perturbations by any small constant meaningless. The scale factor of the input has to be folded into the Δ, for which we use the directional width of X. With these relaxations we give an $(1+\epsilon)$-approximate algorithm for both perceptron classification and logistic regression in a streaming setting. To our knowledge, this is the first one-pass algorithm for any loss-based binary classifier in the streaming model.

Classification for Data Streams Various learning tasks such as clustering [3, 8], regression [**?**, 7] and classification have been studied in the streaming model. Specifically regarding binary classification, there has been extensive work on support vector machines. Assuming the data to be separable, support vector machines aim to find a hyperplane with maximum margin. In a streaming setting, algorithms produce summaries of the data

called coresets, such that a hyperplane with maximum margin on the summary has an ϵ-approximate maximum margin on the original data. The optimization of support vector machines can be formulated in terms of the minimum enclosing ball problem, see Tsang, Kwok and Cheung [10]. Coresets for the minimum enclosing ball problem have been widely studied and algorithms have either storage requirements exponential in d [1,4], do not compute ϵ-approximate coresets [5] or require multiple passes over the data [6,10].

References

[1] P. Agarwal, S. Har-Peled, and K. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004.

[2] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1–2):237 – 260, 1998.

[3] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC)*, pages 250–257, 2002.

[4] T. Chan. Faster core-set constructions and data stream algorithms in fixed dimensions. In *Comput. Geom. Theory Appl*, pages 152–159, 2003.

[5] T. Chan and V. Pathak. Streaming and dynamic algorithms for minimum enclosing balls in high dimensions. In *Proceedings of the 12th international conference on Algorithms and data structures*, WADS'11, pages 195–206, Berlin, Heidelberg, 2011. Springer-Verlag.

[6] K. Clarkson, E. Hazan, and D. Woodruff. Sublinear optimization for machine learning. *J. ACM*, 59(5):23, 2012.

[7] K. Clarkson and D. Woodruff. Numerical linear algebra in the streaming model. In *STOC*, pages 205–214, 2009.

[8] S. Har-Peled and A. Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

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

[10] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006.

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

SFB 876 Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

Subproject A3
Methods for Efficient Resource Utilization in Machine Learning Algorithms

Jörg Rahnenführer Peter Marwedel

Bottleneck Analysis for Machine Learning R Programs

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

R is a multi-paradigm language with a dynamic type system, different object systems and functional characteristics [1]. These characteristics support the development of statistical algorithms at a high level of abstraction. Although R is commonly used in the statistics domain, its performance problems when handling computation-intensive algorithms constitute a major disadvantage. Especially in the domain of machine learning, for example when analyzing high-dimensional genomic data, the execution of R programs is often unacceptably slow. Morandat et al. [2] analyzed R programs from different fields of statistics and were able to show major performance issues. Our goal is to overcome these issues, particularly focusing on machine learning programs. As a first step towards this goal, we used the traceR tool to analyze the bottlenecks arising in this domain. Our results support the development of alternative R interpreters [3] by uncovering the bottlenecks of real-world R code and providing optimization ideas [4].

Runtime Behavior Analysis

We examined the runtime behavior of the R language on a well-defined set of classical machine learning applications [5]. For our analyses we applied those algorithms to real-world data sets from UCI [6].

Figure 1 shows the average mean of the normalized runtime over all benchmarks. Here, we divided the runtime into eleven different categories, which in total form three groups. The first group contains external code parts like the time spent in C or Fortran code

(*External*). This group has the lowest optimization potential since it executes outside of the R interpreter. The second group are the functions provided by the R interpreter like arithmetic operations or built-ins (*Builtin/Special, Subset, Arith*). Their optimization potential varies by the specific function. The third group with the highest optimization potential are the internal tasks of the R interpreter like the garbage collection (*Lookup, Match, Duplicate, GC, MemAlloc, EvalList*). One of the largest interpreter internal

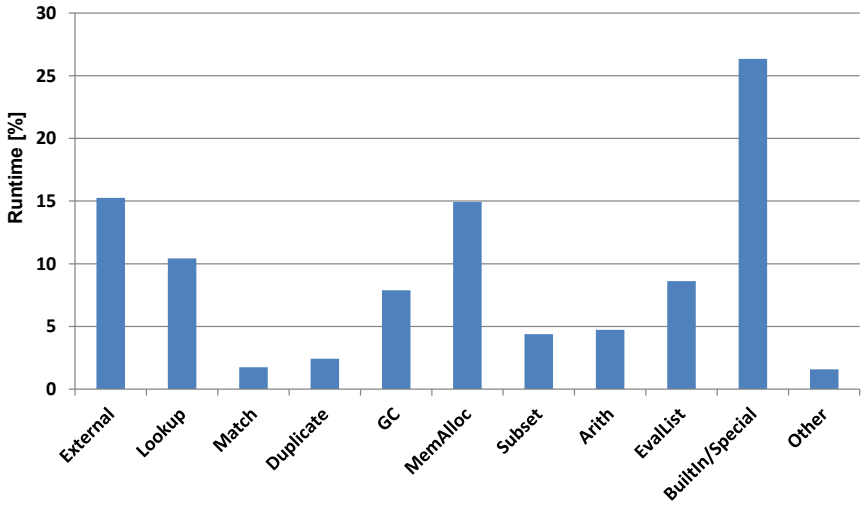


Figure 1: Runtime profile

bottlenecks is the *Lookup* function. The amount of time spent in looking up variables and functions is on average 11% and can amount up to 17% of the total runtime of a benchmark. This time is highly influenced by the amount of environments that build up the search path. Before an R function can be executed or a variable can be accessed the R interpreter has to look up its definition or value through a chain of environments. Each function call adds one more environment to this chain.

On average over 30% of the total runtime of all benchmarks is spent in built-in functions and arithmetic operations. This includes also the time needed for type checks and data conversion to support the dynamic type system of R. The overhead of these preprocessing steps could be optimized by the use of function specialization, which is a common compiler optimization. Since it would be hard to apply this technique to the GNU R interpreter, a more promising approach is to implement it for an alternative R interpreter running on top of a just-in-time compilation based environment, like the JVM.

To pass arguments though a function, the R interpreter creates a list data structure called pairlist. Due to the dynamic nature of R, function arguments can be passed by name, by position or via the ... argument. Our results show that in over 80% of all function calls, the call contained only zero or just one argument. This indicates that the machine learning benchmarks rarely use the full flexibility of argument passing, which

has a positive effect on time spent for matching those arguments (*Match*). Although the length of such an argument list is on average in the range between zero and one – and thus has a low memory footprint – it increases the overall time spent in memory management.

The results of the runtime analysis also indicate that the interpreter-internal tasks for memory management represent big bottlenecks with 8% spent in garbage collection (*GC*) and 15% spent in memory allocation (*MemAlloc*), especially for those benchmarks that spend less time in external code. On average over all benchmarks, the runtime spent in external code (*External*) amounts to 15%. Hence, there is still a high optimization potential on the R code side.

Memory Consumption Analysis

For our memory consumption analysis we first focused on the entire amount of memory allocated during the runtime of each machine learning benchmark. We therefore measured the allocations of new data in memory, but ignore the later removal of this data by the garbage collector. This approach gives a better view on the influence of the memory allocation behavior to the overall runtime.

Figure 2 shows the distribution of the memory allocations over all benchmarks between different categories of data structures. Here, the categories can be split into two groups: Data structures that are primarily used for interpreter-internal tasks like *Pairlists, Promises* and *Environments*, and structures that hold user data (*Vectors*).

The results show that when combined, all interpreter-internal data structures (*Pairlists, Promises, Environments*) sum up to 44% of the total allocated memory. This means that almost half of the allocated memory is used for executing the R program instead for the user data (*Vectors*) it processes. We analyzed also the different memory footprints and the amount of memory each data structure consumes. For the user data, R differentiates between small vectors that can store up to 16 double elements, and large vectors when the number of elements exceeds this limit. On average over all benchmarks more than half of all vectors allocated are small vectors, that furthermore contain only a single element. Single-element vectors have the best optimization potential: A 56 byte (on a 64 bit system) header is needed to manage an object whose size is just 4 or 8 bytes, thus in the worst case the header takes up 14 times as much memory as the user data.

This clearly shows a big potential for introducing scalar values that are not boxed within a vector and do not need a big header. If a scalar value is only used locally within a function, a just-in-time compiler may even be able to keep the value in a CPU register instead of storing it in main memory, saving the time for both allocation and garbage collection.

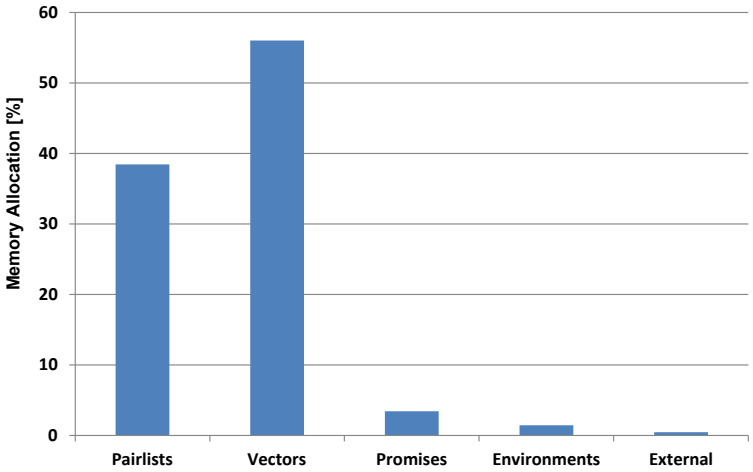


Figure 2: Memory profile

The results of our memory and runtime analyses uncover the bottlenecks of real-world R code and provide optimization ideas, supporting the development of alternative R interpreters. One existing alternative R implementation that targets the JVM is the fastR project [7] which we plan to use for future work as a basis to design optimizations specifically for computationally intensive machine-learning R applications.

References

[1] The R Project for Statistical Computing: R Language Definition, `http://cran.r-project.org/doc/manuals/R-lang.html`, 2013.

[2] Morandat, F. et al.: Evaluating the Design of the R language, In Proceedings of the 26th European Conference on Object-Oriented Programming, 2012.

[3] Kotthaus, H. et al.: A JVM-based Compiler Strategy for the R Language, Research Poster at The 8th International R User Conference, 2012.

[4] Kotthaus, H. et al.: Runtime and Memory Consumption Analyses for Machine Learning R Programs, Statistical Computing Workshop, Schloss Reisensburg, 2013.

[5] BenchR: TU Dortmund, `https://github.com/allr/benchR`, 2013.

[6] Bache, K. and Lichman, M.: UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, `http://archive.ics.uci.edu/ml`, 2013.

[7] FastR: Purdue University, `https://github.com/allr/fastr`, 2013.

Automatic model selection for high-dimensional survival analysis

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

Many different models for the analysis of high-dimensional survival data have been developed over the past years. While some of the models and implementations come with an internal parameter tuning automatism, others require the user to accurately adjust defaults, which often feels like a guessing game. Exhaustively trying out all model and parameter combinations will quickly become tedious or infeasible in computationally intensive settings, even if parallelization is employed. Therefore, we propose to use modern algorithm configuration techniques like iterated F-racing to efficiently move through the model hypothesis space and to simultaneously configure algorithm classes and their respective hyperparameters. In our application we study four lung cancer microarray data sets. For these we configure a predictor based on five survival analysis algorithms in combination with eight feature selection filters. We parallelize the optimization and all comparison experiments with the `BatchJobs` and `BatchExperiments` R packages [1].

Motivation The simultaneous reliable measurement of thousands of genetic variables has initiated the development and application of new statistical survival analysis methods for "small n, large p" problems. Whereas for regression and classification tasks in the area of machine learning, many benchmark data sets are established and comprehensive comparisons between statistical methods have been performed, in high-dimensional survival analysis it is still common practice to compare only a small number of methods on a small number of data sets [2, 7]. There is extreme need for objective and reproducible comparison studies in this field.

Obviously, the quality of the fitted models depends in many cases on the tuning of algorithm parameters. In addition, prefiltering of variables has often proved to be effective which also require the experimenter to set respective control parameters. The exhaustive evaluation of all algorithm and parameter combinations is infeasible, even if parallelization is employed, because single model fits can become computationally expensive.

A modern approach to solve this dilemma is to adaptively make all choices for the current data set at hand through an efficient black-box optimization approach that considers the desired performance measure. While in the past quite generic evolutionary algorithms have been used, modern methods are specifically tailored for the characteristics of this optimization problem, i. e., an expensive-to-evaluate objective function, noisy objective values and mixed numerical and categorical as well as hierarchical parameter spaces. The emerged research field has become known under the label 'algorithm configuration'. One prominent methodology is iterated racing [8].

Data We picked four publicly available lung cancer data sets. Data sets 'GSE31210', 'GSE4573' and 'GSE37745' were extracted from the Gene Expression Omnibus [4] database. The fourth data set 'Jacob' is taken from a website referenced in [9]. In addition to survival times and gene expressions (measured on the Affymetrix HGU1333a or HGU133 Plus 2.0 microarray chip and normalized using RMA [6]), we selected some important clinical covariates for each data set.

Experimental Setup The model building process consists of four steps: (1) split the data into training and test set, (2) apply a preselection filter on the training set, (3) pass the remaining covariates to a survival model and (4) quantify the performance of the model on the test set. In this process various hyperparameters are involved for both filters and models. Besides, the choice of filter and model can also be seen as hyperparameters which leads to a meta-algorithm dispatching on the respective statistical algorithms. Because the process is stochastic we need to repeatedly fit models with the same parameter configuration on multiple problem instances ω to assess meaningful performance measures. Thus the optimization task can be broken down to minimize an algorithm $\mathcal{A}(\omega, \theta)$ where $\theta \in \Theta$ is a configuration of the parameter space.

Iterated racing [8] explores the specific structure of the parameter space Θ and assumes a probability distribution Q over Θ. In an iterative procedure it now samples a set of new candidate configurations S from Q, races the current candidates to a low number of so-called elite configurations and adapts the distribution Q by centering it around the elites as well as reducing its spread. The latter results in exploration in the beginning and exploitation in the later stages of the optimization when the preliminary defined budget B, the maximum number of allowed function evaluation, gets depleted. This procedure is outlined in Algorithm 1.

Algorithm 1 Outline of iterated F-racing.

$Q \leftarrow$ uniform distribution on Θ
$S_{\text{elite}} \leftarrow \emptyset$
while budget B not exhausted **do**
  $S_{\text{new}} \leftarrow \text{sample}(\Theta, Q, N_{\texttt{new}})$ ▷ Sample N_{new} new configurations from Q
  $S_j \leftarrow S_{\text{elite}} \cup S_{\text{new}}$
  $S_{\text{elite}} \leftarrow \text{race}(S_j, B_j)$
  $Q \leftarrow \text{adapt}(Q, S_{\text{elite}})$
  $j \leftarrow j + 1$
end while
return S_{elite}

Results Table 1 compares the obtained results of our tuning approach with the performance measures of four established reference (or baseline) models that a reasonable experimenter might try. These include two boosting approaches implemented in the `R` packages `CoxBoost` and `mboost`, Ridge regression provided by the package `glmnet` and a simple Cox proportional hazards model [3] fitted on clinical covariates only. Our tuned models perform significantly better than all considered reference models at the level $\alpha = 5\%$. Furthermore, many interesting data set characteristics, otherwise inaccessible in high-dimensional settings, can be derived by examining the configurations.

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 for baseline models and configuration approach, cross-validated with three folds.

Outlook The analysis should be extended to consider more filters, models, parameters and data sets. We also plan to compare the iterated F-racing to model-based optimization [5] which allows a more fine-grained control over the optimization process. We have only configured survival analysis models for individual data sets. We think it might also be a worthwhile endeavor to try to configure the modeling for a whole domain of survival analysis data sets, e. g., (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] H. M. Bøvelstad, S. Nyågard, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjaerde. Predicting survival from microarray data–a comparative study. *Bioinformatics*, 23(16):2080–2087, 2007.

[3] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[4] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[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] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264, 2003.

[7] Kai Kammers, Michel Lang, Jan G Hengstler, Marcus Schmidt, and Jörg Rahnenführer. Survival models with preclustered gene groups as covariates. *BMC bioinformatics*, 12(1):478, 2011.

[8] 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.

[9] Kerby Shedden, Jeremy M G Taylor, Steven a Enkemann, Ming-Sound Tsao, Timothy J Yeatman, William L Gerald, Steven Eschrich, Igor Jurisica, Thomas J Giordano, David E Misek, Andrew C Chang, Chang Qi Zhu, Daniel Strumpf, Samir Hanash, Frances a Shepherd, Keyue Ding, Lesley Seymour, Katsuhiko Naoki, Nathan Pennell, Barbara Weir, Roel Verhaak, Christine Ladd-Acosta, Todd Golub, Michael Gruidl, Anupama Sharma, Janos Szoke, Maureen Zakowski, Valerie Rusch, Mark Kris, Agnes Viale, Noriko Motoi, William Travis, Barbara Conley, Venkatraman E Seshan, Matthew Meyerson, Rork Kuick, Kevin K Dobbin, Tracy Lively, James W Jacobson, and David G Beer. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827, 2008.

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 nerve cells were treated in vitro with different compounds of two different types (mercurial compounds and histone deacetylase inhibitors). Then genome-wide 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 train set contains 4-5 technical replicates of three representatives of each type and control compounds. The test set consists of 2 control compounds, 3 mercurials and 3 histone deacetylase inhibitors (HDACs) with 4 technical replicates representing

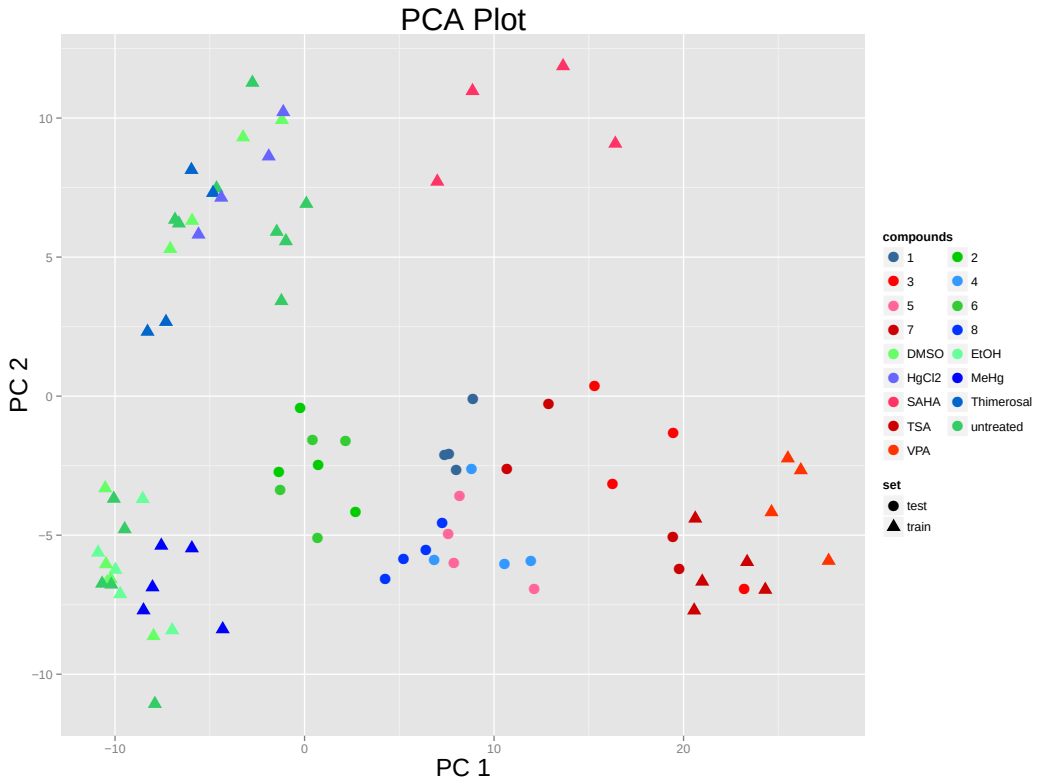


Figure 1: Principal components plot

each type. In the first step of classification task we aim to identify controls. In the second step we would like to distinguish between mercurials and HDACs.

Preliminary analysis

We have applied a large analysis pipeline to evaluate the influence of a compound on the gene expression changes as described already in co-authored articles [1,2], 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). All calculations were performed with the **R** programming language [3]. For example, we calculated the principal components for the train data and projected the expression values of the test data into them to visualize the results of the classification in Figure 1. The triangles represent the train samples, the circles represent the test samples. The green, blue and red color tones display compounds known to be or classified as controls, mercurals, and HDACs, respectively.

Classification

For both classification tasks we used support vector machine (SVM) with linear and gaussian kernels implemented in **kernlab** package. To reduce the number of variables we took 100 genes out of over 50 thousands with highest variation across all samples. In the first step we could classify the controls from the toxicants. In the second step we subtracted confirmed controls from the exposed samples and calculated for each sample and averaged expression level the probability to belong to HDAC type.

For 4 compounds predictions for each replicate were consistent, thus the overall predictions were unanimous. Predictions for two compounds were inconsistent, therefore no clear classification was possible for these toxicants.

Conclusion

In the present project we built classifiers to distinguish consecutively three different types of compounds: controls, mercurials and HDACs. Whilst we succeeded to classify the control compounds, our classifier made errors in separating mercurials from the HDACs. One possible reason for this is the complexity of the cellular response to the toxicants. One common feature like the presence of Me-atom as in case of mercurials or the inhibition of histone deacetylase as in case of HDACs is not necessarily sufficient to define homogeneous classes. The second possible reason for the false classifications may be the small sample size. One possible indicator for this consists in the fact that the classifier built on all 12 but one compound predicted the left toxicant perfectly.

Ongoing work

The cooperation with with Prof. Dr. Jan Hengstler includes several other projects, e.g. cellular response to various concentrations of a toxicant and cellular response after being exposed to toxicant for different periods of time. The results are published in journals [1,2].

Literature

[1] Anne K. Krug, Raivo Kolde, John A. Gaspar, Eugen Rempel, Nina V. Balmer, Kesavan Meganathan, Kinga Vojnits, Mathurin Baquié, Tanja Waldmann, Roberto Ensenat-Waser, Smita Jagtap, Richard M. Evans, Stephanie Julien, Hedi Peterson, Dimitra Zagoura, Suzanne Kadereit, Daniel Gerhard, Isaia Sotiriadou, Michael Heke, Karthick Natarajan, Margit Henry, Johannes Winkler, Rosemarie Marchan, Luc Stoppini, Sieto

Bosgra, Joost Westerhout, Miriam Verwei, Jaak Vilo, Andreas Kortenkamp, Jürgen Hescheler, Ludwig Hothorn, Susanne Bremer, Christoph van Thriel, Karl-Heinz Krause, Jan G. Hengstler, Jörg Rahnenführer, Marcel Leist, Agapios Sachinidis. 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] Nina Balmer,Stefanie Klima, Eugen Rempel, Violeta Ivanova, Lena Smirnova, Raivo Kolde, Matthias Weng, Kesavan Meganathan, Margit Henri, Agapios Sacchinidis, Thomas Hartung, Michael Berthold, Jan Hengstler, Jörg Rahnenführer, Tanja Waldmann, Marcel Leist. Histone modifications and switch from transient drug-induced transcriptome responses to disturbed neurodevelopment of pluripotent human stem cells (submitted).

[3] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2013, URL http://www.R-project.org/.

Subproject A4
Resource efficient and distributed platforms for integrative data analysis

Peter Marwedel Olaf Spinczyk Christian Wietfeld

CiAO/IPv6: Next-Generation Internet Protocols for Data Collection in Wireless Sensor Networks

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

Wireless sensor networks had been a domain for customized and often proprietary networking protocols. Recently, the next-generation Internet Protocol *IPv6* emerged in wireless sensor networks, providing excellent interoperability with traditional IP devices, such as personal computers and routers. This paper describes the design of *CiAO/IPv6*, a highly efficient Internet Protocol stack for wireless sensor networks. Such an infrastructure software forms the technical basis for distributed big-data collection in our everyday life.

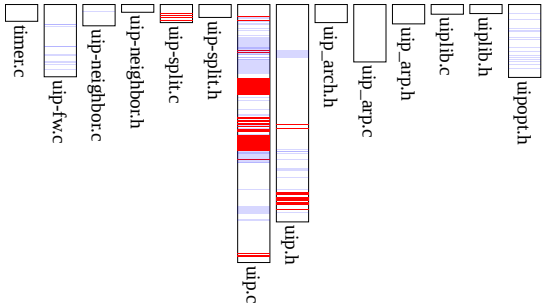
1 Introduction

Traditional wireless sensor networks are built upon customized (for example, *B-MAC* [7]) and proprietary (for example, *SimpliciTI* [4]) networking protocols. These protocols lack interoperability and the flexibility to connect wireless devices anywhere and at any time. The ubiquitous Internet Protocol in its current incarnation (*IPv4*), on the one hand, is running out of addresses and cannot be used to connect masses of new devices – such as wireless sensor networks. The next-generation Internet Protocol *IPv6*, on the other hand, offers connectivity to roughly 10^9 devices and is gaining attention in the wireless sensor networking research community [6].

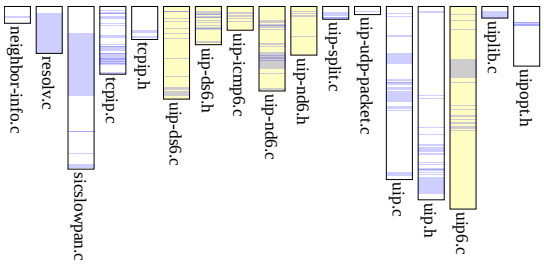
This paper outlines the challenges that the extension by IPv6 pose to low-level infrastructure software in the following section. A new design point is described in Section 3 that overcomes several shortcomings identified in state-of-the-art protocol stacks.

2 State of the Art

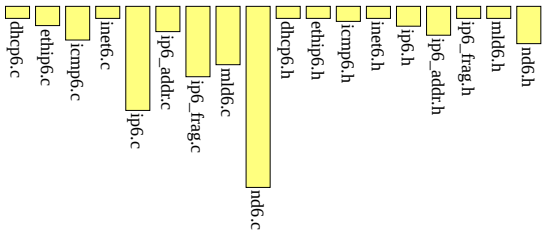
The implementation of IPv6 for resource-constrained embedded systems differs fundamentally from general-purpose networking stacks found in Windows, Linux, and other UNIX-like operating systems. The memory requirements of *embedded* implementations are several orders of magnitude lower to meet the capacities of low-cost embedded devices. *Micro-IP (uIP)* and *lightweight-IP (lwIP)* [2] are probably the most prominent examples of open-source implementations for the domain of embedded (sensing) systems. Both implementations support IPv4 *and* IPv6 – however, code quality is traded for that particular feature. Figure 1 shows the ex-post extension by IPv6 in uIP and lwIP. In uIP, the IPv4-specific code locations are replaced by IPv6 code fragments. This procedure sacrifices dual-stack functionality and is discarded in uIP's successor *Contiki* [3]. In contrast to code replacement, code duplication is used in Contiki to implement IPv6 and IPv4 side-by-side. For example, `uip6.c` and `uip.c` share large amounts of similar code. A hybrid approach was chosen for lwIP, which contains several files dedicated to IPv6 only, which are integrated into the remaining code base by conditional preprocessor directives and opaque function pointer indirection. Thus, neither implementation distinguishes itself by a clean and seamless IPv6 integration.



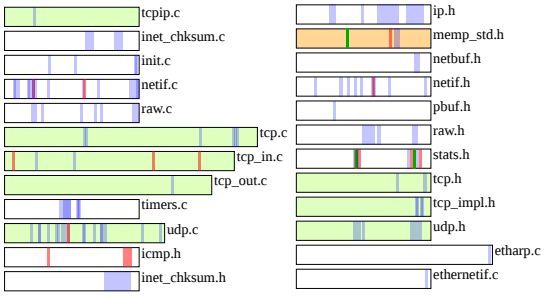
(a) **uIP:** Configurable IPv6 code is highlighted red.



(b) **Contiki:** Files for IPv6 are shown with yellow background color.



(c) **lwIP:** 18 files are dedicated to IPv6 only.



(d) **lwIP:** IPv6 integration points are marked blue.

Figure 1: Code scattering of the IPv6 feature in state-of-the-art implementations for resource-constrained embedded systems [5]

3 Design Approach

The design of an extensible networking stack requires a different methodology than the aforementioned ad-hoc solutions. Configurability has to be the primary design goal for an exchangeable IP layer. *CiAO/IP* [1] follows such a design methodology by exercising aspect-oriented design idioms.

A critical part of the design – with regard to extensibility – is the handling of incoming network packets inside the protocol stack. When a new packet arrives on the network, little is known about its application-layer destination. That information has to be extracted from the packet headers layer-by-layer. Thus, the extraction of transport layer headers, for example TCP and UDP, *depends* on preceding network-layer protocol processing (IP).

An extensible design requires to resolve such dependencies between the protocol layers in order to allow the implementation of well-defined software components that contain functionality of only one single layer. Figure 2 illustrates the aspect-oriented design idiom *upcall dispatcher hierarchy* [8] that resolves dependencies by loose coupling of software components. The colored boxes denote *aspects* that are interposed between several network protocol instances: the control flow is arranged by these aspects to deliver incoming packets to the appropriate protocol instances in the right order. Such a design allows to implement each protocol independently, and the entire system is finally composed using such *upcall dispatcher aspects*. This design idiom, in particular, offers dual-stack functionality, produces no code duplication and no code scattering. Increased efficiency (90% lower memory consumption and 20% higher throughput compared to lwIP), readability and maintainability (60% fewer lines of code compared to lwIP) are the benefits of this approach.

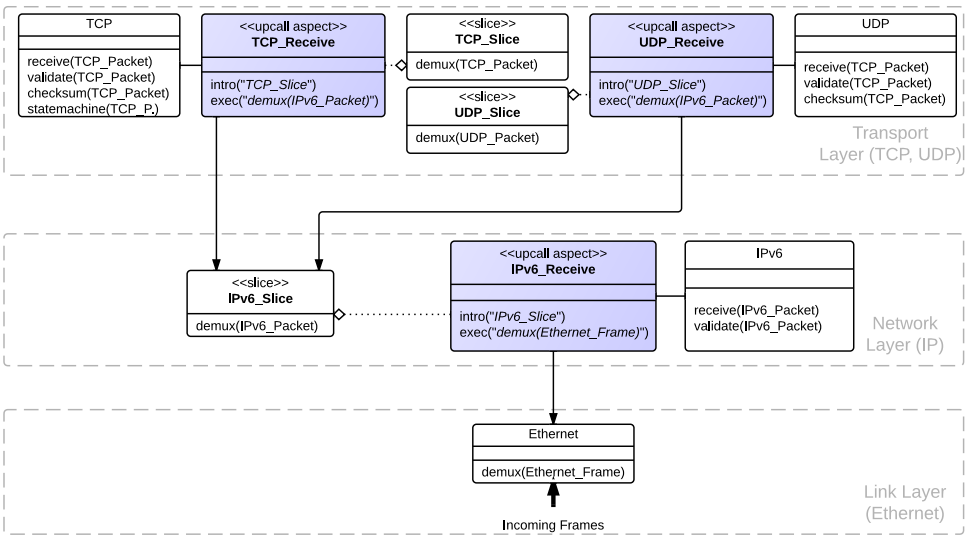


Figure 2: Design for extensibility of the *CiAO/IPv6* stack [5]

4 Conclusions

This paper highlights that state-of-the-art IPv6 implementations for wireless sensor networks offer poor code quality, because of code duplication for IPv4 and IPv6, opaque usage of conditional preprocessor directives and function pointer indirections, leading to scattered IPv6 code fragments. CiAO/IPv6 constitutes a new design point of integrating IPv6 into the existing CiAO/IP stack, and provides excellent code quality by aspect-oriented design idioms. These attributes make CiAO/IPv6 an appealing target for static source code analyses to create a resource model to estimate its runtime and energy costs.

References

[1] Christoph Borchert, Daniel Lohmann, and Olaf Spinczyk. CiAO/IP: a highly configurable aspect-oriented IP stack. In *10th Int. Conf. on Mobile Systems, Applications, and Services (MobiSys '12)*, pages 435–448, New York, NY, USA, June 2012. ACM.

[2] Adam Dunkels. Full TCP/IP for 8-bit architectures. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 85–98. ACM, 2003.

[3] Adam Dunkels, Björn Grönvall, and Thiemo Voigt. Contiki - a lightweight and flexible operating system for tiny networked sensors, 2004.

[4] Larry Friedman. SimpliciTI: simple modular RF network specification. pages 1–34, 2009.

[5] David Gräff. CiAO/IPv6: Aspektorientierte Erweiterung eines Netzwerkprotokoll-Stacks für eingebettete Systeme. Bachelorarbeit, Technische Universität Dortmund, November 2012.

[6] Jonathan W. Hui and David E. Culler. IP is dead, long live IP for wireless sensor networks. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, SenSys '08, pages 15–28, New York, NY, USA, 2008. ACM.

[7] Joseph Polastre, Jason Hill, and David Culler. Versatile low power media access for wireless sensor networks. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, SenSys '04, pages 95–107, New York, NY, USA, 2004. ACM.

[8] Jochen Streicher, Christoph Borchert, and Olaf Spinczyk. Upcall dispatcher aspects: Combining modularity with efficiency in the CiAO IP stack. In *1st AOSD W'shop on Modularity in Sys. Softw. (AOSD-MISS '11)*, pages 23–27. ACM, March 2011.

MobiSIM and MIMOSA - Simulation and Measurement of the Resource Behavior of Embedded Systems

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

The estimation on resource consumption in highly resource-constrained systems is challenging, as it might depend on a broad variety of hardware and software features within a system's product family. Additionally, context parameters have a significant influence on the system's behavior. To address these problems, MobiSIM was developed within the A4 project, and was evaluated in cooperation with A1 using their application patterns. The concept of MobiSIM is to utilize highly detailed and flexible resource models, while using coarse-grained system models, which by that are easy to handle. The benefit of MobiSIM is, that resource models can be easily implemented and tested. Further, we were able to show that for several simulation purposes, the application code can be modeled by a behavioral description, allowing to evaluate application strategies before they are actually implemented on the device.

Further, the formerly presented MIMOSA technology was evaluated in terms of accuracy and linearity.

1 MobiSIM - Simulation of embedded systems

Today, over 500 million smart phones have been sold. Certain classes of social applications, namely *collective apps*, *internet queries* and *crowdsourcing applications*, are widely used to gain information about the user's environment. As an example, *Google Maps*

collects location information to sense and visualize traffic jams. To optimize resource utilization, several strategies might be implemented. For instance, the frequency of positioning requests and the frequency and technology of transmissions to the servers may vary. In dynamic approaches, these parameters may be changed during runtime, according to environment conditions (*context*), like the radio channel quality.

To evaluate different strategies, they usually have to be implemented in a test bed which allows to measure the resource behavior of the system. This can be a problem, since those test beds often simply do not exist.

To evaluate application strategies as well as our resource models, the MobiSIM simulation platform was implemented [2] within the OmNeT++ simulation framework. The core of MobiSIM consists of energy models elaborated within the A4 project. In the first prototypical implementation, a very detailed LTE energy model was introduced, consisting of the fixed-datarate energy model (contribution of Björn Dusza, Figure 1, top-left), which was extended by the timing and energy behavior described in the LTE specification and other sources. Measurements vs. simulation graphs for a single transmission are shown in Figure 1, top-right. All relevant parameters of the model are configurable, so that the model can easily be applied to simulations of different hardware devices. Figure 1, bottom-left, shows the result of a benchmark that was implemented in the simulator and on the real hardware as well. The benchmark sends LTE packages of different sizes and in changing intervals. To see the deviation of the real and the simulated energy, the values were integrated in Figure 1, bottom-right.

To model the application behavior, the simulator offers a C++ based *aplication programming interface* (API) that allows to break down algorithms and strategies to a few lines of code when evaluating applications in the considered domain. Listing 1 shows an example of an application that checks the battery state all 5 minutes and only transmits stored data via LTE, when the battery level is above 10%. Not only energy usage is simulated in this example, but also memory consumption and CPU utilization.

2 Evaluation of the MIMOSA technology

In the 2012 report, the MIMOSA[1] technology was introduced, which facilitates the accurate measurement of low-energy devices [3, pp. 44-47]. At the time of that report, accuracy and linearity had not been evaluated, yet. This was done recently and was presented at the RealWSN workshop [1].

To evaluate the linearity and accuracy, two experiments were carried out. In the first experiment, a set of 10 high-precision ohmic resistors was used to create a well-known

[1] "**M**essgerät zur **i**ntegrativen **M**essung **o**hne **S**pannungs**a**bfall", German for "Measurement device for integrative measurements without voltage drop"

```
if( isTimerMessage(msg) )
{
  if( GETGLOBAL(GLOBAL_BATTERY_LEVEL) > 10 )
  {
    sendLTE(
      GETGLOBAL(GLOBAL_MEMORY_ALLOCATED)
    );
    memoryClear();
  }
  useCPU(10);
  startTimer(3600);
}
```

Listing 1: Example application logic: Transmit via LTE if available and battery-level greater than 10%, clear memory afterwards. Adapted from [2]

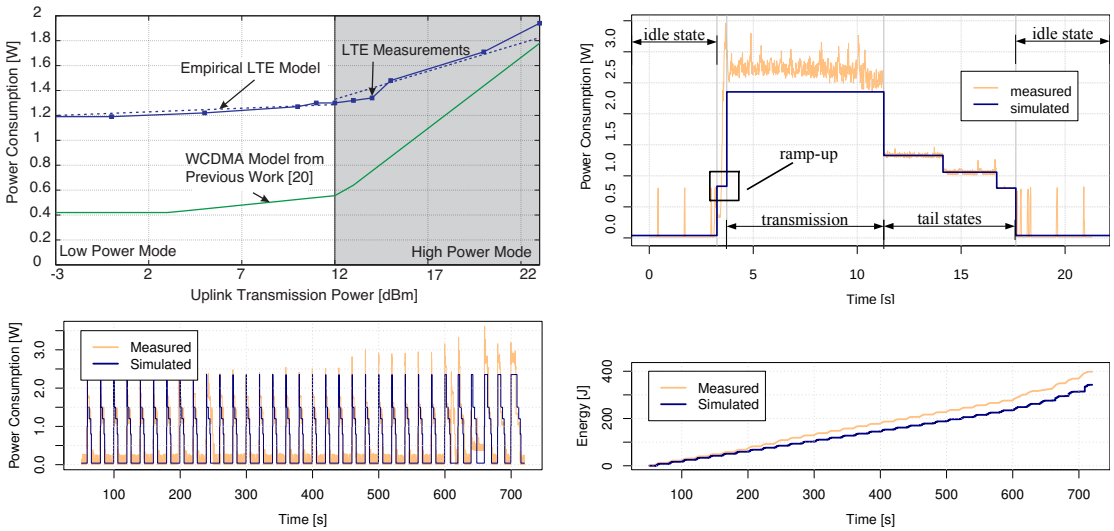


Figure 1: LTE energy model and benchmark, from [2]

load. The energy consumed by this load was measured by MIMOSA, as shown in Figure 2, top-left. Further measurements have shown that MIMOSA behaves linear in all considered measurement ranges. Figure 2, top-right shows the noise measured in the given range, which stays in the order of ± 3.5 nJ and is, as expected, almost constant over the whole range.

In the second experiment we measured the sub-sample linearity, which demonstrates MIMOSAs unique ability to sense short energy pulses shorter than the sample time. For that purpose, pulse in the range from 0.17 μs to 3.17 μs were generated (sample time: 10 μs). We were able to show that MIMOSA also behaves linear here, while keeping the noise even under 2 nJ. The results are shown in Figure 2, bottom.

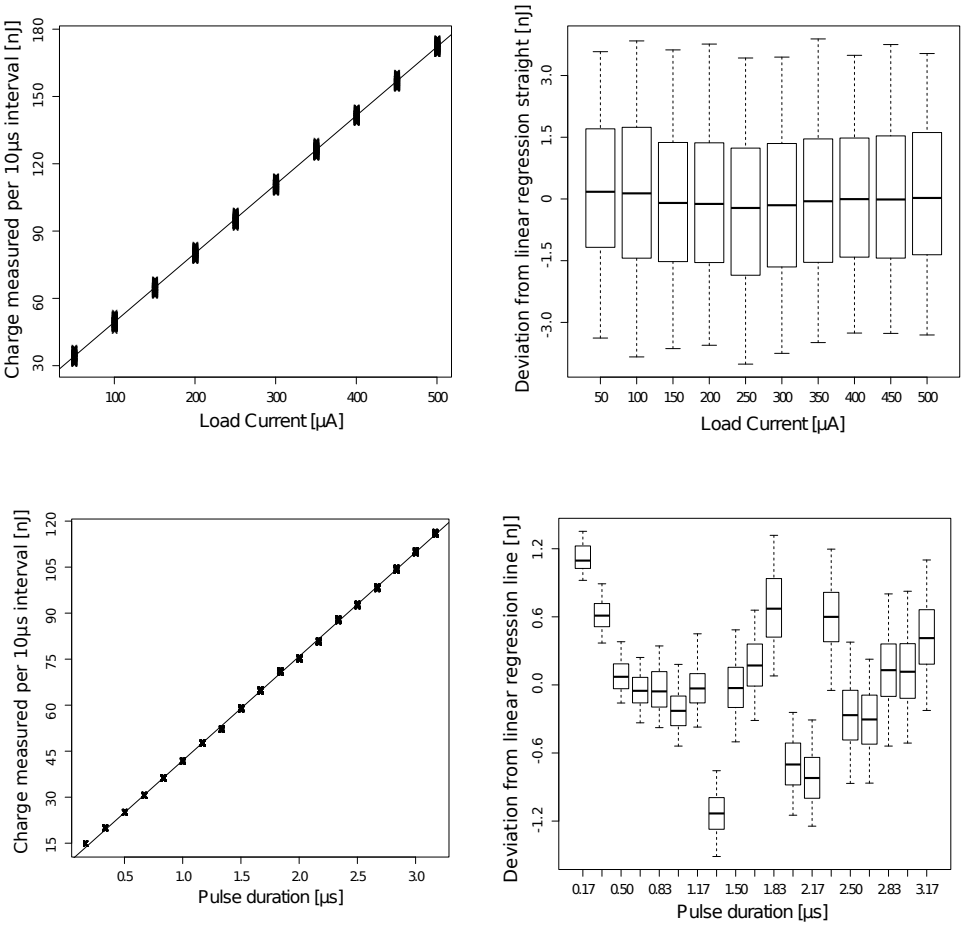


Figure 2: MIMOSA accuracy evaluation, from [1]

References

[1] Markus Buschhoff, Christian Günter, and Olaf Spinczyk. Mimosa, a highly sensitive and accurate power measurement technique for low-power systems. In *Real-World Wireless Sensor Networks*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013.

[2] Markus Buschhoff, Jochen Streicher, Björn Dusza, Christian Wietfeld, and Olaf Spinczyk. Mobisim: A simulation library for resource prediction of smartphones and wireless sensor networks. In *Proceedings of the 46th Annual Simulation Symposium*, ANSS '13. Society for Computer Simulation International, 2013.

[3] Katharina Morik and Wolfgang Rhode (Editors). Technical report for collaborative research center sfb 876 - graduate school. Technical Report 2, TU Dortmund University, September 2012.

Context-Aware Battery Lifetime Modeling for LTE User Equipment

Björn Dusza
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
bjoern.dusza@tu-dortmund.de

In this report, a context-aware power consumption model for LTE user equipment is presented that incorporates not only system parameters (e.g. allocated bandwidth) but also context parameters such as the cell environment and the application characteristics. The results show that these particular influences, that have not been considered by energy models so far, have a significant impact on the power consumption of a device and therefore its expected battery lifetime.

1 Motivation

Improving the energy efficiency of portable communication equipment such as smartphones has recently gained increased attention in the research community. The reason for this is, that battery capacity is not evolving as fast as the power demand of recent devices with bright large displays, multi-core CPUs and multiple communication interfaces. For the quantitative performance evaluation of energy-aware protocols and algorithms, accurate power consumption models are required that incorporate all influencing factors regarding the power consumption. However, extensive literature research discloses that in most cases very simple power consumption models incorporating only a few discrete system states (e.g. idle, active) are used for the performance evaluation of novel schemes.

Figure 1: Chocolate melted on LTE data stick due to waste heat.

2 Empirical Power Consumption Modeling

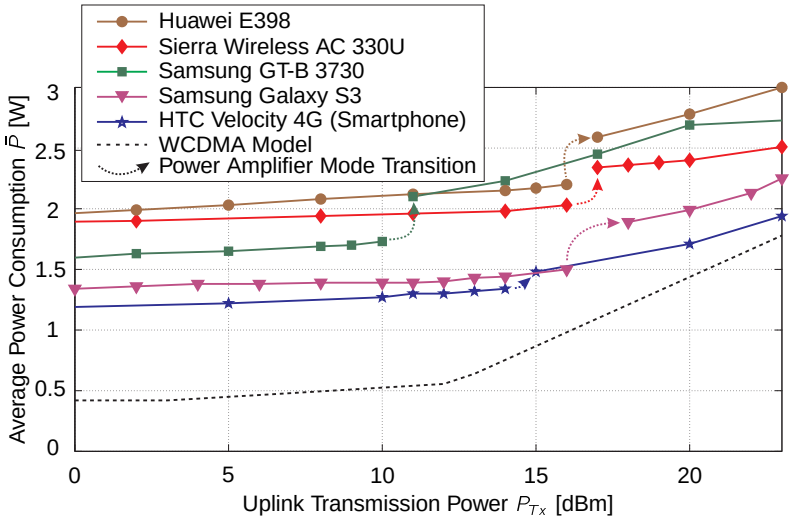


Figure 2: Device-specific Power Consumptions for Different Uplink Transmission Power Values (all in 2.6 GHz).

To provide a more solid base for actual energy efficiency analyses, in [1] a measurement based power consumption model for LTE UE has been presented that describes the relationship between the uplink transmission power P_{Tx} and the actual power consumption $\bar{P}$ of the device under test. The results show that all devices under investigation are characterized by a power consumption curve that can be split into two pieces. The single parts of the curve, representing the low power mode and the high power mode of the power amplifier, can be independently modeled by two linear functions

$$\bar{P}(P_{Tx}) = \begin{cases} \alpha_L \cdot P_{Tx} + \beta_L & \text{for } P_{Tx} \leq \gamma \\ \alpha_H \cdot P_{Tx} + \beta_H & \text{for } P_{Tx} > \gamma \end{cases} \quad (1)$$

with the device specific parameters α, β and γ provided in [1] and the uplink transmission power P_{Tx}.

3 CoPoMo - A Context-Aware Power Consumption Model

For the quantification of potential battery lifetime improvements that can be achieved by energy-aware protocols, the user behavior as well as the environment need to be taken into account. For this purpose, the Context-aware Power consumption Model (CoPoMo) has been developed [2]. The model bases on a four staged Marcov chain (cf. Fig. 3) in which the different states represent different power consumption modes.

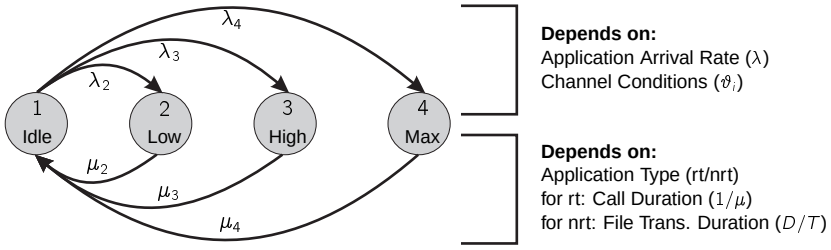


Figure 3: A Markovian Model for Power Consumption of LTE UE.

While in the idle state the User Equipment (UE) is not transmitting any data, the active states (low, high and max.) are characterized by different power costs as well as different data rates. The probability that a UE at a randomly chosen position inside the cell will enter a dedicated state for data transmission is significantly influenced by the environment. This includes the cell radius as well as the frequency band, building density, antenna patterns and a lot more cell specific parameters. Fig. 4 illustrates the spatial distribution of the three active states of the UE (low, high and max) for one example environment. A UE moving through the cell cuts across different areas whereas the probability for entering a state (e.g. high) corresponds to the relative ratio of the three different zones in Fig. 4. Beyond this, different additional parameters such as the average file size, the arrival rate of transmission requests, the bandwidth allocation (i.e. number of physical LTE ressource blocks) and the UE specific parameters α, β and γ described in the previous section are taken into account. In [3], the quantitative impact of single parameter variations on the expected battery lifetime has been investigated in detail.

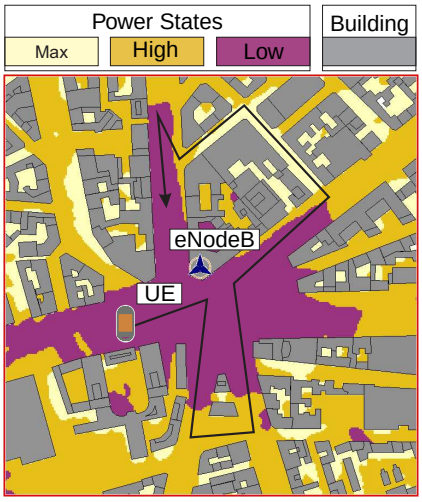


Figure 4: An Illustration of the Spatial Power Mode Distribution in an Example Environment.

4 Example Results

One important example results from [2] that illustrates the need for a detailed power consumption model incorporating system as well as context parameters is illustrated in Fig. 5 where the long term average power consumption P_Σ as well as the approximated battery lifetime for an HTC Velocity 4G smartphone are shown for different application arrival rates. One can see that for low traffic (cf. Fig. 5 (1)) the average power consumption P_Σ converges towards the power consumption in idle state for $\lambda \to 0$. This

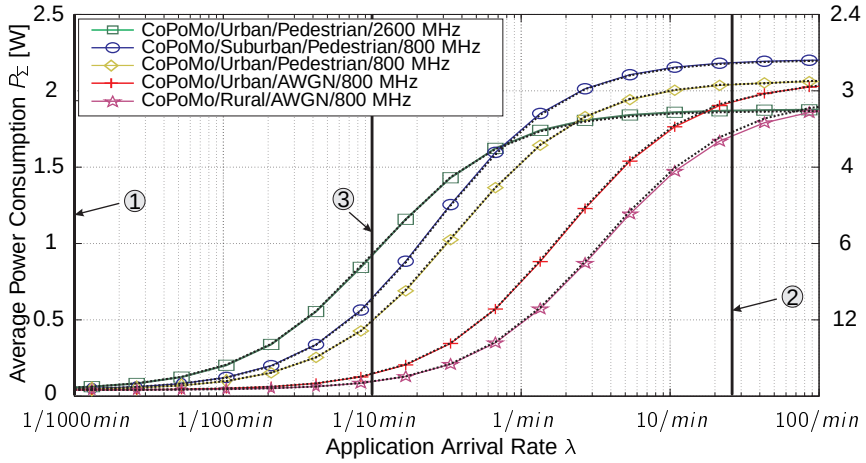


Figure 5: Impact of the Cell Environment on the Average Power Consumption for Different Traffic Characteristics incl. Validation by System Simulation (HTC Velocity 4G, D = 10^8 Bit, $\#PRB = 50$ (continuously allocated), SNR_T =13 dB, $E_{Batt} = 6$ Wh)

is due to the fact that the UE is actually inactive for most of the time. For very high arrival rates λ (cf. Fig. 5 (2)) the UE is almost continuously active and P_Σ converges towards a maximum, which contingents on the context-dependent state probabilities p_i as well as the device parameters for the specific LTE frequency band. For an average application arrival rate of $\lambda = 1/10min$ (cf. Fig. 5 (3)), which could correspond to web-surfing usages, e.g. via multimedia applications such as Instagram, one can observe a significant impact of the context. While in an rural/AWGN/800 MHz scenario a battery lifetime of about 60 hours can be achieved, the battery needs to be recharged after nine hours if the scenario is suburban/pedestrian/800 MHz and after only seven hours in an urban/pedestrian/2600 MHz scenario.

References

[1] Bjoern Dusza, Christoph Ide, Liang Cheng, and Christian Wietfeld. An Accurate Measurement-Based Power Consumption Model for LTE Uplink Transmissions. In *Proc. of IEEE INFOCOM (Poster)*, Turin, Italy, April 2013. IEEE.

[2] Bjoern Dusza, Christoph Ide, Liang Cheng, and Christian Wietfeld. CoPoMo: A Context-Aware Power Consumption Model for LTE User Equipment. *Transactions on Emerging Telecommunications Technologies (ETT), Wiley*, 2013.

[3] Bjoern Dusza, Christoph Ide, and Christian Wietfeld. Quantitative Bewertung des Einflusses von Kontext- und Systemparametern auf die Batterie-Laufzeit von LTE Endgeräten. In *Proc. of the 18th VDE/ITG Fachtagung Mobilkommunikation,*, Osnabrück, Germany, May 2013. VDE press.

Resource-Efficient Spectrum Sharing in Ad-Hoc Femtocell Networks

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

Femtocells are a promising solution regarding the intense demand of today's mobile data traffic. Despite their capacity benefits due to high spatial reuse, the use of femtocells implies a number of challenges, especially Inter-Cell-Interference Coordination with existing macrocells. Many solutions have been proposed to mitigate Inter-Cell-Interference, but all of them require knowledge of the radio resource environment. In the start-up phase of femtocells or in case of black-out recovery, this knowledge is not available because of missing channel quality and measurement reports. In order to allow immediate femtocell activation in these situations, we propose to use a combination of Random Frequency Hopping and Fractional Frequency Reuse, which is able to reduce the amount of interference between femto- and macrocells or equivalently reduce the required transmission power of femtocells while guaranteeing a certain Quality of Service. The approach is evaluated by analytical models and simulations for the Signal-to-Noise-and-Interference Ratio and the Bit Error Ratio.

1 Motivation

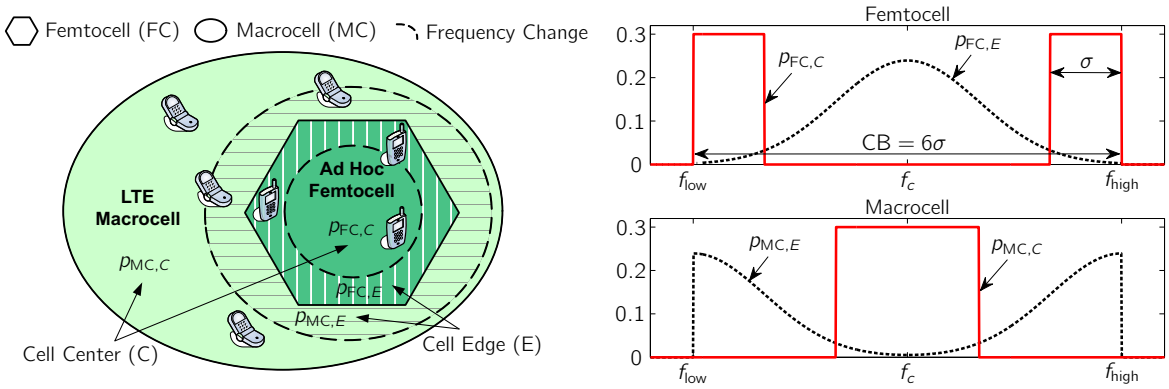
Due to continuously growing user demands for broadband data in mobile radio systems, existing macrocells are no longer able to keep pace with this progress. Even with introduction of new transmission technologies like Long Term Evolution Advanced (LTE-A) or provisioning of additional spectrum, the provided capacity is insufficient. Hence, the integration of small cells such as femtocells is necessary, which can offload excessive traffic

from macrocells. In 2012, femtocell access points already outnumbered traditional base stations [1]. In order to provide sufficient capacity, femtocells typically share the same frequency resources as macrocells. Since femtocells are applied at random positions and are randomly activated, severe Inter-Cell Interference (ICI) is introduced between both types of cells. In the downlink, the macrocell base station interferes with femtocell users, especially femtocell edge users as they suffer from low received powers. In the uplink, macrocell users interfere with the femtocell access point. This interference is even reinforced when closed femtocells are applied, such that macrocell users are not allowed to connect to the femtocell access point even when they are close to it.

Many ICI mitigation approaches have been proposed in order to coordinate the applied resources of the users in both types of cells in a decentralized way. Well known methods are distribution of spectrum allocation and application of macrocell Fractional Frequency Reuse, cf. [1]. The introduction of LTE-A within Rel. 10 has also pushed forward Inter-Cell-Interference Coordination (ICIC) by carrier aggregation in the frequency domain and Almost Blank Subframes (ABS) in the time domain. Moreover, due to the random location and activity of femtocells, self-organizing ICI mitigation is desirable, such as autonomous power control, adaptive channel allocation, and frequency assignment [1]. All of the known ICI mitigation and coordination approaches require knowledge of the interference environment, i.e. which time and frequency resources are allocated by the surrounding cells, in order to avoid these intervals and subbands. Therefore, the methods fail when no wideband Channel Quality Information (CQI) is available because of limited feedback channels, start-up phases of femtocells, or black-out recoveries. In order to provide interference mitigation also in these situations, we propose to use Random Frequency Hopping (RFH) combined with Fractional Frequency Reuse (FFR).

2 Random Frequency Hopping with Fractional Frequency Reuse

The combination of Random Frequency Hopping with Fractional Frequency Reuse is able to minimize ICI without knowing which resources are allocated by surrounding cells. Therefore, each femtocell user transmitting data chooses random carrier frequencies which are changing across time. In every OFDM symbol, a new carrier frequency is selected according to a given probability density function (pdf). In this way, each transmitter is linked to a random hopping pattern, which is signalized to the receiver before transmission. In order to minimize ICI as far as possible, macro- and femtocell users which are located close to each other should use orthogonal pdfs. Therefore, both the femto- and macrocell are divided into cell center and cell edge as shown in Fig. 1a. Each region is associated to its own carrier frequency pdf, depicted in Fig. 1b. In this way, femtocell edge users apply a quasi-orthogonal hopping pdf (p_{FE}) compared to macrocell edge users (p_{ME}). To maintain a high spectral efficiency, frequencies used in the cell



(a) Principle/Scenario

(b) Probability Density Functions

Figure 1: Random Frequency Hopping Combined with Fractional Frequency Reuse

center of the macrocell (p_{MC}) are reused by femtocell edge users (p_{FE}) and frequencies used in the cell center of the femtocell (p_{FC}) are reused by macrocell edge users (p_{ME}).

In order to evaluate the gain of Random Frequency Hopping combined with FFR compared to systems based on wideband-CQI and systems where the interferers transmit on the same subcarriers as the femtocell users, analytical models for the SINR and BER are derived. As shown in [3], the SINR of femtocell users can be determined by

$$\mathrm{SINR} = \frac{1}{\frac{1}{D^2 N} \sum\limits_{MC} \frac{2\sigma_{MC}^2}{2\sigma_{FC}^2 + \varrho^2} \sum\limits_{s=0}^{N-1} \sum\limits_{d=0}^{N-1} \mathbb{E}\left\{\beta_{s,d}^2\right\} + \frac{1}{\mathrm{SNR}}}, \tag{1}$$

where MC denotes the surrounding macrocells, D the OFDM data carrying part, N the number of subcarriers, $2\sigma_{MC}^2$ the power of the macrocell Rayleigh fading, $2\sigma_{FC}^2 + \varrho^2$ the power of the femtocell Rice fading and $\beta_{s,d}$ the interference symbols within the receiver. The more precise the orthogonality of the pdfs between the macro- and femtocell edge users in Fig. 1b, the smaller the interference parameter $\beta_{s,d}$ and hence the higher the SINR. In a similar way, the BER of femtocell users within multi-macrocellular environments is calculated as [2]

$$\mathrm{BER}_s = \frac{1}{2\pi \mathrm{j}} \int\limits_{c-\mathrm{j}\infty}^{c+\mathrm{j}\infty} \mathbb{E}\left\{\mathrm{e}^{-\alpha n_s}\right\} \mathbb{E}\left\{\mathrm{e}^{-\alpha A_{FC}}\right\} \prod_{MC} \mathbb{E}\left\{\prod_{d=0}^{N-1} \cosh(\alpha A_{MC} Q \beta_{s,d})\right\} \frac{\mathrm{d}\alpha}{\alpha}, \tag{2}$$

where n_s represents the Additive White Gaussian Noise, A_{FC} the femtocell fading amplitude, A_{MC} the macrocell fading amplitude, and Q the subcarrier spacing. As (2) does not have a closed form solution, it has to be numerically approximated.

Fig. 2 shows the BER of Random Frequency Hopping according to (2) as a function of the macrocell size when compared to systems without ICI (best case), systems where the macrocell interferers use the same subcarriers as the femtocell users (worst case),

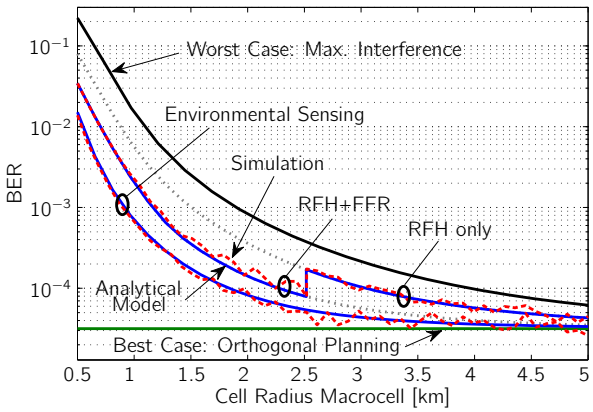


Figure 2: BER of Random Frequency Hopping with FFR

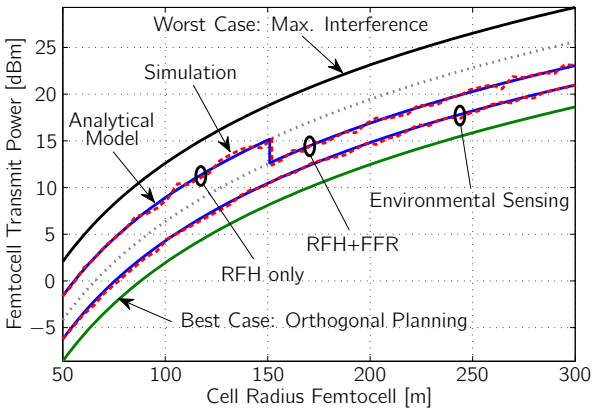


Figure 3: Femtocell Power of Random Frequency Hopping with FFR

and systems based on knowledge of the radio resource environment where femtocell users allocate frequencies with the lowest interference within the given bandwidth (environmental sensing). As the ICI is high for small macrocell sizes (<2.5 km), Random Frequency Hopping is combined with FFR in order to reduce the interference. The ICI is quite low for large macrocell sizes, such that Random Frequency Hopping can be used standalone. It can be seen that the combination of Random Frequency Hopping with FFR is able to reduce the BER up to a factor of 6.5 compared to the worst case and up to a factor of 2.3 compared to Random Frequency Hopping when used standalone. Since Random Frequency Hopping is able to reduce the interference when the transmit power is fixed, it can also be applied to reduce the required transmit power of femtocells when a certain SINR or BER threshold is claimed. Fig. 3 depicts the required transmit power of femtocells as a function of the femtocell size for an SINR upper bound of 10 dB. The ICI is low in case of small cell sizes (< 150m), such that Random Frequency Hopping can be used standalone. For large cell sizes, the ICI is increases, which is counteracted by a combination of Random Frequency Hopping and FFR.

References

[1] J. G. Andrews et. al. Femtocells: Past, Present, and Future. *IEEE Journal on Selected Areas in Communications*, 30(3):497–508, April 2012.

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

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

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

Gabriele Kern-Isberner Joachim Biskup

Secrecy preserving BDI Agents based on Answerset Programming

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

We consider secrecy from the point of view of an autonomous knowledge-based and resource-bound agent with incomplete and uncertain information, situated in a multi agent system. We investigate properties of secrecy and the preservation thereof in this setting and formulate desirable properties. Based on these ideas we develop a flexible BDI-based agent model and define an instance widely based on answer set programming. We show that and how our model and instance satisfy the proposed properties. We implemented our developed extendable framework for secrecy-preserving agents based on JAVA and answer set programming.

1 Principles of Agent based Secrecy

On the topic of secrecy a large body of work exists and diverse definitions of secrecy in various settings with different properties have been developed. For multiagent systems the main research focus herein lies on strong notions of secrecy of a whole (multiagent) system, for an overview see [2,5]. Secrecy is generally imposed by some global definition of secret information from a global, complete view of the entire system. While substantial work on the definition of secrecy exists mechanisms for secrecy preservation in multiagent systems are lacking.

We consider secrecy and secrecy preservation from the point of view of an autonomous knowledge-based agent with incomplete and uncertain information, situated in a multiagent system. Agents reason under uncertainty about the state of the environment, the reasoning of other agents and possible courses of action. For the representation of the

secrecy scenario it is convenient to focus on the communication between two agents, the modeled agent $\mathcal{D}$ which wants to defend its secrets from a potentially attacking agent $\mathcal{A}$. The definition of secrecy is complex and dependent on various aspects which influence the actually obtained secrecy and restriction of information flow. Secrets are not uniform in their content as an agent has different secrets with respect to different agents. Secrets are also not uniform with respect to their strength. That is, an agent wants to keep some information more secret than other. These differences in strength of secrets arise naturally from the value of the secret information. Secrets are also not static, they arise, change and disappear during runtime of an agent such that it has to be able to handle these changes adequately. These considerations lead to the following formulation of properties of secrets: (S1) secrets can be held with respect to specific agents, (S2) secrets can vary in strength, (S3) secrets can change over time.

Defining secrets does not define the preservation of secrecy and its properties. The intuitive formulation of our notion of secrecy preservation can be formulated as: *An agent $\mathcal{D}$ preserves secrecy if, from its point of view, none of its secrets Φ that it wants to hide from agent $\mathcal{A}$ is, from $\mathcal{D}$'s perspective, believed by $\mathcal{A}$ after any of $\mathcal{D}$'s actions (given that $\mathcal{A}$ does not believe Φ already).*

The actual quality of secrecy preservation is highly dependent on the accuracy of the view of $\mathcal{D}$ on the agent $\mathcal{A}$ and its supposed reasoning capabilities as well as on $\mathcal{D}$'s information processing and adaptation of its beliefs and view on $\mathcal{A}$ in the dynamic scenario. To make the importance clear, a completely ignorant agent would never subjectively violate secrecy as it would ignore its violation of secrecy. Likewise underestimating as well as overestimating the capabilities of an $\mathcal{A}$ can lead to a violation of secrecy. In particular a secrecy preserving agent should satisfy the following properties: (P1) The agent is aware of the information communicated to other agents and the meta-information conveyed by it actions, (P2) The agent simulates the reasoning of other agents, (P3) The agent considers possible meta-inferences from conspicuous behavior such as (a) selfcontradiction, (b) refusal, (P4) For all possible states and perceptions the agent does not perform any action that leads to secrecy violation, (P5) The agent only weakens secrets if it is unavoidable due to information coming from third parties and only as much as necessary. The properties (P1) and (P5) are related to the belief change component of $\mathcal{D}$, (P2) and (P3) to the way $\mathcal{D}$ models $\mathcal{A}$ and (P4) to the means-end reasoning behavior of $\mathcal{D}$.

2 Formalizing Secrecy Preserving Agents

We loosely base our agent model on the well known beliefs, desires, intentions (BDI) architecture. However, in our epistemic view of agency, the agent's epistemic state contains a representation of its current desires and intentions which guides its behavior. The functional component of a BDI agent consists of a change operation of the epistemic

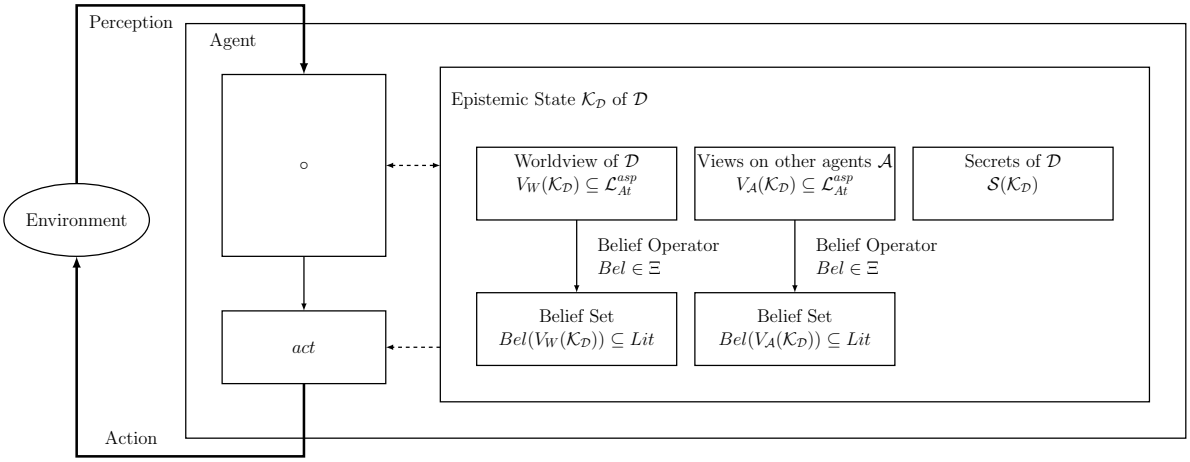


Figure 1: Epistemic Agent Model

state and an action function, executing the next action as determined by the current epistemic state. Our agent model is illustrated in Figure 2.

To define secrets, the information to be kept secret has to be defined. Also, the agent from which the information shall be kept secret has to be defined and lastly the strength of the secret has to be expressed. We make use of the belief operators to express the strength of a secret. A *secret* is a tuple $(\Phi, Bel, \mathcal{A})$ which consists of a formula $\Phi \in Lit$, a belief operator $Bel \in \Xi$ and an agent identifier $\mathcal{A} \in \mathfrak{A}$. The set of secrets of agent $\mathcal{D}$ is denoted by $\mathcal{S}(\mathcal{K}_\mathcal{D})$. Assigning a more credulous belief operator to a secret leads to a stronger protection of secret information. That is, if $\mathcal{D}$ reveals some information, a credulous attacker might infer some secret information while a skeptical one with the same revealed information might not. In the former case the defender should not have revealed the information. Formally, considering two secrets $(\Phi, Bel, \mathcal{A})$ and $(\Phi, Bel', \mathcal{A})$, the former is stronger than the latter iff Bel is more credulous than Bel'. The definition of secrets satisfies (S1), (S2) and (P2).

We consider communicating agents whose actions as well as perceptions τ are speech acts from a set of speech acts $\langle A_s, \{A_{r_1}, \ldots, A_{r_n}\}, type, \Phi\rangle$ specifying the source $A_s \in \mathfrak{A}$, the receivers $A_{r_1} \in \mathfrak{A}$ to $A_{r_n} \in \mathfrak{A}$, the type $type$ and the informational content $\Phi \in Lit$. For each perception $p \in Per$ an agent cycle results in a new epistemic state determined by $\mathcal{K}_\mathcal{D} \circ_\mathcal{D} p \circ_\mathcal{D} \mathsf{act}_\mathcal{D}(\mathcal{K}_\mathcal{D} \circ_\mathcal{D} p)$. Our intuitive idea of secrecy preservation expresses that we want to assure that the secrecy preserving agent always maintains an epistemic state in which it believes that no other agent believes in something that it wants to keep secret. More exactly, it also distinguishes between secrets towards different agents and what it means to it that the information is kept secret. The term "*always maintains*" means that for all possible scenarios of communication the agent acts such that a safe epistemic state is maintained. Let $\mathcal{D} = (\mathcal{K}_\mathcal{D}, (\circ_\mathcal{D}, \mathsf{act}_\mathcal{D}))$ be an agent and Per a set of perceptions. An

epistemic state $\mathcal{K}_D$ is *safe* iff $\Phi \notin Bel(V_A(\mathcal{K}_\mathcal{D}))$ for all $(\Phi, Bel, \mathcal{A}) \in \mathcal{S}(\mathcal{K}_\mathcal{D})$. We call $\mathcal{D}$ *secrecy preserving* with respect to $\Lambda^0_\mathcal{D}$ and Per if and only if for all $\mathcal{K}_\mathcal{D} \in \Omega_{\mathsf{act},\circ}(\Lambda^0_\mathcal{D}, Per)$ it holds that $\mathcal{K}_\mathcal{D}$ is safe. The definition of a secrecy preserving agent satisfies (P4).

3 Results and Discussion

We presented a theoretical, conceptional and practical account of secrecy from the subjective view of an autonomous epistemic agent in [4]. We formulated properties of secrecy and secrecy preservation and developed a framework for and ASP-based instance satisfying them. We have shown in [3] that other many aspects of notions of secrecy such as [1] and [2] can be captured by our underlying model. Moreover, we developed a framework for the implementation of knowledge based agents in which we implemented our general Model and the ASP instance to run experiments. To the best of our knowledge no subjective account of agent based secrecy nor a concrete model or implementation of a secrecy preserving agent system has been presented so far. We see our model and implementation as a good basis for the further theoretical investigation as well as the implementation of secrecy preserving agents. It opens a plethora of possibilities for further investigation. In current work we investigate further properties of secrecy in this model and the relation to other approaches, and integrate advanced deliberation and means-end reasoning techniques in our model and implementation.

References

[1] Joachim Biskup. Usability confinement of server reactions: Maintaining inference-proof client views by controlled interaction execution. In *Databases in Networked Information Systems*, volume 5999 of *LNCS*, pages 80–106. Springer, 2010.

[2] Joseph Y. Halpern and Kevin R. O'Neill. Secrecy in multiagent systems. *ACM Transactions on Information and System Security*, 12:5:1–5:47, October 2008.

[3] Patrick Krümpelmann and Gabriele Kern-Isberner. On agent-based epistemic secrecy. In Riccardo Rossi and Stefan Woltran, editors, *Proceedings of the 14th International Workshop on Non-Monotonic Reasoning (NMR'12)*, 2012.

[4] Patrick Krümpelmann and Gabriele Kern-Isberner. Secrecy preserving BDI Agents based on Answerset Programming. In *Proceedings of the 11th German Conference on Multi-Agent System Technologies (MATES'13)*, volume to appear of *Lecture Notes in Computer Science*. Springer, 2013.

[5] Leendert van der Torre. Logics for security and privacy. In *Data and Applications Security and Privacy XXVI*, volume 7371 of *LNCS*, pages 1–7. Springer, 2012.

Inference-Proof Views Based on Fragmentation and Encryption

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 analysis of an already existing approach which aims at achieving confidentiality by the combined usage of vertical fragmentation and encryption. This analysis relies on a logic-oriented modelling and allows reasoning about the inference-proofness of this approach. Moreover, the already begun research on the creation of inference-proof materialized views by refusing certain values of an original database instance has been continued.

My research during the last year dealt with the analysis of an existing approach to achieve confidentiality by means of the combined usage of vertical fragmentation and encryption presented in [2, 10]. The goal of this approach is both to reduce storage and processing costs by storing data on external servers and to comply with confidentiality requirements – in particular with privacy concerns – in spite of outsourcing data.

For that purpose, a client's database relation is losslessly decomposed into (at least) two vertical fragments each of which is maintained by a different semi-honest server. These fragments are constructed by splitting sensitive data into harmless parts, either by breaking the associations between the values allocated to specific sets of attributes of the database relation or by separating encrypted attribute values from the cryptographic keys needed to decrypt these values. Moreover, the servers storing the data are (postulated to be) mutually isolated from each other and each attacker is assumed to have access to at most one of these servers. Consequently, due to splitting, each attacker (identified with a server) only has accesses to non-sensitive data. In contrast, an authorized user (identified with the client) is able to query all fragments of the losslessly decomposed database relation and can therefore still reconstruct all original data. This scenario is visualized in Figure 1.

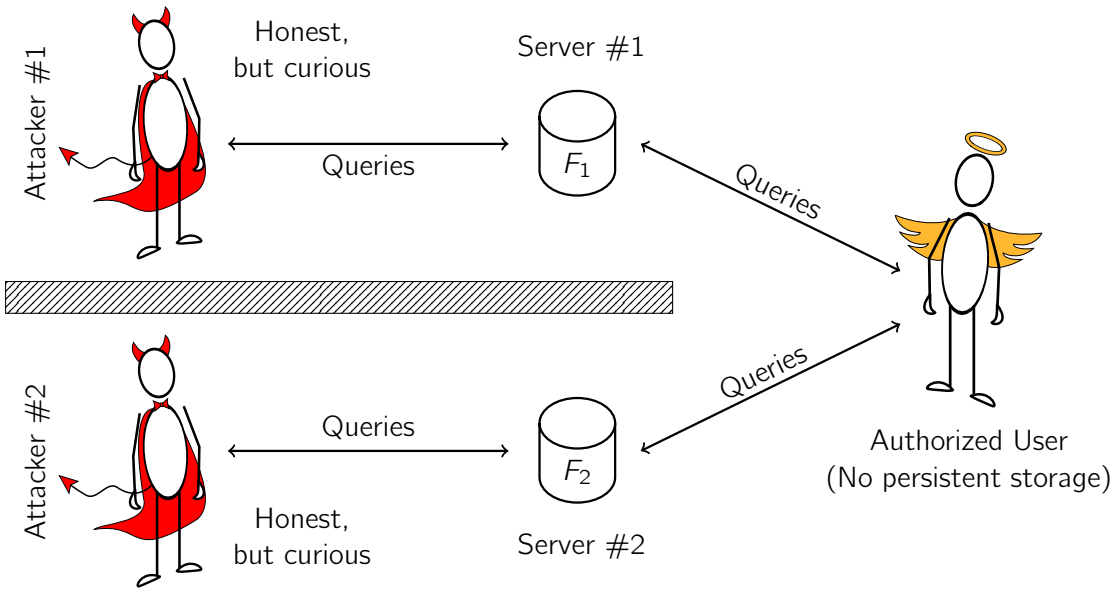


Figure 1: Scenario of the fragmentation approach considered

At first glance, two semi-honest servers seem to "keep the secrets" declared in a confidentiality policy. However, a second thought raises some doubts on the actual achievements: though each server only stores data that is non-sensitive per se, an attacker might still be able to infer sensitive information by exploiting his a priori knowledge [5]. In particular, this a priori knowledge might comprise semantic constraints to be satisfied by the relation being decomposed (see [1]) and individual fact data stemming from the "outside world". In the context of this question, the following problems are analyzed:

- Given a fragmentation, identify conditions on an attacker's a priori knowledge to provably disable this attacker to infer sensitive information.
- Given an attacker's a priori knowledge, determine a fragmentation such that an attacker cannot infer sensitive information.

The results of this analysis are published in [6] and were presented at DBSec 2013 conference in Newark, New Jersey, USA. The analysis is based on a logic-oriented modelling of the fragmentation approach considered within the more general framework of Controlled Interaction Execution (CIE), which is surveyed in [5]. The main contributions of this analysis can be summarized as follows:

- The fragmentation approach considered is formalized.
- A logic-oriented modelling of that approach is provided.
- Sufficient conditions to achieve confidentiality are exhibited.
- A method to compute a suitable fragmentation is proposed.

These results extend the previous work [7] in which a more simple approach to fragmentation proposed in [9] – splitting a relational instance into one externally stored part and one locally-held part without resorting to encryption – is formally analyzed to be inference-proof. In particular, the previous work is extended by a more detailed formal modelling of fragmentation including encryption of values, a more expressive class of sentences representing an attacker's a priori knowledge and a method to compute an inference-proof fragmentation.

From the point of view of inference control, the approaches to vertical fragmentation considered in [6, 7] can also be seen as a mechanism to establish inference control efficiently. For each user querying the database an alternative instance – which is called materialized view – is created by splitting the original database instance according to the approach to vertical fragmentation considered. Then, this user is only allowed to query exactly that fragment which does not contain any information enabling him to infer information to be kept secret.

Consequently, such a generation of inference-proof materialized views allows a safe and as well efficient handling of queries because queries can be answered safely without employing costly mechanisms of (dynamic) inference control based on theorem proving [4]. Each query is simply answered without any monitoring by directly posing it to the materialized view generated for the pertinent user. Of course, the generation of such a materialized view might be of high computational complexity, but can be seen as preprocessing as it is usually generated before granting a user the right to query the database instance.

As proposed in last year's technical report, the creation of inference-proof materialized views by refusing a subset of values of an original database instance has also been a topic of research. This research was motivated by [8], which aims at generating inference-proof materialized views with the help of lies. Last year's thoughts of generating an inference-proof materialized view by refusing a subset of complete tuples of the original instance have now been refined by the insight that often the refusal of some components of certain tuples should be sufficient to preserve confidentiality.

A similar idea is followed in [3], in which the values of single components of tuples are replaced by null-values to preserve confidentiality. But this approach significantly differs from my conceptual ideas since a user is neither assumed to have a priori knowledge, nor is he explicitly aware of which values are refused. Instead, the original database relation may already contain null-values and the user is assumed to be not able to distinguish between these original null-values and those null-values used to refuse values.

A user's explicit awareness of the tuple components whose values are refused might enable this user to infer knowledge to be kept secret by employing so-called meta-inferences. Deduction of knowledge is called a meta-inference if a user succeeds in exploiting an explicit refusal notification to infer sensitive information by simulating the behavior of the algorithm used for generating the inference-proof views [5]. Of course, the elimination of those meta-inferences must be analyzed with scrutiny.

References

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

[2] Gagan Aggarwal, Mayank Bawa, Prasanna Ganesan, Hector Garcia-Molina, Krishnaram Kenthapadi, Rajeev Motwani, Utkarsh Srivastava, Dilys Thomas, and Ying Xu. Two can keep a secret: A distributed architecture for secure database services. In *2nd Biennial Conference on Innovative Data Systems Research, CIDR 2005*, pages 186–199, 2005.

[3] Leopoldo E. Bertossi and Lechen Li. Achieving data privacy through secrecy views and null-based virtual updates. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):987–1000, 2013.

[4] Joachim Biskup. *Security in Computing Systems – Challenges, Approaches and Solutions*. Springer, Heidelberg, 2009.

[5] 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.

[6] 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.

[7] Joachim Biskup, Marcel Preuß, and Lena Wiese. On the Inference-Proofness of Database Fragmentation Satisfying Confidentiality Constraints. In Xuejia Lai, Jianying Zhou, and Hui Li, editors, *14th Information Security Conference, ISC 2011*, volume 7001 of *LNCS*, pages 246–261, Heidelberg, 2011. Springer.

[8] 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.

[9] Valentina Ciriani, Sabrina De Capitani di Vimercati, Sara Foresti, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. Keep a few: Outsourcing data while maintaining confidentiality. In Michael Backes and Peng Ning, editors, *14th European Symposium on Research in Computer Security, ESORICS 2009*, volume 5789 of *LNCS*, pages 440–455, Heidelberg, 2009. Springer.

[10] Vignesh Ganapathy, Dilys Thomas, Tomás Feder, Hector Garcia-Molina, and Rajeev Motwani. Distributing data for secure database services. *Transactions on Data Privacy*, 5(1):253–272, 2012.

Specifying and Enforcing Confidentiality Requirements in Agent Interactions

Cornelia Tadros
Lehrstuhl für Informationssysteme und Sicherheit
Technische Universität Dortmund
cornelia.tadros@tu-dortmund.de

In multiagent systems, several agents (i.e., autonomous computing systems) share information for the purpose of achieving a joint goal, e.g., a sale contract, arrangement of a meeting etc. Whilst sharing of information is a necessary means for the cooperation among the agents it is subject to obligations or interests of individual agents to hide sensitive information from others. As one focus of my research, I proposed a formalization of an abstract multiagent system for the specification of confidentiality requirements and this way related these requirements to others in the literature of information flow control. The formalized system is an abstraction of the multiagent systems I studied in my previous work. Another focus of my research is how an agent can enforce its confidentiality interests in face of the uncertainty about the means of other agents opposing these interests.

Like in my last year's research, my work focused on a scenario of an isolated interaction between two agents, a requesting agent $\mathcal{A}$ and a reacting agent $\mathcal{D}$, outlined by Fig. 1. While my previous work in [3–5] proposed $\mathcal{D}$'s controlled processing of various belief change operations, for finalizing my PhD thesis [8], my aim was to propose a uniform model of security engineering for these work. To this aim, I introduced an abstract model of the multiagent system in the focused scenario as seen from the perspective of security engineering. In this scenario, agent $\mathcal{A}$ iteratively sends requests for offering or demanding information to the other agent $\mathcal{D}$ who reacts to these requests in return. As a unified communication language for various types of interaction, the two agents use propositional logic $\mathcal{L}_{PL}$ over a countably infinite alphabet. The reacting agent $\mathcal{D}$ provides an

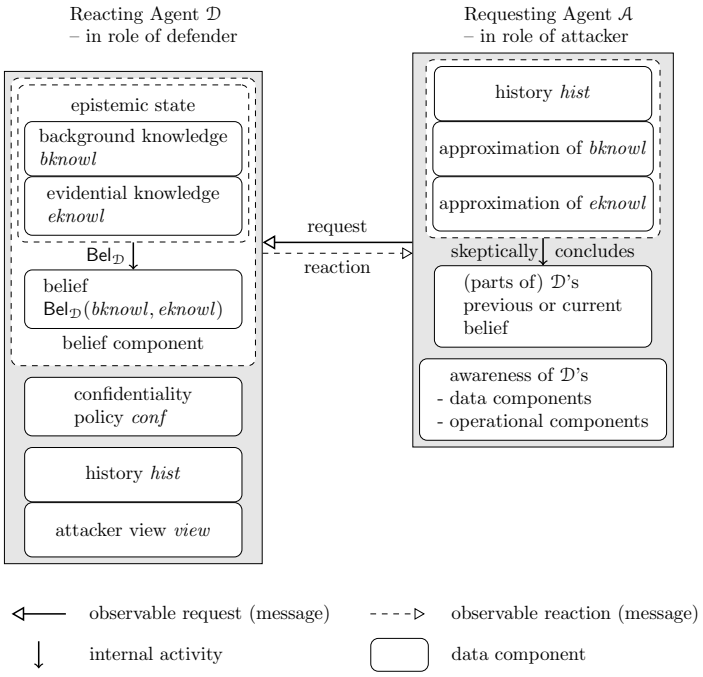


Figure 1: Interaction between two agents under confidentiality requirements

interaction interface to $\mathcal{A}$, defining valid request messages by an abstract set $\mathcal{Req}$, while its internal functionality is abstracted to the control function $\mathsf{cexec} : \mathcal{St} \times \mathcal{Req} \to \mathcal{St}$. The control function separates $\mathcal{D}$'s interaction interface and its internal functionality and this way controls $\mathcal{D}$'s interactions with $\mathcal{A}$ to enforce confidentiality.
Complementing its functionality, $\mathcal{D}$'s data components are defined in the set $\mathcal{St}$ of abstract states each of which is comprised of a belief component, a confidentiality policy and data components for the simulation of $\mathcal{A}$. The belief component defines the agent's belief $\mathsf{Bel}_{\mathcal{D}}(bknowl, eknowl) \subseteq \mathcal{L}_{PL}$ about its environment as the result of reasoning from its *background knowledge bknowl* and its *evidential knowledge eknowl*. The evidential knowledge is specific information $\mathcal{D}$ gathered about the case it reasons about. In contrast, its background knowledge represents its general expertise used in its process of reasoning.
Based on the outlined abstract agent model, extending the previous work in [2], I formalized this abstract agent scenario as a system $\mathcal{R}^{\mathsf{cexec}}_{\mathcal{D},\mathcal{A}}$ in the Runs & Systems framework [6, 7] for the specification of confidentiality requirements. To this end, the functionality of the agent $\mathcal{D}$ defending its confidentiality interests must be specified and its capabilities as a defender as well as the capabilities of the other agent $\mathcal{A}$, opposing $\mathcal{D}$'s interests, must be postulated for security engineering. The specification and postulates are reflected by the model $\mathcal{R}^{\mathsf{cexec}}_{\mathcal{D},\mathcal{A}}$. Then, this model is used to define the semantics of *policy-based secrecy* of a *possibility policy* by means of which confidentiality requirements of $\mathcal{D}$ may be declared. Whereas a possibility policy is a declaration in terms of

the abstract system model, the agent may declare its *confidentiality policy* in the base language $\mathcal{L}_{PL}$ to specify which parts of its current or previous belief are confidential. Like in [2], but with the system model $\mathcal{R}^{\mathsf{cexec}}_{\mathcal{D},\mathcal{A}}$ allowing for general types of interactions between the agents unlike in [2], I show that confidentiality policies may be translated to possibility policies such that the requirement of the confidentiality policy is enforced by policy-based secrecy. This way, I further relate the requirement of the confidentiality policy to others in the research of information flow control [7] and inference control [1] in computer security.

Whereas the model $\mathcal{R}^{\mathsf{cexec}}_{\mathcal{D},\mathcal{A}}$ formalizes $\mathcal{A}$'s postulated reasoning capabilities from the perspective of security engineering, the previous works in [3–5] focus on the simulation of the postulated reasoning. The essential differences between postulated and simulated reasoning are as follows. On the one hand, the postulated reasoning is about the multiagent system specified by abstract states and executions on these states and is used by the security engineer for the specification of confidentiality requirements and their verification. On the other hand, the simulated reasoning is usually formalized in a known logic or one of its fragments and is used by the defending agent $\mathcal{D}$ for computing the attacker's inferences at runtime. Consequently, the simulation of $\mathcal{A}$ needs a representation *view* of $\mathcal{A}$'s information about the system $\mathcal{R}^{\mathsf{cexec}}_{\mathcal{D},\mathcal{A}}$ in the logic used for simulation; an associated operator $\mathsf{skeptical}$ for reasoning in the logic; and finally an account for the further postulates about $\mathcal{A}$ as formalized in system $\mathcal{R}^{\mathsf{cexec}}_{\mathcal{D},\mathcal{A}}$.
Within my PhD thesis, I treat the systems previously studied in [3–5] as an implementation of the abstract model $\mathcal{R}^{\mathsf{cexec}}_{\mathcal{D},\mathcal{A}}$. Beyond the previous studies in [3], I elaborate properties of $\mathcal{D}$'s simulation of $\mathcal{A}$ and prove that they suffice to ensure confidentiality by preserving an invariant $S \notin \mathsf{skeptical}(\mathit{view})$ for all S relevant for the protection of confidential belief.

So far, in the mentioned work in [3–5, 8], we investigate confidentiality preservation under the security engineering postulate that the defender $\mathcal{D}$ is able to determine all the a priori knowledge $\mathcal{A}$ has for reasoning about $\mathcal{D}$'s belief. However, especially in a scenario with several agents, agent $\mathcal{D}$ might be uncertain about $\mathcal{A}$'s state and thus about $\mathcal{A}$'s conclusions about $\mathcal{D}$'s belief in that state. Moreover, due to lack of control over the dissemination of information in the multiagent system, agent $\mathcal{D}$ might be confronted with a violation of its confidentiality interests.
Under these additional challenges, agent $\mathcal{D}$ might not be able to enforce its confidentiality policy by maintaining the invariant $S \notin \mathsf{skeptical}(\mathit{view})$ as in [3–5, 8]. Instead, the agent must be able to reason about possible violations of its confidentiality policy under uncertainty about the states of the other agents. Being autonomous, the agent will utilize this reasoning ability to plan its actions in compliance with its confidentiality policy while during planning the agent reasonably deals with conflicting desires such as cooperative information sharing.
In our ongoing work we address these challenges. Our long-term goal is to enable agent $\mathcal{D}$ to reason on secrecy constraints to the end of decision-making. A secrecy constraint

(ϕ, Bel) intuitively expresses $\mathcal{D}$'s desire that $\mathcal{A}$ should not believe ϕ by means of belief operator Bel. In particular, the operator Bel may not only be skeptical reasoning, whereas $\mathcal{A}$ was postulated to be a skeptical reasoner in previous work.
The end of secrecy reasoning is to classify $\mathcal{D}$'s actions into allowed actions and prohibited actions, capturing its secrecy constraints. In this context, we focus on a scenario of an interaction between $\mathcal{D}$ and $\mathcal{A}$ where $\mathcal{D}$ can choose to execute an *inform-action* to its ends. An inform-action tells $\mathcal{A}$ that some sentence is true; but doing this, $\mathcal{D}$ should comply with its secrecy constraints. The secrecy reasoning of $\mathcal{D}$ should follow the underlying intention to protect sensitive information against other agents in a best possible way – even under the uncertainty inherent in $\mathcal{D}$'s assumption about another agent and even under consideration of violations of secrecy constraints.

References

[1] Joachim Biskup. Inference-usability confinement by maintaining inference-proof views of an information system. *IJCSE*, 7(1):17–37, 2012.

[2] Joachim Biskup and Cornelia Tadros. Policy-based secrecy in the Runs & Systems framework and controlled query evaluation. In Isao Echizen, Noboru Kunihiro, and Ryôichi Sasaki, editors, *Short Paper of IWSEC 2010*, pages 60–77. IPSJ, 2010.

[3] Joachim Biskup and Cornelia Tadros. Inference-proof view update transactions with minimal refusals. In Joaquin Garcia-Alfaro, Guillermo Navarro-Arribas, Nora Cuppens-Boulahia, and Sabrina De Capitani di Vimercati, editors, *DPM 2011/SETOP 2011*, volume 7122 of *LNCS*, pages 104–121. Springer, 2012.

[4] Joachim Biskup and Cornelia Tadros. Revising belief without revealing secrets. In Thomas Lukasiewicz and Attila Sali, editors, *FoIKS 2012*, volume 7153 of *LNCS*, pages 51–70. Springer, 2012.

[5] Joachim Biskup and Cornelia Tadros. Preserving confidentiality while reacting on iterated queries and belief revisions. *Annals of Mathematics and Artificial Intelligence*, pages 1–49, 2013.

[6] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.

[7] Joseph Y. Halpern and Kevin R. O'Neill. Secrecy in multiagent systems. *ACM Transactions on Information and System Security*, 12(1), 2008.

[8] Cornelia Tadros. *Belief Change Operations under Confidentiality Requirements in Multiagent Systems*. PhD thesis, Technische Universität Dortmund, 2013. Submitted.

Subproject B1
Analysis of Spectrometry Data with Restricted Resources

Sven Rahmann Jörg Ingo Baumbach

Automation of the peak extraction process for MCC/IMS measurements

Marianna D'Addario
SFB 876, Project TB1
Computer Science XI, TU Dortmund
marianna.daddario@tu-dortmund.de

The non-invasive technique of MCC/IMS (a coupling of a multi-capillary column (MCC) with an ion mobility spectrometer (IMS)) is nowadays used in a high-throughput context producing a huge amount of data in form of MCC/IMS measurements. Consequently, the processing of MCC/IMS measurements has to rise to new challenges, e.g. time and space reducing processing. As part of the project TB1, this report summarizes the efforts to include a peak candidate detection and a peak picking method in a fully automated peak extraction in order to lower the time consumption of processing MCC/IMS measurements.

1 Introduction: Automated peak extraction.

The MCC/IMS devices detect volatile organic compounds (VOCs) in the air or in exhaled breath. The data produced by these devices is visualized as a heat map, where the two dimensions are the inverse reduced mobility (t in Vs/cm^2) and the retention time (r in seconds). Every elevated area could indicate a measurable intensity of an organic compound and is called a peak. The process of finding all peaks of a measurement is here called *peak extraction*.

The MCC/IMS measurements are already used to identify patterns that signalize known diseases, e.g. lung cancer or diabetes [2,4,8,9]. But the whole peak extraction process is time consuming and involves human interaction, such as hand-picking approximate peak locations, assisted by a visualization of the data. In a high-throughput context, however, it is preferable to have robust methods for fully automated peak extraction.

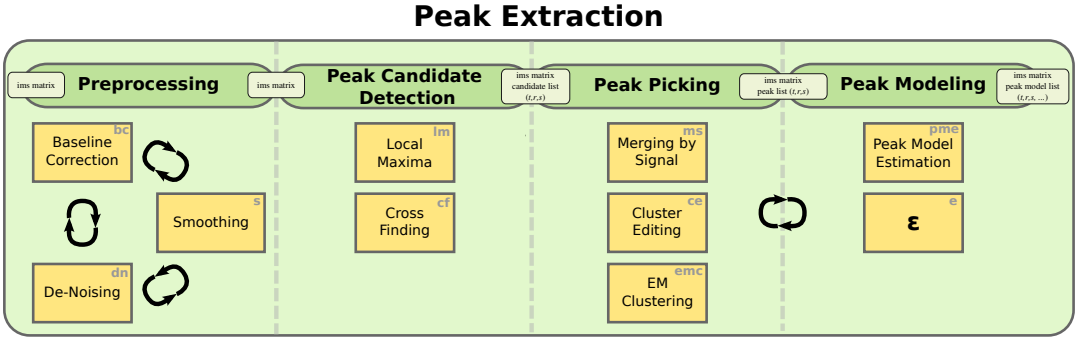


Figure 1: Four steps of the peak extraction. Each step can be implemented by different modules; yellow boxes containing an abbreviation for each module name.

Therefore an automation of the extraction process of an MCC/IMS measurement discovering and quantifying all present peaks seems necessary. The peak extraction process is divided into four steps (Figure 1).

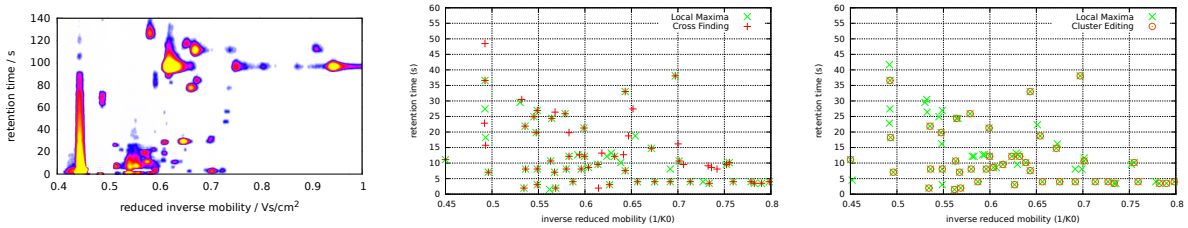
The focus of this technical report relies on the two implemented modules *Local Maxima* for the peak candidate detection step and the *Cluster Editing* for the peak picking step. Further, a measure is needed to determine the quality of the automatic peak extraction with respect to the manually one.

2 Finding the candidate peaks.

Peaks are points in a coordinate system of two dimensions, inverse reduced mobility t and retention time r (see Figure 2(a)). Every point at position (r, t) has a signal intensity $S_{r,t}$. A simple way to report peaks is to detect a candidate for every local intensity maximum with intensity at least M in a surrounding area. To report a point (r, t), we require (1.) that (r, t) is a local maximum in the sense the each of its eight direct neighbors has a lower or equal intensity than $S_{r,t}$ but of at least M, and (2.) that the contiguous area around (r, t) with signal intensity at least I is of sufficient size. In other words, we discard points where the surrounding high-intensity area size consists of too few points. The required number of points is controlled by a parameter $A \geq 9$. By the first condition, (r, t) and its eight neighbors always account for nine points; the parameter A can be used to impose stricter conditions. Figure 2(b) illustrates the difference between the modules *Cross Finding* and *Local Maxima*. In the work of Hauschild *et al.* [5] was shown that the automatic methods for this step can be used for a good separation into two groups of patients, like healthy and not healthy.

3 Picking the peaks.

The main idea of this step is to eliminate those peaks that are too close to another one avoiding multiple peaks produced by the same analyte. Bödeker [3] introduced a



(a) Preprocessed heat map produced by an MCC/IMS device.

(b) Local Maxima versus Cross Finding.

(c) The picking step by Cluster Editing for the candidates found by Local Maxima.

Figure 2: The two consecutive steps peak candidate detection (b) and peak picking (c).

minimum distance in retention time and in inverse reduced mobility such that two peaks exceeding those distances belong to distinct compounds. Hauschild *et al.* [5] suggested a constant $\Delta t =: 0.003\,\mathrm{Vs/cm^2}$ (inverse reduced mobility) and an affine-linear $\Delta r =: pr + c$ (retention time) for a peak at position (r, t), where $c =: 3\,\mathrm{s}$ and $p =: 0.1$.

The idea is to find clusters of peaks, from which a representative will be picked, by solving an instance of the weighted cluster editing problem [1, 7]: Let $G = (V, E)$ be a weighted, undirected graph without loops with a symmetric similarity weight function $w\colon \binom{V}{2} \to \mathbb{R}$, such that $E = \{\{u, v\} : w(u, v) \geq 0\}$. The graph can be modified by removing an existing edge or adding a non-existing $\{u, v\}$. For both modification a cost of $|w(u, v)|$ incurs. The total cost is then the sum of all modifications. The aim is to find a set of edge modifications with minimum cost such that the resulting graph consists of disjoints cliques (i.e., is transitive).

Every candidate peak is a vertex $u = (r_u, t_u)$. The similarity $w(u, v)$ between two vertices u, v depends on their distances on the r- and t-axis $(\Delta r, \Delta t)$. The distance measure is

$$d^2(u, v) =: \frac{1}{2}\left[\left(\frac{t_u - t_v}{\Delta t}\right)^2 + \left(\frac{r_u - r_v}{\Delta r}\right)^2\right]$$

and the similarity weight function (with a constant scaling factor b) is

$$w(u, v) =: \begin{cases} 2^{b(1-d^2(u,v))} - 1 & \text{if } d^2(u, v) \leq 1, \\ 1 - d^2(u, v) & \text{otherwise.} \end{cases}$$

The range for $w(u, v)$ is therefore $[-\infty, 2^b - 1]$. The weighted cluster editing problem is solved with the *yoshiko 2.0* software (`http://www.cwi.nl/research/planet-lisa`). Figure 2(c) shows the chosen representatives for the found peak candidates.

4 Measuring the quality.

To conform to the standards of B&S Analytik the recommended `.xls` format [6] was used to store the list of the found peaks. To measure the quality of an automati-

cally picked peak list this will be compared to a list of peaks manually picked by an expert. Peaks detected by both methods, manual and automatic, within one measurement are true positives (TP). Manually annotated peaks that are not detected by the automatic procedure are false negatives (FN) and automatically detected peaks not found in the manual annotation are false positives (FP). We compute the sensitivity SENS $=:$ TP/(TP + FN) and the positive predictive value PPV $=:$ TP/(TP + FP). Their geometric mean $G =: \sqrt{\text{SENS} \cdot \text{PPV}}$ summarizes both measures. Further, the Jaccard index between two peak lists is $J :=$ TP / (FN + TP + FP) $\in [0, 1]$. From this, we derive a distance measure $d := 1/J - 1 \in [0, \infty]$. The distance and geometric mean can be calculated separately for each combination of methods and measurement.

References

[1] S. Böcker, S. Briesemeister, and G. W. Klau. Exact algorithms for cluster editing: Evaluation and experiments. *Algorithmica*, 60(2):316–334, 2011.

[2] B. Bödeker, W. Vautz, and J. I. Baumbach. Peak comparison in MCC/IMS-data – searching for potential biomarkers in human breath data. *International Journal for Ion Mobility Spectrometry*, 11(1-4):89–93, 2008.

[3] B. Bödeker, W. Vautz, and J. I. Baumbach. Peak finding and referencing in MCC/IMS-data. *International Journal for Ion Mobility Spectrometry*, 11(1):83–87, 2008.

[4] A. Bunkowski, B. Bödeker, S. Bader, M. Westhoff, P. Litterst, and J. I. Baumbach. MCC/IMS signals in human breath related to sarcoidosis – results of a feasibility study using an automated peak finding procedure. *Journal of Breath Research*, 3(4):046001, 2009.

[5] A. C. Hauschild, D. Kopczynski, M. D'Addario, J. I. Baumbach, S. Rahmann, and J. Baumbach. Peak detection method evaluation for ion mobility spectrometry by using machine learning approaches. *Metabolites*, 3(2):277–293, 2013.

[6] S. Maddula, K. Rupp, and J. I. Baumbach. Recommendation for an upgrade to the standard format in order to cross-link the GC/MSD and the MCC/IMS data. *International Journal for Ion Mobility Spectrometry*, 15(2):79–81, 2012.

[7] S. Rahmann, T. Wittkop, J. Baumbach, M. Martin, A. Truss, and S. Böcker. Exact and heuristic algorithms for weighted cluster editing. In *Computational Systems Bioinformatics Conference*, volume 6, pages 391–401, 2007.

[8] M. Westhoff, P. Litterst, L. Freitag, W. Urfer, S. Bader, and J.I. Baumbach. Ion mobility spectrometry for the detection of volatile organic compounds in exhaled breath of patients with lung cancer: results of a pilot study. *Thorax*, 64(9):744–748, 2009.

[9] M. Westhoff, P. Litterst, S. Maddula, B. Bödeker, S. Rahmann, A. N. Davies, and J. I. Baumbach. Differentiation of chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control group by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, 13(3-4):131–139, 2010.

Massively parallel read mapping with GPGPUs

Johannes Köster
Genominformatik
Institut für Humangenetik, Medizinische Fakultät
Universität Duisburg-Essen
johannes.koester@uni-due.de

This work presents a novel read mapping algorithm that exploits the capabilities of modern GPGPUs. Thereby, a novel q-gram index with low memory footprint is introduced.

The read mapping problem occurs when sequencing genomes of organisms with next-generation sequencing technologies. The genome is fragmented into small pieces, which are sequenced by the sequencing machine. The output consists of hundreds of millions of small (e.g. 100 nucleotides) DNA sequences, called *reads*, whose origins in the donor genome are unknown. The read mapping problem is then to find the origin of each read in a known reference genome (e.g. the human reference genome). Calculating the full alignment of each read against the reference genome needs quadratic time per read and is thus infeasible for large genomes and current sequencing depth. Hence, state of the art read mappers mostly use either a q-gram index [4] or a Burrows-Wheeler Transformation (BWT) [1] to speed up the process at the cost of accuracy.

This work presents an approach to exploit the massive parallelism of GPGPUs for read mapping. While read mapping is inherently parallel, above data structures both are problematic on GPGPU architectures: q-gram indices need extensive amounts of memory, BWT based algorithms defeat data-parallelism by branching and recursions.

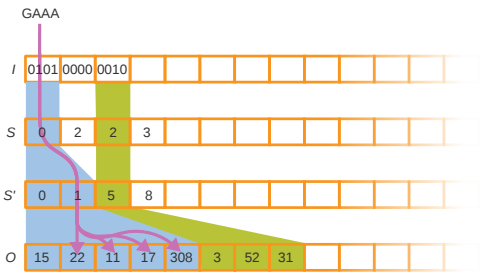


Figure 1: The q-group index

1 The q-group index

Here, a q-gram is a sequence of DNA nucleotides (represented with the four letters A, C, G, T) of length q. Commonly, each letter is encoded with two bits (e.g. A=00, C=01, G=10, T=11), such that a q-gram can be represented as a bit-vector. Therefore, a 16-gram fits into a 32-bit integer. A q-gram index is build over a given text T, and allows to retrieve for any q-gram its occurences in the text. A traditional q-gram index consists of two arrays. The first contains one entry for every possible q-gram, pointing to a starting position in the second array which contains the occurences of the q-grams in a given text. An interval in the occurence array defined by two subsequent entries g amd $g+1$ in the first array shall contain all occurences of the q-gram represented by g. Therefore the size of a q-gram index is $\mathcal{O}(2^{2q} + |T|)$ which is dominated by the exponential size of the first array.

In the following we present the key idea of the q-group index introduced in this work (also see Figure 1). We assign to each q-gram g a q-group $\lfloor g/w \rfloor$. The w should be chosen to match the size of the machine words encoding the q-grams, i.e. $w = 2q$. Then, we introduce a new top-level array, I, that contains a bit-vector for each possible q-group. The i-th bit of the bit-vector is one iff q-gram g with $g \bmod w = i$ occurs in the text. A second array S contains a starting position for each q-group which points to the third array. The third array S' is a sequence of starting positions in the fourth array O which contains the occurences of each q-gram in the text. The occurences and starting positions are aligned in a way such that the occurences of the j-th q-gram g of the q-group G in the text can be retreived as the interval

$$O[S'[S[G] + j]] \ldots O[S'[S[G] + j + 1]$$

Hence, the occurences can be retrieved with a constant amount of operations, similar to a traditional q-gram index. However, the size of the q-group index grows significantly smaller in q, i.e.

$$\mathcal{O}(2^{2q}/q + 2|T|).$$

Figure 2 shows the asymptotic size of the q-group index compared to a traditional q-gram index depending on $|T|$ and q.

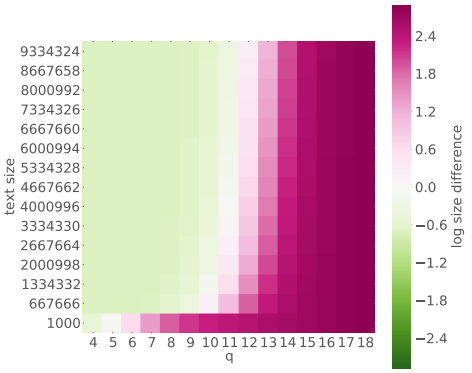


Figure 2: Difference between logarithmic size of q-gram index and worst case logarithmic size of q-group index for different q and text lengths.

2 Mapping Strategy

The implemented read mapping algorithm consists of three major steps: filtration, validation and postprocessing. Filtration and validation are implemented as data-parallel OpenCL-kernels, while I/O and postprocessing happens in parallel on the CPU. Figure 3 illustrates the workflow. First, reads are loaded from disk until a buffer is filled. Then the first reference sequence is loaded (i.e. the first chromosome of the reference genome to map against). The q-grams in the reference are filtered with a q-gram index of the buffered reads. Remaining q-grams, together with the occurences retreived from the index denote potential mapping positions of the reads. These are validated in the next step, using a bit-parallel alignment algorithm [3]. Finally, validated mappings are post-processed, e.g. removing duplicates, calculating mapping probabilities and writing them to disk in the SAM/BAM format [2].

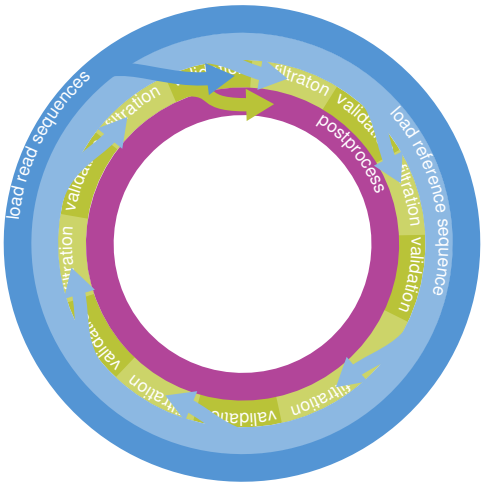


Figure 3: Algorithm workflow.

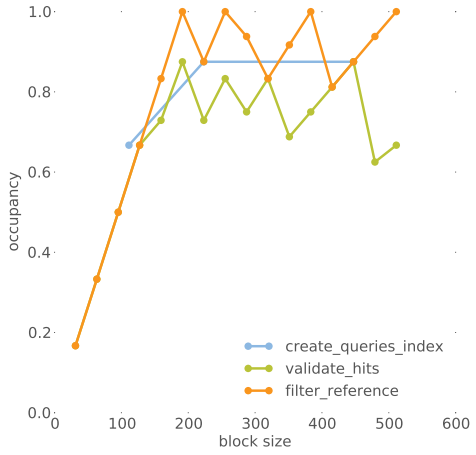


Figure 4: Occupancy of implemented OpenCL kernels depending on the thread block size.

All levels are designed to hide latency when possible. For example the next bunch of reads or reference sequences is loaded from disk, while the first buffer is processed on the GPU. Further, the GPU kernels hide latency by running other thread groups while one is reading from memory. The latter ability can be measured by the occupancy, that can be seen as the fraction of active processing cores on the GPU. Hence an occupancy of 1 means that all cores are busy 100% of the computation time. A small occupance is not desirable as it implies unused resources. Figure 4 shows the observed patterns of occupancy for our OpenCL kernels. As can be seen, all kernels reach very high occupancy rates. The q-group index based filtration step (filter_reference) even reaches an occupancy of 100%.

References

[1] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[2] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[3] Gene Myers. A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM*, 46(3):395–415, May 1999.

[4] David Weese, Manuel Holtgrewe, and Knut Reinert. RazerS 3: Faster, fully sensitive read mapping. *Bioinformatics*, 28(20):2592–2599, October 2012.

Various novel Methods for Analysis of Spectrometry Data

Dominik Kopczynski
Lehrstuhl für Algorithm Engineering
Technische Universität Dortmund
Dominik.Kopczynski@tu-dortmund.de

An ion mobility (IM) spectrometer coupled with a multi-capillary column (MCC) measures volatile organic compounds (VOCs) in the air or in exhaled breath. Each peak in an MCC/IM measurement represents a certain compound, which may be a known or unknown. For clustering and classification of measurements, the raw data matrix must be reduced to a set of peaks. Each peak is described by its coordinates (retention time in the MCC and reduced inverse ion mobility) and shape (signal intensity, further shape parameters). This fundamental step is referred to as *peak extraction*. We present new approaches for particular challenges in peak extraction.

1 Introduction

While ion mobility (IM) spectrometry (IMS) is an established technology to detect volatile organic compounds (VOCs) in the air or exhaled breath, the more recent combination with multi-capillary columns (MCCs) has opened new applications in biotechnology and medicine [1]. The analytes, metabolites present within exhaled breath, are pre-separated using the MCC, analogously to gas chromatography (GC) before mass spectrometry (MS). An MCC/IM measurement is described by a two-dimensional matrix with retention time R and reduced inverse mobility T as dimensions where the rows are spectra and the columns are chromatograms. Several processing steps are necessary to extract information of peaks within a complete measurement (IMSC). We developed a framework called "Peax" including a set of exchangable processes. Two particular processes *De-Noising* and *Cross Finding* are explained in detail in the following.

2 De-Noising

In general it is observable that spectra from different spectrometry devices show a high signal-to-noise ratio. This results in well formed peaks where noise nearly does not disturb the peaks shapes. Peaks with a small signal suffer from a higher perturbation. Thus we need a filter that differenciates between true signals and noise. We present a novel method utilizing the EM algorithm to determine whether the intensity $S(r,t)$ at coordinates (r,t) belongs to a peak region or can be solely explained by background noise. To be independent in application, we developed a generic approach that does not need any information about models of the desired signal. Our results show that the background noise is normally distributed. As a preliminary for computation in the expectation step, we have to set up an auxiliary matrix A where the average of a certain data point considering an arbitrary margin ρ is stored. Let

$$A_{r,t} \coloneqq \frac{1}{(2\rho+1)^2} \cdot \sum_{r'=r-\rho}^{r+\rho} \sum_{t'=t-\rho}^{t+\rho} S_{r',t'}$$

for all $r \in R, t \in T$. Typically $\rho = 4$ is an appropriate value for our data. The algorithm needs start parameters, the ordinary mean μ_{N} and standard deviation σ_{N} of the measurement signals are appropriate for this. At this point we introduce two additional models: the first model shall describe the desired data. Because the data is biased with a higher weight towards low signal data points, we use the inverse Gaussian distribution with mean parameter μ_{D} and shape λ_{D} to achieve a skewed model. The second additional model is the uniform distribution, since some data points can not be described by the first two models. As typical for the EM algorithm, all models have a weight $\omega_{\mathrm{N}}, \omega_{\mathrm{D}}, \omega_{\mathrm{B}}$ (noise, data, background). Let $\omega_{\mathrm{N}} = |\{a \in A | a \leqslant \mu_{\mathrm{N}} + 3 \cdot \sigma_{\mathrm{N}}\}|/|A|$. To obtain both remaining weights, we assume that 99.9% of the remaining weight belongs to the data, thus $\omega_{\mathrm{D}} = (1-\omega_{\mathrm{N}}) \cdot 0.999$, and $\omega_{\mathrm{B}} = (1-\omega_{\mathrm{N}}) \cdot 0.001$, respectively. To compute the start parameters for the data model, let $\mu_{\mathrm{D}} = \sum_{r,t} A'_{r,t}/|A'|$ and $\lambda_{\mathrm{D}} = (\sum_{r,t}(1/A'_{r,t} - 1/\mu_{\mathrm{D}}))^{-1}$ where $A' = \{a \in A | a \geqslant \mu_{\mathrm{N}} + 3 \cdot \sigma_{\mathrm{N}}\}$.

Having computed the matrix A and all start parameters, the algorithm begins with the expectation step. To estimate the probability $W_{r,t,\mathrm{N}}$ that a data point is created by noise, we use the following equation

$$W_{r,t,\mathrm{N}} \coloneqq \frac{\omega_{\mathrm{N}} \cdot \mathrm{G}(A_{r,t}, \mu_{\mathrm{N}}, \sigma_{\mathrm{N}})}{\omega_{\mathrm{N}} \cdot \mathrm{G}(A_{r,t}, \mu_{\mathrm{N}}, \sigma_{\mathrm{N}}) + \omega_{\mathrm{D}} \cdot \mathrm{IG}(A_{r,t}, \mu_{\mathrm{D}}, \lambda_{\mathrm{D}}) + \frac{\omega_{\mathrm{B}}}{|R| \cdot |T|}} \tag{1}$$

for all $r \in R$, $t \in T$. The probabilities for data $W_{r,t,\mathrm{D}}$ and $W_{r,t,\mathrm{B}}$ are computed analogously. In the maximization step we estimate parameters for both models normal distribution (μ_{N} and σ_{N}) and inverse Gaussian distribution (μ_{D} and λ_{D}) using maximum likelihood estimators. Let

$$\omega_{\mathrm{N}}^* \coloneqq \frac{\sum_{r,t} W_{i,t,\mathrm{N}}}{|R| \cdot |T|}, \tag{2}$$

$$\mu_{\mathrm{N}}^* \coloneqq \frac{\sum_{r,t} W_{r,t,\mathrm{N}} \cdot A_{r,t}}{\sum_{r,t} W_{r,t,\mathrm{N}}}, \tag{3}$$

$$\sigma_{\mathrm{N}}^* \coloneqq \frac{\sum_{r,t} W_{r,t,\mathrm{N}} \cdot (A_{r,t} - \mu_{\mathrm{N}}^*)^2}{\sum_{r,t} W_{r,t,\mathrm{N}}}, \tag{4}$$

$$\mu_{\mathrm{D}}^* \coloneqq \frac{\sum_{r,t} W_{r,t,\mathrm{D}} \cdot A_{r,t}}{\sum_{r,t} W_{r,t,\mathrm{D}}}, \tag{5}$$

$$\lambda_{\mathrm{N}}^* \coloneqq \frac{\sum_{r,t} W_{r,t,\mathrm{D}}}{\sum_{r,t} W_{r,t,\mathrm{D}} \cdot (1/A_{r,t} - 1/\mu_{\mathrm{D}}^*)} \tag{6}$$

for all $r \in R, t \in T$. The weight of the data model ω_{D} is computed analogously to Equation (2). Equation (1) in expectation step and equations (2), (3), (4), (5) and (6) are repeated until μ_{N} and σ_{N} converge. After the hidden variables are estimated, finally the new signal matrix must be applied, let $S'_{r,t} \coloneqq S_{r,t} \cdot (1 - W_{r,t,\mathrm{N}})$ for all $r \in R, t \in T$.

3 Cross Finding

Cross Finding is a method that finds peaks within a preprocessed IMSC and outputs a peaklist described by their parameters. It is an improvement of the work by Fong *et al.* [2].

At the beginning two auxiliary matrices D^{R} and D^{T} have to be constructed, both with the same dimensions $|R| \times |T|$. Matrix D^{T} stores partial derivatives with respect to reduced inverse mobility, let $D^{\mathrm{T}}_{r,t} \coloneqq S_{r,t+1} - S_{r,t}$; analogously D^{R} stores derivatives of chromatograms. Since the analysis for both matrices is equal, we concentrate only on how D^{T} is analysed.

In each derived spectrum (for fixed retention time r), we mark downward zero crossings; these are indices t with $D^{\mathrm{T}}_{r,t-1} \geqslant 0$ and $D^{\mathrm{T}}_{r,t} < 0$. The resulting indices t are called *active positions* for retention time r.

While we scan through the spectra, we maintain two data structures. The first one is an *active set* containing lists of active positions connected across several spectra. The second one is a *finalized set*, where lists from the active sets are moved when they have been processed. Initially both sets are empty.

We want to connect active positions between consecutive retention times, i.e., we want to find active positions for spectrum $r+1$ corresponding to active positions in spectrum r (see Figure 1(left)). To find the correspondences, we use a variant of global alignment between the two sorted integer lists A and A^+ containing the active positions. The score of aligning $A[i]$ to $A^+[j]$ depends on the distance between $A[i]$ and $A^+[j]$, let $\mathtt{score}(i,j) \coloneqq (1 + |A[i] - A^+[j]|)^{-1} \in [0,1]$ be the score function. To prevent that two positions with a high distance are aligned, we introduce a gap score $\gamma = 0.2$. We can solve the alignment problem very efficiently by only considering (i,j) with $|A[i] - A^+[j]| \leqslant 5$,

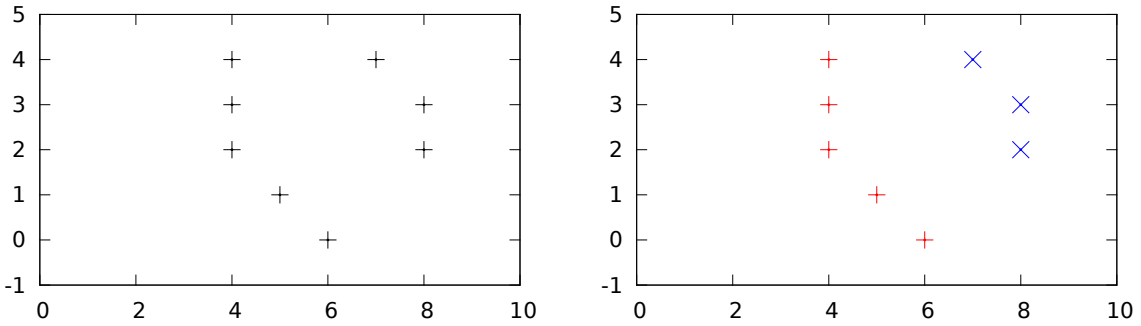


Figure 1: Cross finding: Active positions (marking potential peak maxima) are initially unaligned (left) and then connected by alignment across spectra (right; shown as red + and blue x). The same procedure is repeated over all chromatograms giving horizontal bands instead of vertical bands. Intersecting the results from both dimensions results in peak candidates.

since we dont need to consider pairs of $A[i]$ and $A^+[j]$ with distance larger than 5 index units.

The alignment approach leads to three possible scenarios between the aligned position pairs: (1) If $A^+[j]$ is not aligned to any $A[i]$, it becomes a "new" active position, and a new list (containing only $A^+[j]$) is inserted into the active set. (2) If $A^+[j]$ is aligned to some $A[i]$, the corresponding list containing $A[i]$ is already in the active set and extended by $A[j]$. (3) Each $A[i]$ that is not aligned to any $A^+[j]$ finalizes its corresponding active list, and the list is moved into the finalized set.

The process will be finished by finalizing each remaining list in the active set. We obtain several position lists pointing out consecutive maxima throughout each spectrum; see Figure 1(right).

Matrix D^R will be processed analogously. By taking the intersection of these consecutive maxima lines found from both matrices (which can be visualized as crosses; hence the name "Cross Finding"), we achieve the corresponding positions of the peaks local maxima. Each reported point whose signal exceeds an arbitrary threshold is a candidate for a peak location.

References

[1] R. Koczulla, A. Hattesohl, S. Schmid, B. Bödeker, S. Maddula, and J. I. Baumbach, *International Journal for Ion Mobility Spectrometry* **14**, 177 (2011).

[2] S. S. Fong, P. Rearden, C. Kanchagar, C. Sassetti, J. Trevejo, and R. G. Brereton, *Analytical Chemistry* **83**, 1537 (2011).

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

Peter Marwedel Heinrich Müller Alexander Zybin

Automatic Energy-Aware Design Space Exploration for GPGPUs

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

This report presents a novel approach for automatically determining the most power- or energy-efficient Graphics Processing Units (GPUs) with respect to given parallel computation problems. Different objectives and constraints can be considered. For the use-case of a mobile detection of biological nano structures the real-time constraint has to be met and the energy-consumption has to be minimized to maximize the number of possible measurements.

1 Introduction

General Purpose computing on Graphics Processing Unis (GPGPU) becomes more and more important and energy consumption has to be considered, especially but not only on mobile devices. The focus of the work lies on the GPU hardware with the objectives speed and energy consumption, which is important to transfer the application, the detection of biological, virus like nano structures ([5], [4]), from a desktop to a mobile device. Therefore the platform design was investigated using a design space exploration to explore which hardware requirements fit the problem best. For this step the hardware was simulated with the cycle accurate simulator GPUGPU-Sim [1], giving control over many hardware parameters, e.g. clock speed, global/local memory size for different operations or work group sizes. To determine the power- or energy-consumption GPUWattch [3] was used. A genetic-algorithm approach was used to find the parameters for a real-time and energy-saving hardware platform. As a result the best possible hardware can be selected without expensive testing of actual GPUs.

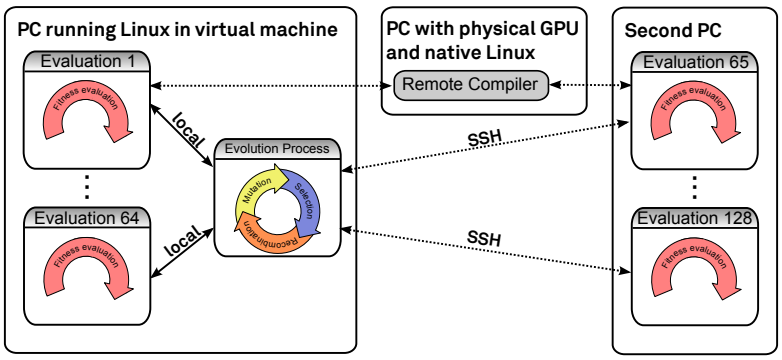


Figure 1: Example instance of the evolutionary process. One PC is running the evolution and part of the evaluations, a second PC is running the rest of the evaluations and a third PC is used for remote compilation.

Table 1: Examined architectures

Architecture	SMs	Core Clock (MHz)	DRAM Clock (MHz)	Number of Registers
GF-108	1-2	700-810	1600-1800	16k-32k
GF-106	3-4	590-790	1800-4000	16k-32k
GF-104	6-7	650-675	3400-3600	32k-64k
GF-100	11-15	610-780	3200-4000	32k-64k

2 Automatic Energy-Aware Design Space Exploration

A framework for an automatic energy-aware design space exploration of GPGPUs has been developed and published [6]. For a given GPGPU program or a set of programs the best suited GPU can be discovered. A parameter set of GPGPU-Sim parameters is used to encode the specification of each single GPU. The overall design space consists of all possible parameter sets. As the search space contains also unreasonable GPUs some effort has been made to easily configure the design space to the research objective. The design space can be set up as fine or coarse granulated as wanted, down to a list of every single GPU configuration that should be considered. An example of a design space can be seen in table 1, four different NVIDIA architectures from GF-108 up to GF-100, each with a distinct number of streaming multiprocessor cores (SMs).

An evolutionary algorithm was used to optimize the power- or energy-consumption. An example of the overall evolution process is shown in figure 1. The evaluation process - selection, recombination and mutation - is running on one PC in a number or threads. The evaluation step can be distributed to different threads, either locally or remotely. All necessary files are distributed automatically to the evaluation threads. Each of the threads is then running the GPGPU program using GPGPU-Sim to calculate the number of needed cycles and GPUWattch to calculate the average power consumption.

Table 2: Simulated power consumption of the fastest GPU compared to the optimized results

Benchmark	Average Power Cons. GTX 480	Average Power Cons. Proposed	Power Saved
HotSpot	49.7 Watt	11.5 Watt	77%
k-Means	26.0 Watt	7.5 Watt	71%
k-Nearest Neighbors	91.6 Watt	13.4 Watt	85%
Needleman-Wunsch	50.1 Watt	41.3 Watt	18%

3 Results

For the evaluation different GPU architectures (cf. table 1) have been examined. A subset of the OpenCL programs in the Rodinia benchmark set [2] was used to demonstrate the generality of the proposed method: A physics simulation named HotSpot, two common data mining algorithms (k-Means and k-Nearest Neighbors (k-NN)) and the bioinformatics algorithm by Needleman-Wunsch were selected. Each benchmark was evaluated individually. The power savings are listed in table 2. The power savings are varying from 18 to 85 percent. An evaluation of the energy-consumption of the GPUs has also be done but has not yet been published. For the energy consumption savings from 45 percent (Needleman-Wunsch) up to 93 percent (k-Means) were measured.

4 Future Work

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. Tradeoffs in energy efficiency between local and remote processing can be examined. The framework can be used in the projects second phase to identify energy efficient GPUs in a sensor network with a centralized high performance computation (HPC) system. To model the whole process e.g. the energy consumption of the network communication of the system needs to be considered. This could be examined in collaboration with the project A4.

5 Conclusion

A framework for an automatic energy-aware design space exploration of GPGPUs has been developed. This framework can be used in a wide variety of applications. For example to identify the most energy-efficient GPUs in huge high performance computation

systems or in small cyber physical 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. Also the framework can be used for multi-objective optimization, for example to identify the tradeoff between detection rate and energy consumption in a hardware/software co-design. 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.

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] Shuai Che, M. Boyer, Jiayuan Meng, D. Tarjan, J.W. Sheaffer, Sang-Ha Lee, and K. Skadron. Rodinia: A benchmark suite for heterogeneous computing. In *Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on*, pages 44–54, 2009.

[3] Jingwen Leng, Tayler Hetherington, Ahmed ElTantawy, Syed Gilani, Nam Sung Kim, Tor M Aamodt, and Vijay Janapa Reddi. Gpuwattch: Enabling energy optimizations in gpgpus. *International Symposium on Computer Architecture*, 2013.

[4] Pascal Libuschewski, Dominic Siedhoff, Constantin Timm, Andrej Gelenberg, and Frank Weichert. Fuzzy-enhanced, real-time capable detection of biological viruses using a portable biosensor. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSIGNALS)*, February 2013. Publication.

[5] Pascal Libuschewski, Dominic Siedhoff, Constantin Timm, and Frank Weichert. Mobile detektion viraler pathogene durch echtzeitfähige gpgpu-fuzzy-segmentierung. *Bildverarbeitung für die Medizin 2013*, pages 326–331, March 2013.

[6] Pascal Libuschewski, Dominic Siedhoff, and Frank Weichert. Energy-aware design space exploration for gpgpus. *Computer Science - Research and Development (CSRD)*, 2013. DOI: 10.1007/s00450-013-0237-5.

An OpenMP-Inspired Approach For Heterogeneous Embedded MPSoC

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

Future embedded systems will increasingly make use of heterogeneous multiprocessor systems-on-chip (MPSoC). Using these systems efficiently requires methods and tools that support the extraction and implementation of parallelism that is typically found in embedded applications.

Most infrastructures available today that support the coordination and execution of parallelized software, however, are designed according to the requirements of high performance computing. This wastes optimization opportunities essential for resource-constrained embedded systems.

In this report, we describe our approach of a new infrastructure inspired by the OpenMP API in order to support the definition and implementations of pipeline parallelism, which is commonly found in complex embedded applications. Our toolflow combines our approach with advanced parallelization extraction and mapping tools which enables a more efficient approach to exploit parallelism for typical embedded applications on heterogeneous MPSoCs.

Introduction and Motivation

The requirements to create software for current embedded multi-processor system-on-chips (MPSoCs) pose a number of additional challenges compared to high-performance computing (HPC) applications. On the one hand, MPSoCs feature a large variety in the characteristics of the available cores, such as clock frequency, processor pipeline structures, or memory hierarchies. While current systems provide a number of processor cores, most other resources of embedded MPSoC systems, such as energy, memory size, and communication bandwidth are strictly limited.

On the other hand, embedded software exhibits different structures of control flow and data dependencies compared to high-performance computing applications. Whereas many HPC applications are executable efficiently by implementing coarse-grained task- and finer-grained data-level parallelism, a significant percentage of embedded applications exhibit more complex data dependencies. This can lead to a significant loss of efficiency when applying typical HPC parallelization tools to embedded environments.

Embedded software often operates on streams of data, e.g., in media or signal processing applications. For these, in most cases, a pipeline structure can be identified. Considering the versatility of embedded hardware and software, an optimizing approach to implement parallel software for embedded MPSoCs should combine the best of both worlds. It should be possible to make use of coarse-grained task- and fine-grained data-level parallelism while exploiting the additional opportunities pipeline parallelism offers.

To take advantage of heterogeneous MPSoCs, applications need to be adapted. In general, two approaches are applied to solve this issue efficiently. The first one is to create new parallel software which utilizes the different processing units of the target heterogeneous platform best but requires a large amount of time to develop, is very error-prone and also generates code which is hard to understand. The second one is to parallelize already existing sequentially written legacy code, thus code is reused which significantly reduces the development time required. In the following, we present a new approach which supports the application designer during parallelization of already existing sequentially written applications for embedded heterogeneous target platforms.

Framework

This section briefly describes our API which is inspired by OpenMP [8] to support heterogeneous architectures. They enable the application designer to parallelize applications for resource-restricted multiprocessor platforms. Most of th embedded software is written in sequential C. Thus, only OpenMP for C [7] is considered[1]. OpenMP uses `#pragma` directives as annotations which have to be added by the developer to implement parallel regions. Programs parallelized with OpenMP use the *fork-join model*, where a running task spawns new subtasks (*fork*) and waits until all subtasks have terminated (*join*). Thus, implicit data and control-flow synchronization points are added at the end of each parallel region.

OpenMP supports task-level parallelism(`omp parallel sections`) and data-level parallelism which can be implemented by the `omp parallel for` pragma. Many embedded applications profit more from pipeline parallelism which is currently not supported by OpenMP. Therefore, we introduce a new directive to implement pipeline parallelism(`parallel pipeline for`). In addition, we added clauses to enable an manual

[1] OpenMP for Fortran is not considered in this report.

task to processor set mapping (*processors*) which is mandatory to utilize heterogeneous systems. Further, a static mapping of iterations to tasks is introduced by two new clauses (*chunks*, *iterations*).

We implemented a tool which is able to process these annotations as a source-to-source transformation. The tool is based on the MACC framework [10] using the ICD-C [1] compiler framework's abstract syntax tree and versatile program and data flow analyses. Thus, communication between tasks can be implemented automatically. Furthermore, an automatic parallelizer PAXES [3–5] focusing on resource-restricted embedded systems is also available and tightly coupled with our tool. PAXES generates annotated source code which is then processed by the source-to-source transformation tool.

In addition, we modified an existing toolchain to compile the transformed source code into executables for the target platforms. These platforms are composed of ARM processors connected through a shared memory simulated with MPARM [2] or Virtualizer [12] simulators. With the last one, we created a heterogeneous system comparable to ARM's big.LITTLE [9] platform. As operating system RTEMS [11] with R^2G [6] is used. To implement necessary communication between parallel tasks, we developed a lightweight runtime library where shared memory communication with software FIFOs is supported.

Conclusion and Future Work

This report presented our infrastructure for heterogeneous resource-restricted embedded systems. The proposed API enables applications written in C and annotated easily to benefit from modern heterogeneous embedded systems. With the proposed framework, developers are now able to express pipeline parallelism without rewriting the entire application. Loops can now be balanced offline according to the performance characteristics of the available processing units. In addition, data dependencies are detected automatically by our framework which relieves the developer from the burden of manually specifying dependencies and synchronization points.

Since data dependencies between tasks can be extracted fully automatically, we plan to optimize communication between tasks. In addition, we also intend to optimize the scheduling of synchronization points. Special platform-dependent communication methods like, e.g., hardware FIFOs, could be used to further increase the performance of the application.

References

[1] ICD-C Compiler framework. `http://es.icd.de/`, February 2013.

[2] Luca Benini, Davide Bertozzi, Alessandro Bogliolo, Francesco Menichelli, and Mauro Olivieri. MPARM: Exploring the Multi-Processor SoC Design Space with SystemC. *Journal of VLSI Signal Processing Systems*, 41(2):169–182, September 2005.

[3] Daniel Cordes, Michael Engel, Olaf Neugebauer, and Peter Marwedel. Automatic extraction of multi-objective aware parallelism for heterogeneous mpsocs. In *Proceedings of the Sixth International Workshop on Multi-/Many-core Computing Systems (MuCoCoS 2013)*, MuCoCoS 2013, Edinburgh, Scotland, UK, sep 2013.

[4] Daniel Cordes, Michael Engel, Olaf Neugebauer, and Peter Marwedel. Automatic extraction of pipeline parallelism for embedded heterogeneous multi-core platforms. In *Proceedings of the Sixteenth International Conference on Compilers, Architectures, and Synthesis for Embedded Systems (CASES 2013)*, CASES 2013, Montreal, Canada, oct 2013.

[5] Daniel Cordes, Michael Engel, Olaf Neugebauer, and Peter Marwedel. Automatic extraction of task-level parallelism for heterogeneous mpsocs. In *Proceedings of the Fourth International Workshop on Parallel Software Tools and Tool Infrastructures (PSTI 2013)*, PSTI 2013, Lyon, France, oct 2013. (accepted for publication).

[6] Andreas Heinig. R^2G: Supporting POSIX like semantics in a distributed RTEMS system. Technical Report Technical Report 836, TU Dortmund, Faculty of Computer Science 12, December 2010.

[7] ISO. The ANSI C standard (C99). Technical Report WG14 N1124, ISO/IEC, 1999.

[8] OpenMP. The OpenMP API specification for parallel programming. `http://www.openmp.org/`, December 2012.

[9] P. Greenhalgh, ARM. Big.LITTLE Processing with ARM Cortex-A15 & Cortex-A7. `http://www.arm.com/files/downloads/big.LITTLE\_Final.pdf`, February 2013.

[10] Robert Pyka, Felipe Klein, Peter Marwedel, and Stylianos Mamagkakis. Versatile System-level Memory-aware Platform Description Approach for embedded MPSoCs. In *Proc. of LCTES10*, Stockholm, Sweden, 2010.

[11] RTEMS. RTEMS Operating System | Real-Time and Real Free, June 2013.

[12] Synopsys. Synopsis Virtualizer, May 2013.

Results of Automating the Analysis of PAMONO Biosensor Data

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

1 Introduction

This technical report summarizes the so-called *Synthesize/Optimize* framework that has been developed in project B2 with the goal of automating quantitative microscopy experiments. It accelerates typical workflows arising in the evaluation and development of prototypical sensor technology by providing generic methods for detecting and classifying small objects in noisy data. These methods automatically adapt to changing sensor setups (physical sensor parameters) with minimum user interaction. As a practical use-case of the framework we employ the PAMONO biosensor for indirect detection of nano-objects (e.g. biological viruses) via optical microscopy, that has also been developed in project B2.

2 Synthesize/Optimize Framework

Figure 1 depicts the *Synthesize/Optimize* framework to be summarized now. Details are in an upcoming paper. The description given here concretizes the framework for the analysis of PAMONO data, however concretizations for other microscopy tasks are conceivable as well. The input of the framework are real sensor images which are firstly used to drive the generation of synthetic images with known ground truth segmentation and classification (cf. in section 3). The parameters of an object detector and a learning algorithm are determined by optimizing analysis results for the ground truth data. These parameters and the learned model are then applied to the real sensor input data to analyze it.

The employed object detector is a significantly extended version of the GPGPU-based object detector from [6]. Extensions include image stabilization [4] and intensity-curvature based features [11] which are used in the subsequent learning stage.

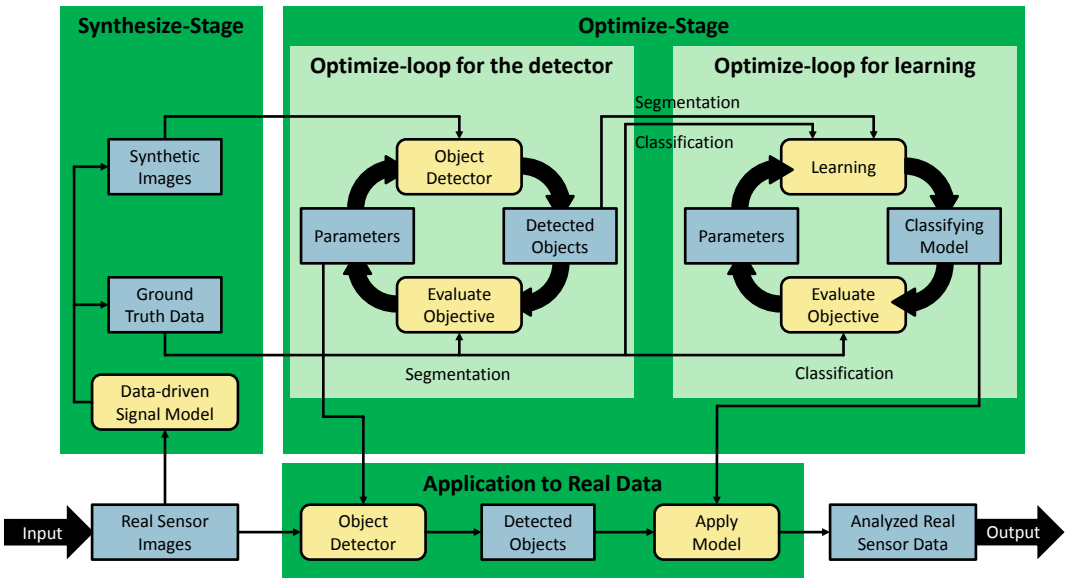


Figure 1: *Synthesize/Optimize* framework.

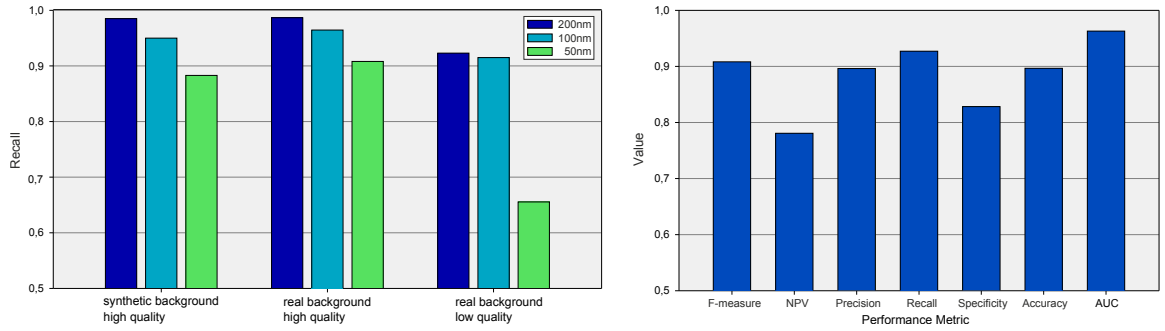
The learning stage integrates an advanced version of [10] under the *Synthesize/Optimize* framework. Its goal is tuning the hyperparameters of a learning algorithm, which is subsequently used to learn an optimal model to finally classify the real input data. Examples of examined learning algorithms and corresponding hyperparameters are the number k of regarded neighbors in k-Nearest Neighbors [9], the number of features available for splitting at each node in a Random Forest [2] and the regularization parameter and choice of kernel function in a Support Vector Machine [7]. Furthermore, the synthetic data abundance is exploited for obtaining unbiased estimates of performance [8]. To avoid problems arising from imbalanced class distributions, the training set is balanced before being input to the learning algorithm [5].

3 Sensor Model

The Synthesis stage of the *Synthesize/Optimize* framework relies on a suitable sensor model. Image acquisition of the PAMONO biosensor can be modeled as

$$I(x, y, t) = N(\mu, \sigma) + (B \cdot V)(x + x_j(t), y + y_j(t), t). \tag{1}$$

Here $I(\cdot, \cdot, t)$ is the image as delivered by the sensor at time t, and x and y are its spatial coordinates. The N term is a coordinate-independent draw from a random Gaussian distribution with mean μ and standard deviation σ. It models the additive noise incurred on the sensor. The second term consists of a background signal B and the virus signal V, with B being significantly larger in terms of amplitudes than V and V being a multiplicative modulation term that assumes value 1 for nonvirus areas and low amplitude variations about 1 in virus areas. The functions $x_j(t)$ and $y_j(t)$ denote the x and y components of a time-dependent jitter on the spatial coordinates which results from small displacements



(a) (b)

Figure 2: (a) Detector recall. (b) Classification performances.

of the sensor surface with respect to the camera. The background signal B consists of a surface component S and an artifacts component A:

$$B(x, y, t) = S(x, y) \cdot A(x, y, t) \tag{2}$$

The surface component S is modeled as being constant over time, containing an image of the empty sensor surface and constant diffraction patterns. The S component contributes the high amplitudes to I. The artifacts component A models nonvirus nanoobjects and any other effects impeding virus detection. A has a similarly small magnitude as the virus signal V and is also modeled as a multiplicative modulation.

The sensor model is used to generate two types of synthetic PAMONO data: One can take a fully synthetic approach by ignoring the A term in equation (2), taking any sensor image as S and applying synthetic jitter and noise. Such datasets primarily test the detection part because classification does not have to separate any artifacts from real viruses. The second approach uses a virus-less real background measurement as B, hence incoporating a real artifacts signal. For both types of signals, virus templates are uniformly distributed in the x, y, t volume and inserted into I as according to equation (1).

4 Results

Validation of the results attained by the framework focuses on its two central components: the detection and the learning part, cf. Figure 1.

Figure 2(a) summarizes results for the detector, displaying the dependence of detector recall on the background type (fully synthetic as well as high and low quality real background, cf. section 3) and particles size. Lower particle size corresponds to a lower SNR in the data. As can be seen, detector recall is always above 0.9, except for the 50 nm particles. Recall and particle size exhibit an nonlinear relationship. The impact of background quality (which is equivalent to the quality of the employed gold plate in the sensor setup) is low, except for the 50 nm particles. For larger particle sizes, the detector handles deterioration of the background well. For 100 nm particles on low quality real

background a recall of 0.92 is attained. The given values are recall estimates, attained for analyzing unseen synthetic testing sets with ground truth available. Manual validation confirmed these estimates to be within 5% of the recall attained on the actual input, where no ground truth was available. The estimates are more exact for higher quality real background.

Figure 2(b) summarizes classification results attained by an optimized Random Forest, averaged over different datasets with 100nm particles on low quality real background. Note that due to page limits only Random Forest results are presented; an investigation determining Random Forest as the best classifier in the PAMONO context can be found in an upcoming paper. The objective function for optimizing the hyperparameters of the Random Forest was f-measure. As shown in the figure, performance indices are generally above 0.89, except for negative predictive value and specificity, indicating a deficiency in classifying negative examples. This is assumed to be related to balancing the skewed class distribution: The negative class is often underrepresented in the data.

5 Future Work

Future work will investigate using Markov Random fields [1] as a combined approach for 1) modeling the sensor (generating synthetic data) and 2) classifying pixels during detection. Furthermore, a denoising procedure in the style of the spatial BM3D algorithm [3] is to be developed to exploit the spatiotemporal characteristics of PAMONO data.

References

[1] Simon A. Barker. *Image Segmentation using Markov Random Field Models*. PhD thesis, University of Cambridge, 1998.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16:2080–2095, 2007.

[4] E. Haber and J. Modersitzki. Numerical methods for volume preserving image registration. *Inverse Problems*, 20:1621–1638, 2004.

[5] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21:1263–1284, 2009.

[6] P. Libuschewski, C. Timm, D. Siedhoff, F. Weichert, H. Müller, and P. Marwedel. Improving nanoobject detection in optical biosensor data. In *Proceedings of the 15th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2011.

[7] K-R. Müller, S. Mika, G. Ratsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12:181–201, 2001.

[8] T. Scheffer and R. Herbrich. Ubiased assessment of learning algorithms. In *Proc. IJCAI*, 1997.

[9] G. Shakhnarovich, T. Darrell, and P. Indyk, editors. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[10] D. Siedhoff, F. Weichert, P. Libuschewski, and C. Timm. Detection and classification of nano-objects in biosensor data. In *Microscopic Image Analysis with Applications in Biology (MIAAB)*, 2011.

[11] 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

Expression Languages for Data Stream Analysis

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

Conditions are a key component of every programming language. A high throughput can only be achieved, if the evaluation of conditions have a low impact on the overall runtime of the program execution. This is especially important for real-time data stream analysis. if conditions are defined on data streams, every data-item needs to be evaluated and even a small reduction of runtime per evaluation can lead to a better performance. A common use case is to check wether or not a set of attributes are available. If the number of attributes is high, this will have an high impact on the performance regardless of the rest of the analysis.
The *streams*-framework [1] includes a specialized Expression Language (EL) to state conditions in the XML-container definition, but also as a convenient way of implementing meta-processors on exchangeable contexts. The EL is used for all conditioned processors, for example the If-processor and Skip-processor. Therefore, the EL will be used in almost every analysis.
A new version of the EL was implemented in version 0.9.10. The improved performance was shown for the first time in the DEBS Challenge [3]. The central idea of the new version of the EL is that only the values will change over time, but not the structure of the expression (Fig. 1). Therefore, the expression could be compiled at start-up time. Until version 0.9.10, every expression was parsed and interpreted for every evaluation of the expression.

The key element of expressions are contexts. A context $C_i \in \mathcal{C}$ maps a string $k \in \mathcal{K} \subset \mathcal{S}$ to an object $o \in O$, $C_i(k) = o$.
The *streams*-framework defines 4 types of contexts. The container context C_c holds the global states of the application. The process context C_{p_i} holds all the information and

states concerning process P_i. Every processor of the process P_i can read and write to this context. With the use of services, information from other processes or even from remote containers and data sources could be stored in the process context and then be used in expressions. The data context C_{d_i} represents the current data item d_i and the static context C_{v_i} represent static values, e.g. $C_{v_i}(k) = v_i, \forall k \in \mathcal{K}$.
Expressions are functions defined on contexts $C_i \in \mathcal{C}$. They garantuee that the returned value of a context for given k or of an operation on subexpressions over the contexts are of type T. A type is a set of elements of the same type, $t \in T$. The special element *null* (empty set) is part of every type T.

$$E^T(\mathcal{C}\big|\mathcal{K}) = \begin{cases} C(k) & C(k) \in T, C \in \mathcal{C}, k \in \mathcal{K} \\ \otimes^{T',T}(\mathcal{C}\big|\mathcal{K}) & \otimes^{T',T}(\mathcal{C}\big|\mathcal{K}) \in T \\ null & otherwise \end{cases} . \quad (1)$$

The simplest non static expression is the simple context access : $E^O(\mathcal{C} = \{C\}|k) = C(k)$.
An operation on expressions is defined as:

$$\otimes^{T_1,T_2}(\mathcal{C}\big|E_1, E_2, \mathcal{K}_1, \mathcal{K}_2) = E_1^{T_1}(\mathcal{C}\big|\mathcal{K}_1) \otimes^{T_2} E_2^{T_1}(\mathcal{C}\big|\mathcal{K}_1). \quad (2)$$

Operations on expressions are expression themselves. They can be combined with other expressions to define arbitrary expression trees (Fig. 1). The evaluation of expressions starts at the "root"-expression. The contexts $\mathcal{C}$ will be passed "in-order" from the top and accessed at the leafs of the expression tree only.
Examples of operations are the greater-condition and the AND-condition:

$$>^{\mathbb{R},\mathbb{B}}(C) = \begin{cases} false & E_1^{\mathbb{R}}(C) == null \vee E_2^{\mathbb{R}}(C) == null \\ E_1^{\mathbb{R}}(C) > E_2^{\mathbb{R}}(C) & \text{otherwise} \end{cases}$$

$$\wedge^{\mathbb{B},\mathbb{B}}(C) = \begin{cases} false & E_1^{\mathbb{B}}(C) == null \vee E_2^{\mathbb{B}}(C) == null \\ E_1^{\mathbb{B}}(C) \wedge E_2^{\mathbb{B}}(C) & \text{otherwise} \end{cases}$$

In the *streams*-framework contexts are evaluated by $\%\{\langle contextname\rangle.\langle key\rangle\}$. Figure 1 shows an example for an AND-condition and the resulting expression tree. The example also shows the usage of brackets and single quotation marks. Brackets define a subexpression with the same return type as the included expressions. Single quotation marks are necessary to define the return type of the expression as string (see (2)). If static contexts are used, e.g. $\%\{data.k_3\} ==$ 'test', the single quotation marks are not mandatory for the expression.
Table 1 shows all the implemented expressions in the *streams*-framework. Until now, only operators with return type Boolean are implemented. Nevertheless, it would be possible to define arithmetic operations like $+, -, ^*$ with return type Double or even operations with other return types.

$$(\%\{data.k_1\} > \%\{process.k_2\}) \text{ AND } (\text{'}\%\{data.k_3\}\text{'} == \text{'}\%\{process.k_4\}\text{'})$$

.

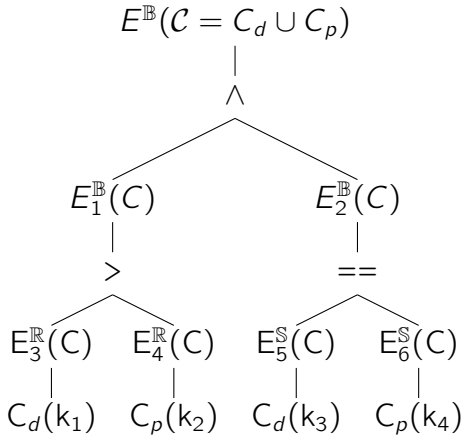


Figure 1: AND-condition and the resulting expression tree.

Expression	Return type	Data types	Description
==, !=	Boolean	Double, String, null	equal operators
>, >=, <, <=	Boolean	Double	inequality operators
@rx	Boolean	String	regular expression operator
AND, and, OR, or	Boolean	Boolean	logical operators
$\%\{context.key\}$	Double	Context	Double expression
$\text{'}\%\{context.key\}\text{'}$	String	Context	String expression
$[0-9]^*$	Double	String	Static Double expression
$.^*$	String	String	Static String expression

Table 1: Implemented operators in the *streams-framework version 0.9.13.*

With the given definition of expression trees, there could be cases where the evaluation of large expressions lead to a higher average runtime per evaluation. Figure 2 shows an example for an degenerated expression tree. There are 2 possibilities to reduce the expected average runtime per evaluation. This could be achieved by the reduction of the tree depth or by optimizing the the traversal order of the evaluation.

The depth of the tree could be reduced by set operators. A set operator is defined as:

$$\bigoplus^{T}(\mathcal{C}\big|\mathcal{K}, k^*) = \bigoplus_{k_i \in \mathcal{K}} C'(k_i) \otimes C^*(k^*). \tag{3}$$

Examples are a combination of a context access and the AND operator ($\bigoplus$ and $\otimes$) $\%\{context.\{regexp\}^* > 3\}$ and $\%\{context.\{regexp\} > 3\}$ or a combination of a context access and the OR operator ($\bigoplus$ and $\otimes$) $\%\{context.\{regexp\}+ > 3\}$.

The traversal order could either be optimized iteratively or by learning "good" expression

$\%\{data.k_1\}$ AND $\%\{data.k_2\}$ AND $\%\{data.k_3\}$ AND $\%\{data.k_4\}$

.

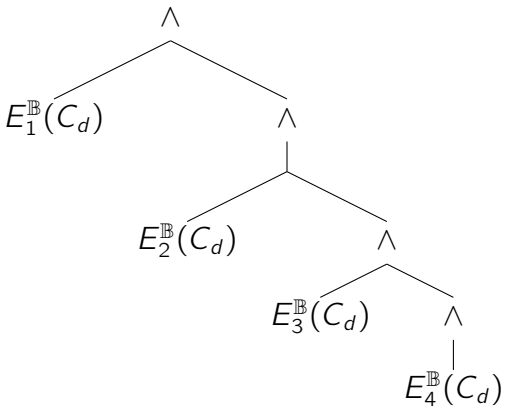


Figure 2: Simplified expression tree for a chain of AND-operators.

trees for the current data of the data stream ([4], [2]). With minor changes to the implementation a rearrangement of the tree nodes would be possible. It's still to be clarified if every evaluation of a data item should be analysed, how often the traversal order should be changed or what is the best algorithm for the rearrangement.

References

[1] Christian Bockermann and Hendrik Blom. The streams framework. Technical Report 5, TU Dortmund University, 12 2012.

[2] Graham Cormode, Flip Korn, S Muthukrishnan, and Divesh Srivastava. Finding hierarchical heavy hitters in data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pages 464–475. VLDB Endowment, 2003.

[3] Avigdor Gal, Sarah Keren, Mor Sondak, Matthias Weidlich, Hendrik Blom, and Christian Bockermann. Grand challenge: the techniball system. In *Proceedings of the 7th ACM international conference on Distributed event-based systems*, pages 319–324. ACM, 2013.

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

A Group Technology-based Concept for Adapting Flow Production to High-Mix, Low-Volume Production

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

Assembly Line Balancing is a central task in designing efficient flow production systems. In the setting of high mix, low volume production, flow systems can hardly be realized as large numbers of product variants lead to significant idle times in production processes. This Tech Report presents a concept for adapting Assembly Line Balancing approaches to those settings by integrating means of Group Technology.

1 The Assembly Line Balancing Problem

Consider an assembly line consisting of S (work-)stations. Products are launched into the line in constant intervals, i.e., the takt-time. At each station $s \in S$ a subset t_s out of all assembly tasks T can be performed at each product. The Assembly Line Balancing Problem (ALBP) deals with optimally assigning all assembly tasks T to the stations S such that a set of predefined constraints is fulfilled. ALBP can be divided in four main types: ALBP-1 aiming at minimizing S for given takt-time, ALBP-2 minimizing the takt-time given S, ALBP-E optimizing line efficiency and ALBP-F creating a feasible assignment [1] [7]. The central constraints of all ALBP types are tasks' precedence restrictions that ensure that all assembly steps are conducted in the right order.

The Simple Assembly Line Balancing Problem (SALBP) was first mentioned by Salveson and Jackson [9] [6]. SALBP focus on single-product assembly lines, i.e. all products

assembled on a given assembly line are identical in terms of required assembly tasks T. Due to the effects of buyer-oriented markets, especially Mass-Customization, the number of different products and the number of variants of each product are steadily increasing. As many markets are saturated, i.e., sales-growth is waning, the production volume for each product variant is decreasing. Consequently, single-product assembly lines will lead to low utilizations of each line so that mixed-model assembly lines (MiMAL) are preferred [4]. MiMAL assemble a defined set of product variants in an arbitrary sequence and do not require setup-times between two consecutive products. The General Assembly Line Balancing Problem (GALBP) optimizes the task-assignment for this type of assembly lines. In order to simplify the problem setting, GALBP are reduced to SALBP before solving by generating a combined precedence graph for all variants produced by the assembly line [10] [8]. A review on solution procedures is given by Boysen et al. [2] [3]. In the GALBP-case, solution quality can be measured in terms of the resulting idle times: due to the differences between different product variants, different assembly tasks may be required for different variants. This leads to idle times, when assembly tasks do not have to be performed. The higher the sum of idle times, the less efficient the assembly line operates. Due to this reason, ALBP-E is of major importance in practice. With the number of product variants increasing, efficiency will steadily decrease, when the GALBP solution procedures described above are applied [4]. To increase efficiency in the case of high-mix, low-volume production so that flow production can be applied, new balancing approached are demanded.

2 Assembly Line Balancing for High-Mix, Low-Volume Production

In order to overcome the drawbacks of state-of-the-art ALBP solutions the MiMAL's central paradigm has to be challenged, i.e., that no setup times occur between consecutive products. Allowing for product-dependent line setups decreases line inefficiency due to idle times, but causes inefficiency due to line setup. To optimize the overall efficiency, the optimal trade-off between inefficiency resulting from both sources has to be determined. Group Technology contains methods to achieve this goal. It aims at identifying families of similar products on the basis of product and process data. Data Mining approaches such as clustering can be applied to identify these families. Products of a certain family are similar to each other, i.e., for assembling these products the same tasks have to be performed, the same precedencies apply and the same components have to be shipped to line. Due to this, members of a single family can easily be produced in an arbitrary sequence without causing setup times. Products from different groups differ in the aforementioned properties. Therefore, setup times solely result from changes between different families. A solution procedure for ALBP incorporating Group Technology is depicted in Fig.1.

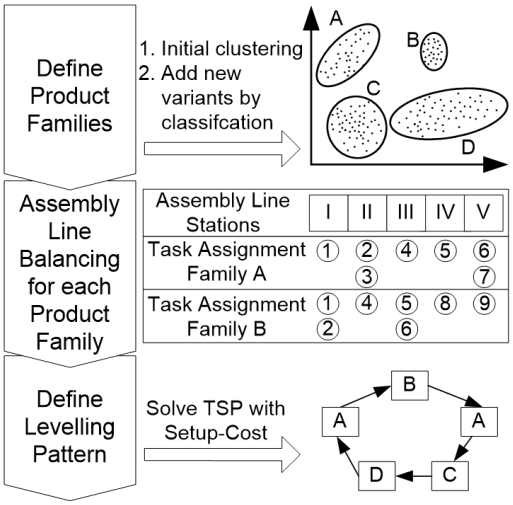


Figure 1: ALBP-Concept for High-Mix, Low-Volume Production Setting [5]

The concept's first part contains all necessary steps to accomplish both, initially creating similar product families as well as adding new product variants to those families once they are established. Especially the latter is relevant in practice as new product variants are designed according to customer preferences during the life-cycle of every product type. Relevant data for family creation is (1) product data, specifying product structure and properties, (2) assembly process data that describes the assembly tasks that have to be conducted, the precedence restrictions between each pair of tasks and required equipment to fulfill the tasks. Moreover, (3) data on logistic's processes is evaluated when forming the families in order to assure that the demanded materials and components as well as their provision processes are similar so that no logistic setup times occur within a family.

For each product group the ALBP is solved in the second step of the methodology. This contains constructing the combined precedence graph for all family members. Constraints that are included derive from the three data sources introduced above: product, assembly processes and logistic processes. As ALBP are NP hard in general [8], solutions heuristics have to be applied for defining the optimal task assignment per family. This step has to repeated as soon as new variants are included that require new assembly tasks or changed task sequences.

Finally, the optimal production sequence of the different product groups has to be determined and capacity has to be assigned to each of them. For computing the optimal sequence setup-times between all groups are determined. Sequence optimization is accomplished by solving a travelling salesman problem (TSP) on basis of families' setup-times. Lastly, to assign capacity slots to each family, customer demand is evaluated for the product variants of each family.

3 Conclusion

The methodology presented in this tech report provides an approach to apply flow production in high-mix, low-volume production settings. Methods of Group Technology are applied in order to control the effects of large numbers of product variants so that line inefficiencies due to idle times can be avoided.

References

[1] J. Bautista, R. Suarez, M. Mateo, and R. Companys. Local search heuristics for the assembly line balancing problem with incompatibilities between tasks. *Proceedings of the 2000 IEEE International Conference on Robotics and Automation*, pages 2404–2409, 2000.

[2] N. Boysen, M. Fliedner, and A. Scholl. A classification of assembly line balancing problems. *European Journal of Operational Research*, 183:674–693, 2007.

[3] N. Boysen, M. Fliedner, and A. Scholl. Level-scheduling bei variantenfliessfertigung: Klassifikation, literaturueberblick und modellkritik. *Journal fuer Betriebswirtschaft*, 57(1):37–66, 2007.

[4] Y. Bukchin and I. Rabinowitch. A branch-and-bound based solution approach for the mixed-model assembly line-balancing problem for minimizing stations and task duplication costs. *European Journal of Operational Research*, 174:492–508, 2006.

[5] J. Deuse, B. Konrad, and F. Bohnen. Renaissance of group technology: Reducing variability to match lean production prerequisites. *Proceedings of the IFAC MIM 2013*.

[6] J. R. Jackson. A computing procedure for a line balancing problem. *Management Science*, 3:261–271, 1956.

[7] A. Lerttira and P. Yarlagadda. Assembly line balancing the comparison of comsoal and msnsh technique in motorcycle manufacturing company. *Advanced Materials Research*, 605-607:166–174, 2013.

[8] S. Matanachai and C. Yano. Balancing mixed-model assembly lines to reduce work overload. *IIE Transactions*, (1):29–42, 2001.

[9] M. E. Salveson. *Journal of Industrial Engineering*, 6(3):18–25, 1955.

[10] N. T. Thomopoulos. Mixed model line balancing with smoothed station assignments. *Management Science*, 16(9):593–603, 1970.

Anomaly Detection for Distributed Sensor Measurements

Marco Stolpe
Lehrstuhl für Künstliche Intelligenz, LS 8
Technische Universität Dortmund
marco.stolpe@tu-dortmund.de

In project's B3 accompanying rolling mill case study, sensors located at different processing stations measure parameters like rolling temperature, force and speed. Based on these parameters, the quality of steel blocks at the end of the milling process should be predicted as early as possible and in real-time. Here, we pose the task as that of distributed anomaly detection. We present two computationally efficient methods for the preprocessing of time series data and the aggregation of resulting feature values in the context of communication-efficient kernel methods.

1 Introduction

The automatic detection of anomalous events like tsunamis, network security breaches, traffic jams or quality deviations in production often requires the analysis of sensor measurements from different locations. Moreover, particular locations can have different types of sensors, e.g. measuring ocean level, water temperature and wind speed. An anomalous event might be indicated by patterns of measurements occuring at single or across different points in time and space. While finding such combinations of measurements can be a challenging data mining task by itself, in many of the aforementioned scenarios resources as time, memory, energy and bandwidth are limited and thus prohibit the transmission of all data to a central server. Based on the premise that communication is more expensive than computation, the first section presents a solution for the efficient aggregation of locally available sensor measurements and the second a new communication-efficient method for distributed anomaly detection in the vertically partitioned data scenario.

2 Efficient Aggregation of Sensor Measurements

In the rolling mill case study, the readings of different sensor types over a meaningful time interval, e.g. a single processing step, can be represented as a multivariate time series. Such time series potentially contain quality-related patterns. The learning task is therefore the identification and extraction of exactly those patterns that allow for training a highly accurate classifier. Since trying all possible patterns would have exponential running time, Mierswa [1] developed a genetic programming approach for the automatic generation of feature extraction processes, successfully classifying music by genre. However, the method can only handle univariate time series of fixed length, while we have multivariate time series of different lengths. Moreover, genre-related features can be expected to occur in most parts of a song, allowing for the correct labeling of windows in a subsample. In contrast, assigning the same label to all parts of a production process is incorrect if only a single part of the process causes a quality deviation. More promising seems to be the shapelet approach [2] which searches for subsequences most correlated with the label. However, it cannot handle multivariate time series.

What we propose in [3] instead is the segmentation of time series into meaningful intervals, based on domain-knowlegde, and describing such intervals by simple statistics, e.g. by their minimum, maximum, mean and standard deviation. These values can be aggregated further, leading to a representation of time series at different levels of granularity. The calculated features are encoded in a single vector per sensor and processing step. All feature vectors belonging to the processing of a single steel block are then concatenated, representing a single training example. The approach has several advantages. First of all, simple statistics remain interpretable and can be efficiently calculated, even on resource-constraint devices. Then, the encoding of time series as fixed-length feature vectors allows for the application of standard supervised learning methods. Furthermore, the approach can handle multivariate time series, as relevant combinations of features belonging to different types of series may be detected by standard feature selection methods. And last, the local aggregation may reduce communication in the distributed setting.

The cleansing of time series and all of the aforementioned processing steps have been implemented in RapidMiner [4], based on domain-knowledge provided by our project partners. Instead of creating a single monolithic process, all steps were analyzed and divided into reusable subprocesses, separating generic from domain-dependent parts as far as possible. The explorative analysis of processes for 470 steel blocks represented by 218 extracted features on a 40x30 SOM yielded a proper identification of different operational modes in the steel factory, as validated by experts from the domain. The results could also be verified by applying k-Means with the DTW distance measure on the raw time series data and comparing the classification results. Though being much faster, the calculation of simple statistics has similar accuracy. Moreover, the visualization of label information on a SOM suggests that quality deviations are local outliers between

arbitrarily shaped clusters of high quality processes. While standard classification methods are unable to detect local outliers, the 1-class SVM described in the next section is especially suited for the detection of such anomalies.

3 Vertically Distributed Core Vector Machine

In comparison to distributed settings where whole observations are stored at different nodes, also known as the horizontally partitioned data scenario, the measurements belonging to the processing of a single steel block are vertically distributed, i.e. assessed and preprocessed at different sensors. While the creation of a single data table as described in the previous section requires the transmission of all data to a central server, communication is severly constraint in settings such as wireless sensor networks [5]. The vertical distribution is particularly challenging, since the detection of quality deviations might depend on a combination of attributes from different nodes. Even more challenging is the use of non-linear kernels, which usually combine several or even all of the available dimensions. Therefore, though training local classifiers and combining their predictions can be highly communication efficient, the accuracy might suffer from non-linear dependencies between the label and combinations of non-local feature values.

A state-of-the-art method that combines local models and a global one is the distributed 1-class SVM by Das et al. [6]. Local 1-class models are trained at each data node, while a global 1-class model is trained at a central node, but only on a sample of the data. During application, the local models send outlier candidates - i.e. local outliers - to the central node which checks if they are also global outliers. Since the number of outliers is assumed to be small, the method is highly communication efficient in the detection phase. However, in comparison to training only local models, communication during training is still high. Moreover, users have difficulties specifying the number of points to sample.

An equivalent method for 1-class learning is the Core Vector Machine (CVM) [7], which yields a $(1+\varepsilon)$-approximation of the minimum enclosing ball (MEB) around all training examples in feature space with high probability. The algorithm incrementally draws fixed-sized samples from the whole dataset and, for each sample, determines the furthest point from the current center in feature space. It stops when all points can be covered by the current MEB, otherwise it calculates a new one. Since it incrementally samples only as many observations as needed, it already solves the problem of having to specify a fixed sample size. In [8], it is demonstrated that by using a combination of local RBF kernels, the furthest point calculation can be distributed. With this Vertically Distributed Core Vector Machine (VDCVM), communication costs during training can be reduced by up to an order of magnitude. Nevertheless, on most of the tested data sets, it yields a similar accuracy as the distributed 1-class SVM.

4 Conclusion and Future Work

Since the segmentation and aggregation of time series values at different processing stations is inherently parallel, they can be seen as distributed preprocessing steps before the calculation of local RBF kernels on the extracted features. It is planned to integrate and combine both methods in the stream framework [9] and to evaluate them on time series from the rolling mill case study.

References

[1] Ingo Mierswa and Katharina Morik. Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58:127–149, 2005.

[2] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proc. of the 15th ACM SIGKDD*, KDD '09, pages 947–956, New York, NY, USA, 2009. ACM.

[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 *46th CIRP Conf. on Manufacturing Systems*, volume 7, pages 193–198. Elsevier, 2013.

[4] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: rapid prototyping for complex data mining tasks. In *Proc. of the 12th ACM SIGKDD*, KDD '06, pages 935–940, New York, NY, USA, 2006. ACM.

[5] K. Bhaduri and M. Stolpe. Distributed data mining in sensor networks. In Charu C. Aggarwal, editor, *Managing and Mining Sensor Data*. Springer, Berlin, Heidelberg, 2013.

[6] K. Das, K. Bhaduri, and P. Votava. Distributed anomaly detection using 1-class SVM for vertically partitioned data. *Stat. Anal. Data Min.*, 4(4):393–406, 2011.

[7] I. Tsang, J. Kwok, and P. Cheung. Core Vector Machines: Fast SVM Training on Very Large Data Sets. *J. Mach. Learn. Res.*, 6:363–392, December 2005.

[8] M. Stolpe, K. Bhaduri, K. Das, and K. Morik. Anomaly detection in vertically partitioned data by distributed core vector machines. In *Proc. of the ECML PKDD 2013*. Springer, 2013.

[9] C. Bockermann and H. Blom. Processing Data Streams with the RapidMiner Streams-Plugin. In *Proc. of the 3rd RapidMiner Community Meeting and Conference*, 2012.

Subproject B4
Analysis and Communication for dynamic traffic prognosis

Michael Schreckenberg Christian Wietfeld

Improving a Microscopic Traffic Simulation using Real-Time Information on Weather Conditions

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

This report illustrates the benefit of real-time weather data for highway traffic simulation and describes the modelling and integration of weather conditions into a complex microscopic traffic information system. Using stationary measured weather data as an example, the achieved results show the potential extended Floating Car Data (xFCD) can have for traffic simulation.

As a part of the OLSIM traffic information system [1], we have access to real-time highway traffic data provided by the German state of North Rhine-Westphalia. Data from various weather stations located directly at the highways are provided as well. We decided to use these stationary measured data to examine, if weather data provided by vehicles equipped with xFCD transmitters can be helpful to improve traffic simulations.

The whole weather station network comprises of more than 300 units, most of them located on bridges or at hilly sections, where ice and snow can cause major problems during winter. Each station is at best equipped with sensors for

- air and surface temperatures,
- rain intensity or water film thickness and a derived categorical value that describes the surface condition,
- wind intensity and direction.

Most of the stations do not provide all of these sensor types, but very often, sensors for temperatures, water-film thickness and surface condition are available.

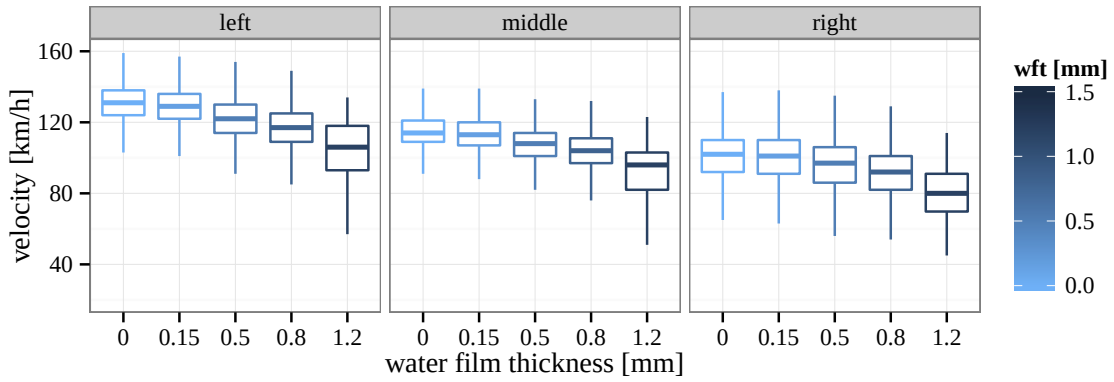


Figure 1: Passenger car velocity on each lane in northbound direction depending on the water film thickness measured at weather detector "750m". The depicted data were gathered on working days only.

For the purpose of simulation, we decided to have a closer look at a particular section of a highway, that has a suitable number of weather stations with good data availability and quality. After an analysis of the available data, we chose a section on the German highway A2 between the exits Bielefeld-Sennestadt and Bielefeld-Zentrum. The A2 is a transit highway with a high percentage of long-distance traffic and generally has three lanes in each direction. Due to a hilly landscape between both exits and the risk of ice buildup during winter, the section is equipped with eight weather stations. Two traffic loops provide the traffic data. Traffic and weather data were available from July 2012 to April 2013 for this section.

The key results of the empirical analysis are

- an increasing water film thickness is the main reason for weather-related velocity drops,
- air and surface temperatures show only a small influence on the driven velocities. Slightly lower mean velocities can be noticed at temperatures between 0°C and −10°C, presumably for safety reasons,
- apparently, the pure amount of rain measured by rain intensity sensors does not impact the velocity that much.

Following these results, we decided to focus our modelling on the impact given by the water film thickness. As depicted in fig. 1, it can cause massive velocity drops from about 30%, if one compares the mean velocities on the left lane with 0mm and 1.2mm water film thickness. Examples in our data set show that such a decrease can take place within only three minutes. As every velocity drop means increasing travel times, such an information can be very useful for a traffic information system like OLSIM.

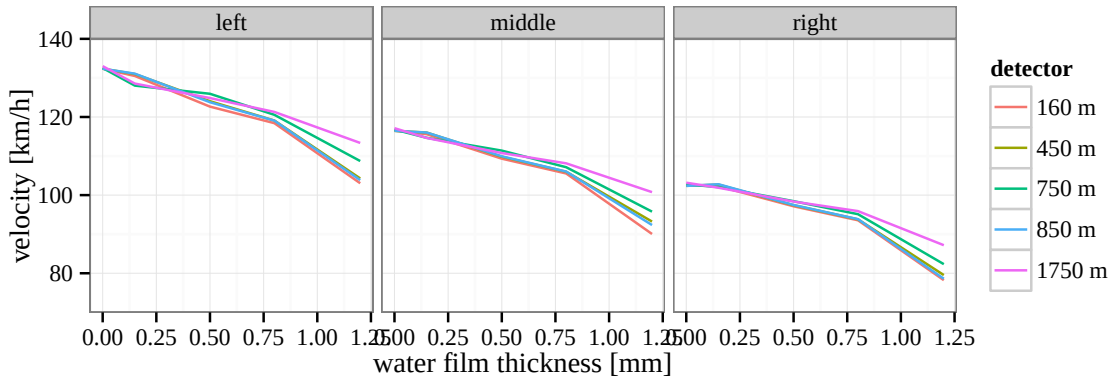


Figure 2: Passenger car velocity and water film thickness at different weather stations. The different colors illustrate the distance between weather detector and traffic detector.

Prior to any simulation attempt, one has to answer a very practical question: Often weather sensors and traffic detectors are not located at the same position of the highway. Which distance between them can be tolerated for simulation purposes, such that the described weather impact is still there? That means, an area of effect has to be defined. According to fig. 2, a maximum distance of about 800m between both sensors seems to be justifiable here. This result sounds promising for a planned use of xFCD vehicles, as one only needs less than 5% equipped vehicles to get a comparable coverage like in this stationary example [4].

The recently finished version 4 of OLSIM allows for simulating multi-lane highway traffic with different realistic cellular automata models. We used a model invented by Lee *et al.* [5], that features limited breaking capabilities, for our simulations, combined with additional rules for multi-lane traffic on highways with asymmetric lane usage (as on Germany highways) [2]. To simulate weather conditions, every vehicle located up to 800m away from a weather sensor performs an additional update step

$$v_n(t) = \max\left(v_n(t-1) - D_n, v_n^* - D(w)\right)$$

to calculate its velocity v_n in the next timestep. Here, v_n^* is the velocity given by the models [2, 5]. The deceleration because of rain is described by $D(w)$, a binary value which is stochastically determined with different probabilities $\rho(w)$ for different values of the water film thickness w. The first term ensures that no vehicle is forced to break harder than its maximum deceleration D_n.

Fig. 3 shows recent results from simulation runs of the described highway section. As clearly visible, we were able to find appropriate parameters for $\rho(w)$ to reproduce the impact of rain in our traffic simulation.

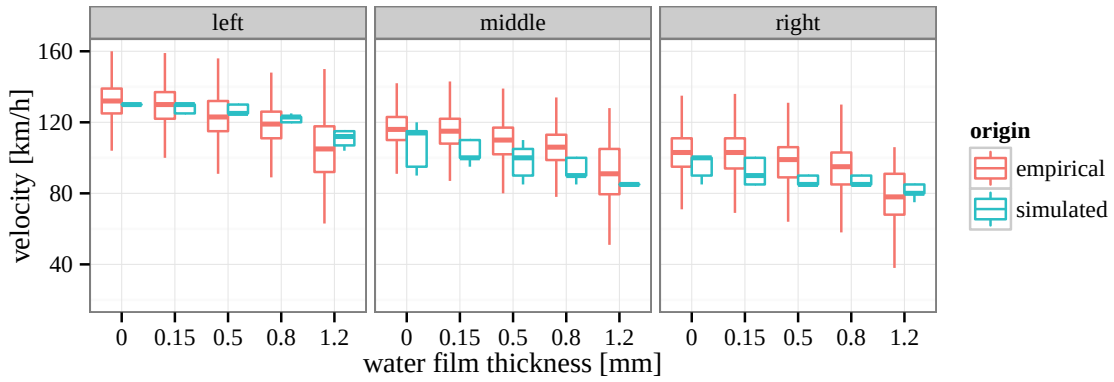


Figure 3: Comparison of empirical results from weather detector "160m" and simulation results from the "weather-improved" cellular automaton model.

The results shown in this report and a detailed explanation of the additional weather rule set will shortly be published in [3].

References

[1] Johannes Brügmann, Wolfram Luther, and Michael Schreckenberg. Real-Time Traffic Information System Using Microscopic Traffic Simulation. In *EUROSIM 2013 – 8th EUROSIM Congress on Modelling and Simulation*, pages 448 – 453, Cardiff, Wales, September 2013. EUROSIM, IEEE – The Institute of Electrical and Electronics Engineers, Inc.

[2] Lars Habel. Modellierung, Implementierung und Validierung von streckenabhängigem Fahrverhalten und asymmetrischen Fahrstreifenwechseln für das Modell von Lee, Barlovic und Pottmeier. Diplomarbeit, Universität Duisburg-Essen, 2011.

[3] Lars Habel, Timo Knaup, Christoph Ide, Christian Wietfeld, and Michael Schreckenberg. Improving a Microscopic Traffic Simulation using Real-Time Information on Weather Conditions. *Journal of Statistical Mechanics*. to be submitted.

[4] Christoph Ide, Brian Niehoefer, Timo Knaup, Daniel Weber, Lars Habel, Christian Wietfeld, and Michael Schreckenberg. Efficient Floating Car Data Transmission via LTE for Travel Time Estimation of Vehicles. In *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, pages 1–5, 2012.

[5] Hyun Keun Lee, Robert Barlovic, Michael Schreckenberg, and Doochul Kim. Mechanical Restriction versus Human Overreaction Triggering Congested Traffic States. *Phys. Rev. Lett.*, 92:238702, Jun 2004.

Efficient LTE Communication for the Dynamic Traffic Prognosis

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

In this report, the performance of Channel-Aware Transmission (CAT) for efficient Long Term Evolution (LTE) Machine-Type Communications (MTC) is evaluated based on resource allocation measurements in the field. By means of a spectrum analyzer, the LTE resource utilization in terms of occupied Resource Blocks (RBs) of the uplink is quantified as a function of the channel conditions. The measurement results are used to quantize the potential gain of CAT. It is shown that by means of the scheme, the influence of MTC on other LTE users can be significantly reduced. Furthermore, the impact of a cluster-based MTC on the LTE resource utilization is presented.

1 Channel-Aware Transmission Scheme

In order to minimize the impact of MTC on Human-to-Human (H2H) communications, we presented CAT in [1]. The conclusions presented in [1] are based on laboratory measurements. In addition, the impact of different channel conditions on the radio resource utilization in a real LTE network is analyzed in [2]. The channel quality has a major impact on the performance of cellular wireless communication systems like Long Term Evolution (LTE). This is due to two main concepts: The channel-aware choice of the Modulation and Coding Scheme (MCS) and the retransmission of erroneously received packets. To exemplify these effects, the impact of the radio channel on the achievable throughput is shown in Fig. 1 for different transport layer protocols and different fading conditions. These measurements are performed in the laboratory by means of an LTE base station emulator and a channel emulator. The figure also illustrates

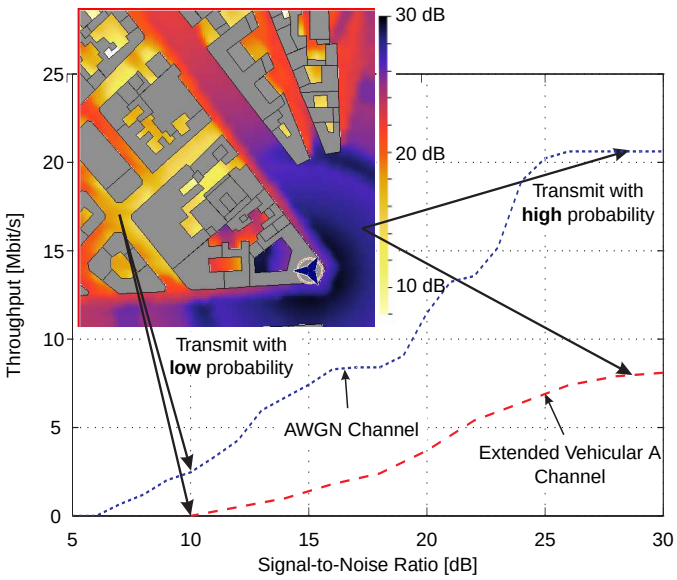
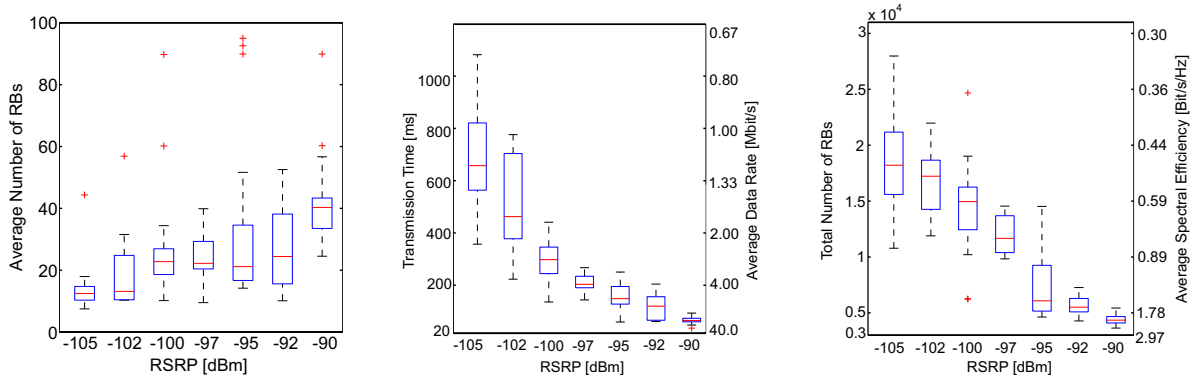


Figure 1: Radio Channel Dependent Throughput (Laboratory Measurements) and Example Ray Tracing Simulation Illustrating the Idea of CAT.

the idea of CAT. For good channel conditions (see Ray Tracing map) extended Floating Car Data (xFCD) should be transmitted with a higher probability in contrast to bad channel conditions.

2 LTE Field Resource Measurements

For capturing the radio resource requirements in time dimension (transmission duration) and frequency dimension (number of RBs) of an LTE User Equipment (UE) in a real-world LTE deployment, a real-time spectrum analyzer is used. The channel quality is thereby represented by the Reference Signal Received Power (RSRP) value. For bad channel conditions (RSRP = -105 dBm), only 14 RBs median are assigned to the user (cf. Fig. 2a). The value increases to 40 RBs for a good RSRP of -90 dBm. This effect is unexpected, because the scheduler typically assumed in literature usually apply a resource or rate fair scheduling. For a rate fair scheduling, users with good channel conditions would get even a lower number of RBs in contrast to cell edge users with worse conditions. However, the opposite is obviously done in the real LTE network for a small payload size (100 kByte). The total number of allocated RBs for different RSRP values (it combines the resource utilization in frequency (cf. Fig 2a) and time (cf. Fig 2b) dimension) is shown in Fig. 2c. This figure of merit describes how intensive the LTE air interface is utilized by the data transmission.



(a) Average Number of RBs (b) Transmission Time (c) Total Number of RBs

Figure 2: Results of LTE Resource Measurements in the Field (100 kByte Payload).

3 Performance Evaluation of Channel-Aware MTC

The resource utilization measurement results are used to determine the benefits of CAT. For this purpose, the Markovian model for resource utilization modeling of LTE for many users (cf. [1]), is parameterized by the results of required RBs from Section 2. By means of the model, the blocking probability of H2H traffic (cf. Fig. 3) for H2H and MTC traffic is calculated. It can be seen from the figure that CAT leads to a significantly lower blocking probability. The average blocking probability is smaller for the results from the laboratory measurements because Additive White Gaussian Noise (AWGN) channel conditions are assumed (cf. Fig. 1 for data rates). The channel conditions in the field measurements are given by the measurement positions (cf. [2]). Here, many locations are characterized by Non-Line-Of-Sight (NLOS) conditions.

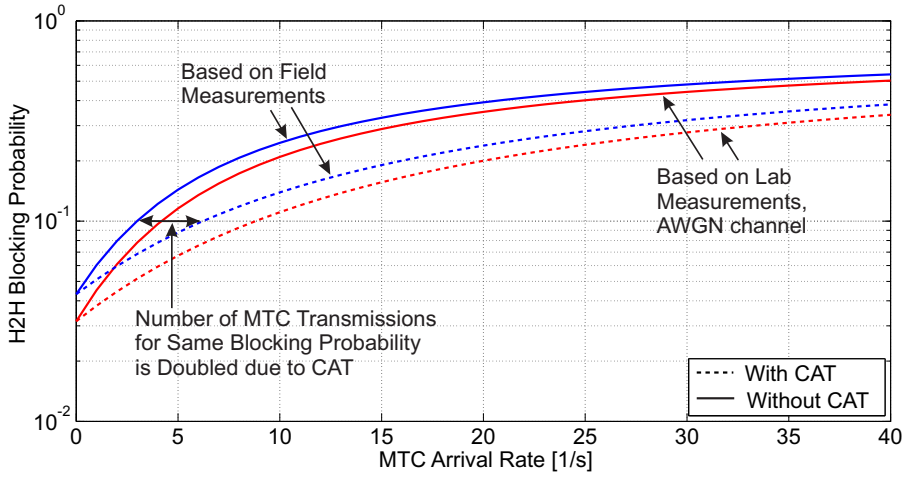


Figure 3: H2H Blocking Probability for Channel-Aware MTC. Comparison Between Lab and Field Measurements.

4 Impact of Cluster-based MTC on the LTE Utilization

Furthermore, in [3] clustering algorithms for a efficient LTE MTC are presented and the performance is evaluated by means of a novel simulation model. The mean LTE Physical Uplink Shared Channel (PUSCH) utilization over all eNodeBs (5 for highway and 14 for urban scenario) and the utilization of the LTE cell with the highest load are illustrated as box plots in Fig. 4. It can be seen from the figure that No Clustering (NC) leads to a high utilization and high peaks (especially for the urban scenario). In addition, the median of the average PUSCH utilization can be reduced from 1.7% to 0.3% for an example highway scenario.

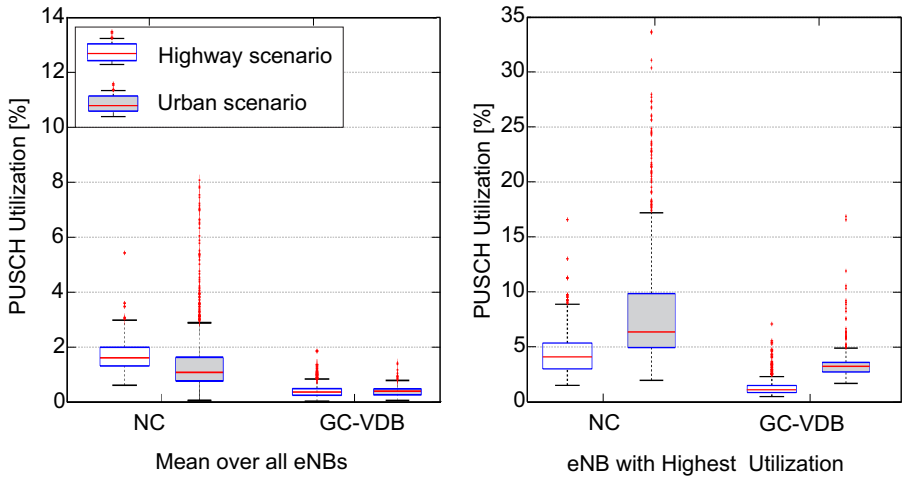


Figure 4: LTE PUSCH Utilization for No Clustering (NC) and under Cluster-Based xFCD Collection (GC-VDB Algorithm (cf. [3])).

References

[1] C. Ide, B. Dusza, M. Putzke and C. Wietfeld, *Channel Sensitive Transmission Scheme for V2I-based Floating Car Data Collection via LTE*, IEEE International Conference on Communications (ICC), Ottawa, Canada, Jun. 2012.

[2] C. Ide, B. Dusza and C. Wietfeld, *Performance of Channel-Aware M2M Communications based on LTE Network Measurements*, IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), London, UK, Sep. 2013.

[3] C. Ide, F. Kurtz and C. Wietfeld, *Cluster-Based Vehicular Data Collection for Efficient LTE Machine-Type Communication*, IEEE 78th Vehicular Technology Conference (VTC-Fall), Las Vegas, USA, IEEE, Sep. 2013.

Precise Vehicle Positioning in Urban Canyons using Local Interference Compensation

Brian Niehöfer
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
brian.niehoefer@tu-dortmund.de

The steadily increasing traffic density is causing enormous negative effects such as jams, accidents or CO_2 emissions. Hereby, a reliable and dynamic congestion forecast is the key, to avoid and/or compensate such locale bottleneck situations. To enable and further improve such predictions, the Communication Networks Institute (CNI) focuses on a lane-specific positioning of vehicles to enable a more detailed and lane-accurate traffic prediction (e.g. detecting short-dated roadworks) in the future. Thereby, the scientific challenge is to predict, quantify and compensate the inevitable local impacts to the positioning accuracy when using ordinary GPS receivers especially in urban canyons, so namely areas with a high probability of congested traffic situation. Hence, the CNI developed a combined simulation framework, the so-called *Local Interference Compensation (LOCATe)* to overcome those limitations with a back-end solution and furthermore provides a Software-Defined GPS receiver solution to evaluate the performance gain in experimental measurements.

1 LOCATe - Local Interference Compensation for Global Navigation Satellite Systems

Former publications already clarified and quantify the impact of local circumstances to the satellite positioning accuracy and the necessity to compensate them to increase the potential application field of Global Navigation Satellite Systems [1] [4]. In contrast

to other scientific approaches, the Local Interference Compensation (LOCATe), focuses on the avoidance of position-specific and unavoidable failures within GNSS positioning techniques without any additional hardware in the embedded resource-constrained front-end. Figure 1 visualizes the approach. Based on a detailed 3D model of the direct

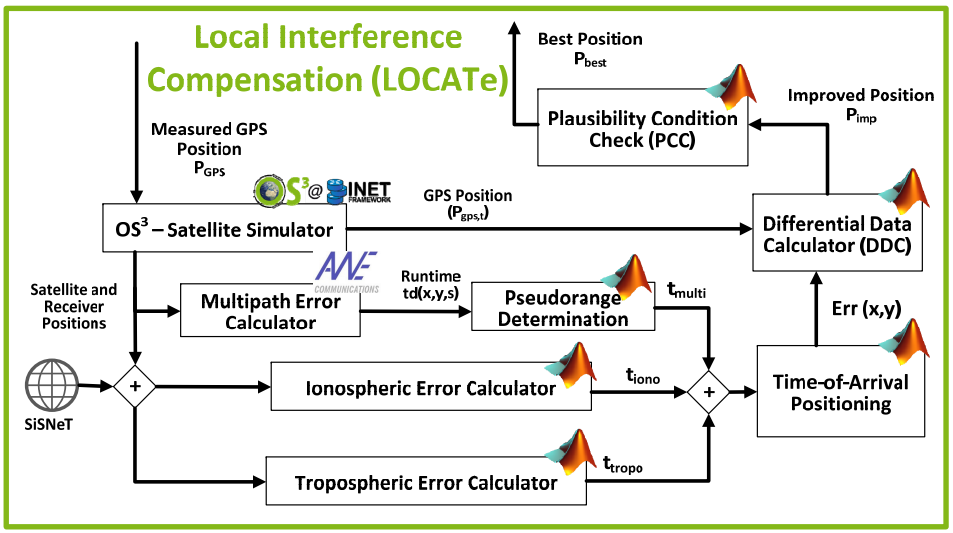


Figure 1: Architecture of LOCATe

receiver surroundings, the LOCATe system simulates accurately satellite movement of any actual or future GNSS constellation using the developed *Open Source Satellite Simulator (OS³)* [3]. Afterwards, the impulse responses are evaluated using ray tracing technology. Taking all influences into account, atmospheric as well as multipath effects. A Time-of-Arrival routine is used to calculate the real receiver's position and by that the differences to the supposed one by the GNSS. The difference in between is provided back to the simulation framework.

2 Real-Life Validation using an Advanced Software-Defined GPS Receiver Solution (ASDR)

Beneath already performed simulation results evaluating the possible performance of LOCATe [4], the CNI also focuses on a real-life evaluation. Therefore, a measurement equipment based on *Software-Defined Radio* as well as the gained experimental results in using LOCATe will be introduced in the upcoming section.

2.1 Advanced Software Defined GPS Receiver (ASDR)

Software Defined Radio (SDR) might be understood as a platform for the introduction of new technologies and services into existing live networks. Especially when thinking on multiple satellite positioning systems the adaptability of SDRs clarifies the effectiveness and the advantage for future satellite-based positioning techniques. In addition, an SDR implementation allows a realistic evaluation of developed services, because no 'external' measurement equipment (*Black Box*) may influence the results. Hence, the CNI implemented the GPS positioning algorithms and functionalities to evaluate the developed LOCATe system. Additionally, some modifications enhance the functionalities of already known GPS-SDR approaches.

First of all, the filtering routine was integrated within the GNU Radio implementation. This comes along with the advantage to monitor the filtering steps during runtime in the field. A further modification addresses a dual link architecture to allow back-end applications to increase the positioning performance via post-processing and send those back to the SD-GNSS receiver. In addition, the so-called Smart Constellation Selection (SCS) uses known circumstances in the receiver to increase the positioning performance without additional hardware or resource-intensive algorithms or communication.

2.2 Experimental Validation Tests

To enable a valid performance analysis of the LOCATe system, the CNI uses two points specified by the land surveying office at the campus of the university in Dortmund as measurement reference. Both points are matching to the definition of the introduced urban canyons and by that, standing for the most challenging environments for satellite positioning systems. The results of more than 500 measurements are shown in Figure 2. Up to 23.6% of all positions can be determined lane-specific ($< 1.5m$, visualized by the red solid line) using LOCATe in contrast to just 7% with pure GPS and reduce the average error by more than 45%, and even more important: LOCATe lowers the occurring peak value significantly beneath $9m$, what again clarifies the benefit.

In addition, filter algorithms may be used to further improve the accuracy gain. As an example, the CNI added a randomly chosen and adapted particle filter (well-known method in GPS positioning techniques) to the already smoothed results from LOCATe and by that decreases overall error by 63.3% in average. Furthermore, a combined handling using LOCATe and additional filters, increases the lane-specific detection by four times. It should be mentioned, that this filter is just used to clarify the additional performance possibility and shall not indicate the finest choice to work with LOCATe. This topic will be investigated in the next weeks in addition to the impact of the modeling level of detail to the accruing accuracy gain.

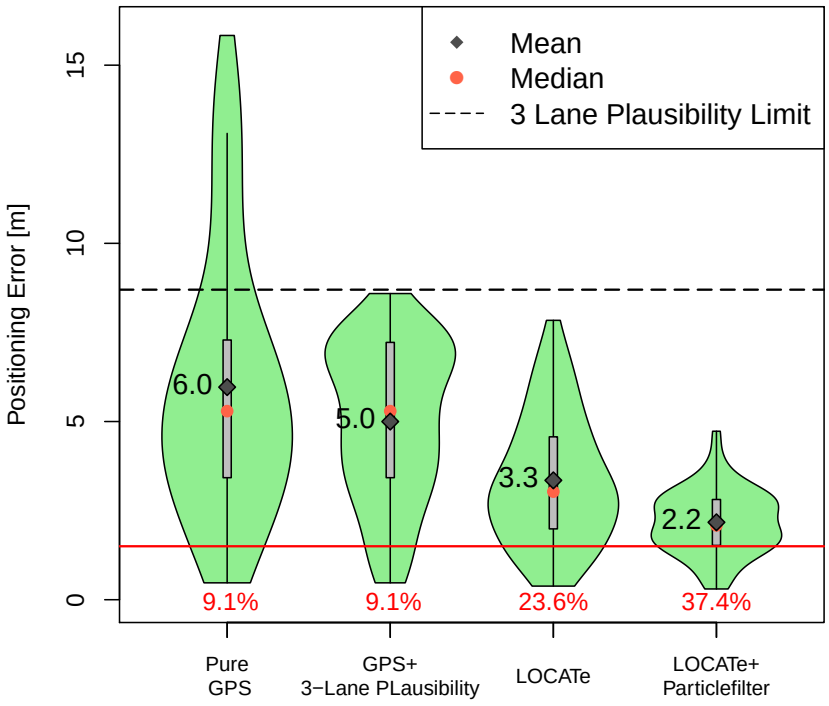


Figure 2: Performance Gain using LOCATe and Additional Randomly Chosen Particle Filter in Comparison to pure GPS Data

References

[1] Groves, P.D., Jiang, Z., Wang, L. and Ziebart, M.K.. *Intelligent Urban Positioning using Multi-Constellation GNSS with 3d Mapping and NLOS Signal Detection*, ION GNSS, Nashville, USA 2012.

[2] Niehoefer, B., Schweikowski, F. and Wietfeld, C., *Smart Constellation Selection for Precise Vehicle Positioning in Urban Canyons using a Software-Defined Receiver Solution*, submitted to 20th IEEE Symposium on Communications and Vehicular Technology (SCVT), 2013, Namur, Belgium.

[3] Niehoefer, B., Subik, S. and Wietfeld, C. *The CNI Open Source Satellite Simulator based on OMNeT++*, 6th OMNeT++ Workshop (SimuTools), 2013, Cannes, France.

[4] Niehoefer, B., Lehnhausen, S. and Wietfeld, C. , *Combined Analysis of Local Ionospheric and Multipath Effects for Lane-Specific Positioning of Vehicles within Traffic Streams*, 6th ESA Workshop on Satellite Navigation (NaviTech), 2012, Noordwjik, Netherlands

SFB 876 Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

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

Katharina Morik Alexander Schramm

Analyses of lncRNAs in neuroblastoma cells after knockdown of the JARID1C histone demethylase

Kathrin Fielitz
Oncology Lab – Childrens Hospital Essen
Universität Duisburg-Essen
Kathrin.Fielitz@stud.uni-due.de

Neuroblastoma is the most common solid extracranial malignancy of childhood, accounting for 15% of the deaths attributed to malignancies in children. Since neuroblastoma shows a large clinical heterogeneity, there is urgent need of identifying genes usable for outcome prediction or targeted therapy. We previously identified upregulation of the histone demethylase, JARID1C, as a marker of aggressive neuroblastoma. JARID1C modulates target gene expression by altering accessibility of the transcriptional machinery to chromatin. A possible role for genes, which are not translated into proteins, has been emerging for neuroblastoma as well as for other cancer types. In the current project we follow the hypothesis that JARID1C could influence neuroblastoma biology by regulating lncRNAs.

Neuroblastoma is the most common extracranial malignancy of childhood, showing a large heterogeneity in clinical courses. Unfavorable neuroblastomas proceed quickly, aggressively and fatally, while tumors with a favorable biology sometimes even regress without treatment. Important clinical features determining tumor aggressiveness include age at diagnosis, tumor stage and MYCN amplification status, since this oncogene is associated with poor patient outcome. In spite of the consistently ongoing advancements in developing and improving diagnostic and therapeutic opportunities, it remains difficult to make a reliable prediction for the individual course of the disease. In the past few years, microarray and real-time PCR based approaches have been heavily used to identify genes, which could serve as surrogate outcome markers or novel medical target structures [1] [2]. The entire human genetic information consists of 20.000 protein-coding genes, but

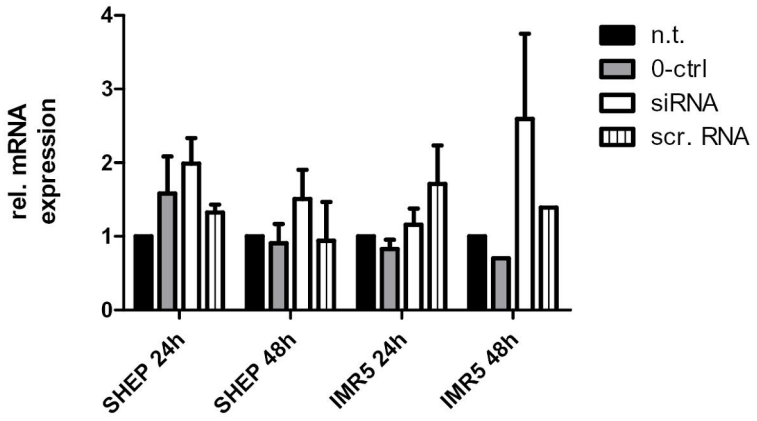


Figure 1: Expression of MALAT1 RNA after JARID1C knock-down in SHEP and IMR5 NB cell lines. The graph shows average valus +/-SEM

these represent only 2% of the human genome. By now, it is established that 70% of the genome are transcribed into non (protein)-coding RNA [3]. The best described class of these non-coding RNAs is the group of microRNAs (miRNAs). In neuroblastoma, global miRNA expression levels as well as specific roles e.g. for the tumor-suppressive miR-137and the oncogenic miR17-92 cluster have been reported by us and others [4] [5]. A less well described family of untranslated RNAs are long non-coding RNAs (lncRNA), comprising transcripts >200 nucleotides. The lncRNAs play an important role in cell biology and cancer [6]. Bernard and coworkers were able to show that the long non-coding RNA Malat1 (metastasis associated lung adenocarcinoma transcript 1) increases differentiation and activity of hippocampal neurons [7]. Yet, Malat1 was originally described as a prognostic parameter for patients with stage 1 lung adenocarcinoma or squamous cell carcinoma [8].
We already identified upregulation of the histone demethylase JARID1C in aggressive NB independent of MYCN . We hypothesized that JARID1C exerts its functions at least in part due to regulation of lncRNAs. In pilot experiments, we were able to show an increase of Malat1 expression following JARID1C knockdown [Fig.1]. This finding motivated us to analyze a more general role for JARID1C in modulating lncRNA expression.

RNA was extracted from cells with either normal or transiently down-regulated JARID1C levels and the expression of lncRNAs was analysed by qPCR. lncRNA levels in primary tumors (n=274) were quantified by reanalyses of HuEx exon arrays. Only those expression values were considered, for which the difference in Ct did not exceed a threshold of 0.5 between control samples . The cq values (log-scale) were denoted as cqi_{si}, cqi_{nt} and cqi_{0ctrl} for a given lncRNA i. The consistency between controls was defined as

$$|cqi_{0ctrl} - cqi_{nt}| \quad (1)$$

After this filtering step, expression changes in siJARID-treated cells were calculated compared to controls as log-fold-change

$$cqi_{si} - 1/2(cqi_{0ctrl} + cqi_{nt}) \quad (2)$$

Several lncRNAs were significantly regulated upon knock-down of JARID1C. Amongst these were the lncRNAs NEAT1 (log-fold-change of 2,3 compared to control, p= 3,83x 10^{-6}) and Malat1 (log-fold-change of 1,5 compared to control, p= 0,006862).

We compared these results to the expression of the respective lncRNAs in the primary tumors. We were able to show a significant correlation between upregulated lncRNAs after JARID1C knock-down and lncRNAs upregulated in primary tumors (Tab.1). We will further analyse the contribution of JARID1C-regulated lncRNA to neuroblastoma biology focussing on the top targets depicted in table 1.

lncRNA ID	No treatment	control	siRNA treatment	Foldchange	Tumorcount	p-val expressed	JARID1C correlation	p-val correlation
XIST	3,7877	4,08549	4,7192	0,7825	119	$1,48x\ 10^{-5}$	0,5153273	$7,39x10^{-18}$
CLUHP3	1,5177	1,59549	2,0692	0,5125	274	$7,96\ x\ 10^{-176}$	0,36085	$6,5x10^{-9}$
TUG1	4,5177	4,53549	4,6792	0,1525	274	$3,15x10^{-301}$	0,30594	$1,23x10^{-6}$
SNHG17	1,1177	1,01549	2,5292	1,4625	274	$6,66x10^{-178}$	0,298476	$2,25\ x10^{-6}$
NEAT1	6,0877	6,54549	8,6092	2,2925	274	$3,46x10^{-157}$	0,29847	$3,83x10^{-6}$
H19	0,6377	0,69549	1,1792	0,5125	274	$2,31x10^{-168}$	0,2823749	$8,43x10^{-6}$
MARS2	3,0677	2,89549	3,5492	0,5675	265	$1,04x10^{-82}$	0,261823	$3,95x10^{-5}$
SNHG1	3,7477	3,78549	4,3392	0,575	273	$1,46x10^{-135}$	0,2258117	0,000406
MYCNOS	-1,6022	-1,7295	-1,3807	0,285	64	0,0327898	0,1827501	0,00409
STX18-AS1	-1,7822	-1,3295	-0,0107	1,545	271	$1,44\ xx10^{-127}$	0,1768636	0,005389
MALAT1	8,7877	9,27549	10,549	1,5175	274	0	0,1716322	0,006862

Table 1: **The expression of lncRNAs in primary NB (exon array) and in NB cells after JARID1C knock-down** - no treatment, control, siRNA treatment: expression of lncRNAs in NB cells after JARID1C knockdown and in controls; foldchange: factor by which the expression in siRNA treated cells differs from the controls; tumorcount: number of tumors the lncRNA was expressed in; pval-expressed: p-value of the lncRNA measurement in the microarray; JARID1C correlation: correlation between the JARID1C expression in the microarray and the lncRNA in the micro-array; pval-correlated: p-value of the correlation between JARID and lncRNAs in the array

References

[1] Schramm, A.; Schulte, J.H.; Klein-Hitpass, L.; Havers, W.; Sievers, H.; Berwanger, B.; Christansen, H.; Warnat, P.; Brors, B.; Eils, J.; Eils, R.; Eggert, A.: Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling.

In Oncogene vol.: 24, 7902-7912, 2005
[2] Schramm, A.; Vandesompele, J.; Schulte, J.H.: Translating expression profiling into a clinically feasible test to predict neuroblastoma outcome. In Clinical Cancer Research, vol. 13: 1459 – 1465; 2007
[3] Diederichs, S.: Nichtcodierende RNA in malignen Tumoren – Eine neue Welt von Tumor-Biomarkern und Zielstrukturen in Krebszellen. In: Pathologe. Vol.: 31, 258 – 262; 2010
[4] Althoff, K.; Beckers, A.; Odersky, A.; Mestdagh, P.; Köster, J.; Bray, I.M.; Bryan, K.; Vandesompele, J.; Speleman, F.; Stallings, R.L.; Schramm, A.; Eggert, A.; Sprüssel, A.; Schulte, J.H.: MiR-137 functions as a tumor supressor in neuroblastoma by down-regulating KDM1A. In International Journal of Cancer, vol.: 133, 1064 – 1074; 2013
[5] Mestdagh P, Boström AK, Impens F, Fredlund E, Van Peer G, De Antonellis P, von Stedingk K, Ghesquière B, Schulte S, Dews M, Thomas-Tikhonenko A, Schulte JH, Zollo M, Schramm A, Gevaert K, Axelson H, Speleman F, Vandesompele J.: The miR-17-92 microRNA cluster regulates multiple components of the TGF-b pathway in neuroblastoma. In Molecular cell. Vol.: 40; 762 – 773; 2010
[6] Gutschner, T.; Diederichs, S.: The Hallmarks of cancer: a long non-coding RNA point of view. In RNA Biology vol.: 9, 703 – 719; 2012
[7] Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, Zhang MQ, Sedel F, Jourdren L, Coulpier F, Triller A, Spector DL, Bessis A..: A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. In: The EMBO Journal, vol: 29, 3082 – 3093; 2010
[8] Ji, P.; Diederichs, S.; Wang, W.; Boing, S.; Metzger, R.; Schneider, P.M.; Tidow, N.; Brandt, B.; Buerger, H.; Bulk, E.: MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene, vol.: 22; 8031 – 8041; 2003

Investigation of the glutamine transporter SLC1A5 as a new Neuroblastoma druggable target

Melanie Schwermer
Oncology Lab – Childrens 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 find new neuroblastoma markers and therapy targets to enable a better outcome prediction. The search for such genes was here based on "Exon Array" analysis. We identified a gene, SLC1A5, whose expression correlates with poor survival and aggressive tumor biology. SLC1A5 is an amino acid transporter that can be inhibited by small molecule antagonists. Here, we analysed the effects of abrogating SLC1A5 function in vitro.

Neuroblastoma (NB) is the most common and deadly solid tumors 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 an individual therapy, it is essential to extend the knowledge of the molecular mechanisms causing 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 will deal with the identification of NB relevant genes.

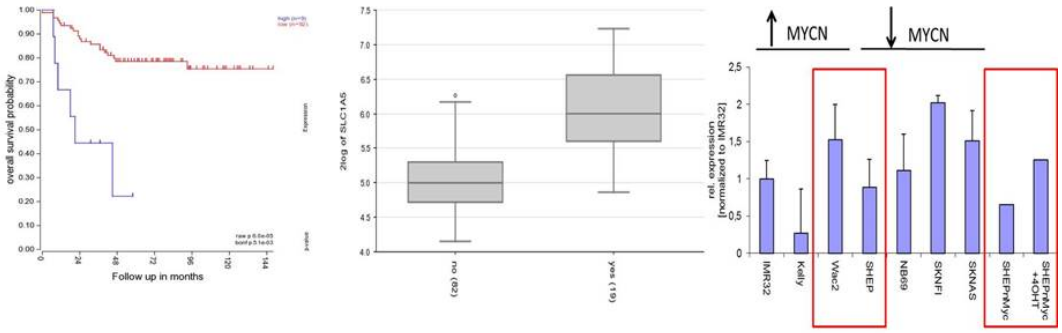


Figure 1: **SLC1A5 expression correlates with a poor outcome and MYCN status.** The Kaplan-Meier curve reveals an upregulation of SLC1A5 in patients with a fatal course of disease (bonf p: 5,1*10-3). SLC1A5 mRNA levels are also elevated in MYCN amplified tumors (p: 1,2*10-13; fig1 a and b were generated using R2; http://hgserver2.amc.nl/cgi-bin/r2/main.cgi). The association between SLC1A5 and MYCN could be validated in ectopically MYCN-expressing SHEP cells (SHEP: low MYCN level; Wac2: high MYCN level) and in MYCN inducible SHEP cells (SHEPnMyc). SLC1A5 is also highly expressed in NB single copy cell lines (NB69, SKNFI and SKNAS).

Gene expression profiling identified SLC1A5 (ASCT2) as potential NB druggable target. SLC1A5 expression correlates with a poor outcome and with MYCN amplification.

We could confirm correlation between MYCN and SLC1A5 expression in NB cell lines. SLC1A5 is higher expressed in cell lines with ectopic MYCN expression (Wac2) compared to the parental cell line (SHEP) and the expression is also increased in MYCN inducible SHEP cells (SHEPnMyc) after MYCN induction (SHEPnMyc +4OHT). Interestingly, also cell lines with low MYCN levels have a high SLC1A5 expression and vice versa (Fig.1) SLC1A5 belongs to the SLC (solute carrier) transporter super family. Currently approximately 300 SLC-transporters, which can be divided into 43 families have been identified [5]. SLC1A5 is a member of the SLC1 family, which comprise five high affinity glutamate transporter and two amino acid transporter including the SLC1A5 transporter [6]. SLC1A5 shows a high affinity for glutamine, but has also considerable affinity to other amino acids, including alanine, serine and threonine. Amino acid transport across the plasma membrane is important for cellular metabolism [7]. Cancer cells have a high proliferation rate, so that they depend on amino acid supply to sustain their biosynthetic pathways. To analyze SLC1A5 as a potential target in Neuroblastoma, we blocked the transporter by using an SLC1A5 inhibitor, gamma-L-Glutamyl-p-Nitroanilide (GPNA). The inhibitor causes a decrease of cell viability after 72 hour in NB cell lines. The strongest effect could be observed in Wac2 cells. After 48 hour the treatment of 1 mM GPNA leads to a decrease of the p70 ribosomal protein S6 kinase-1 (S6K1) and of the eukaryotic translation initiation factor 4E-binding protein 1 (4EBP) and also of their phosphorylated forms (Fig.2). S6K1 and 4EBP are phosphorylated by the ser-

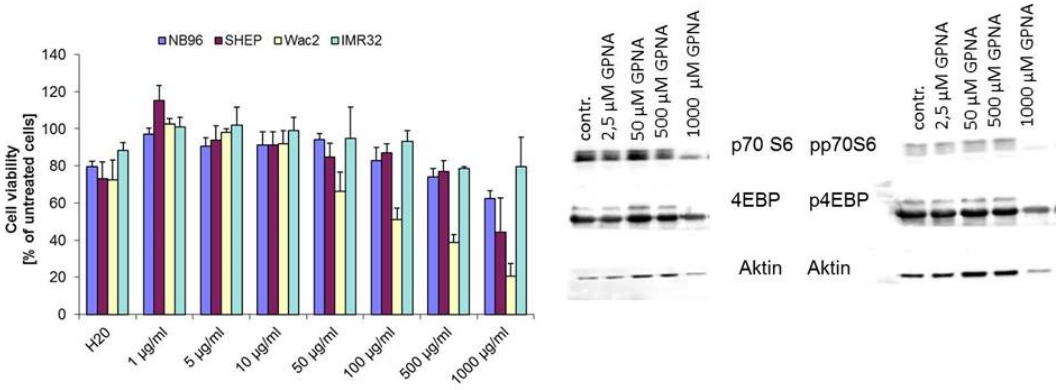


Figure 2: **Inhibition of SLC1A5 causes a decrease in cell viability and has an inhibitory effect on the mTor pathway** After 72 h GPNA treatment the cells have a reduced cell viability compared to untreated cells. Wac2 seems to be the most sensitive cell line (yellow bar). The Western blot (here shown for SHEP) reveals a decreased expression of the p70 ribosomal protein S6 kinase-1 (S6K1) and of the eukaryotic translation initiation factor 4E-binding protein 1 (4EBP) as well as of their phosphorylated forms after 48 h of GPNA treatment.

ine/threonine kinase mammalian target of rapamycin (mTor). The mTor pathway plays a role in growth, survival and proliferation processes [8]. Based on these results we concluded that the inhibition of SLC1A5 seems to be cell line specific and that GPNA has to be used in high doses. Thus the clinical application of GPNA for NB treatment as a single agent is limited. Currently we are investigating combination therapies inhibiting amino acid supply in conjugation with cytotoxic drugs to exploit tumor cell dependency on amino acid supply for NB therapy.

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] Hediger, Matthias A.; Romero, Michael F.; Peng, Ji-Bin; Rolfs, Andreas; Takanaga, Hitomi; Bruford, Elspeth A. The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. Journal of Physiology 2004; 447(5): 465–468.

[6] Arriza JL, Kavanaugh MP, Fairman WA, Wu YN, Murdoch GH, North RA, Amara SG Cloning and expression of a human neutral amino acid transporter with structural similarity to the glutamate transporter gene family. J Biol Chem. 1993; 268(21):15329-32.

[7] Wasa, M.; Wang, H.-S; Okada, A. Characterization of L-glutamine transport by a human neuroblastoma cell line. In: AJP: Cell Physiology 2002; 282(6): C1246

[8] Fuchs, Bryan C.; Finger, Richard E.; Onan, Marie C.; Bode, Barrie P. ASCT2 silencing regulates mammalian target-of-rapamycin growth and survival signaling in human hepatoma cells. In: Am. J. Physiol., Cell Physiol 2007; 293(1):C55-63.

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

Roland Fried Wolfgang Rhode

Calculating energy-dependent limits on neutrino point source fluxes with stacking and unfolding techniques in IceCube

Fabian Clevermann
Experimentelle Physik 5
Technische Universität Dortmund
fabian.clevermann@udo.edu

The stacking method is a standard technique to search for possible neutrino sources in which several sources of the same type are bundled into one catalogue so that the possible signal from their different positions can be superimposed for data analysis. Flux limits can be placed on models assuming specific neutrino energy spectra for the source class. To improve this result and obtain separate flux limits at different energies, this work uses a new approach that combines the stacking with an unfolding of the energy spectrum of the neutrino events at the source positions of the investigated catalogue. Because the unfolding algorithm is independent of an assumed model or spectrum, the results are model independent. No sources have been discovered yet, so the number of potential signal neutrinos contributing to the unfolded result will be very small. The novel software TRUEE is used to obtain unfolding results with few events, which can then be used to infer limits on additional astrophysical contributions to the detected atmospheric neutrino flux. We present the resulting sensitivity for a given source catalogue with this method using data collected by the IceCube detector when it was partially constructed in its 59-strings configuration.

1 Stacking analysis

The likelihood method for a single source is described in [1]. To include multiple sources, the signal term gets modified to include M sources with theoretical weights $W(j)$, relative source efficiencies $R(j,\gamma)$ and the signal PDF's $S_{i,j}$ for the i^{th} event w.r.t. the j^{th} source:

$$L(x_s,\gamma,n_s)=\prod_i\left[\frac{\frac{n_s}{n_{\mathrm{tot}}}\sum_{j=1}^{M}W(j)R(j,\gamma)S_{i,j}(x_i,E_i,\gamma)}{\sum_{k=1}^{M}W(k)R(k,\gamma)}+\left(1-\frac{n_s}{n_{\mathrm{tot}}}\right)B(x_i,E_i)\right]$$

For this work, 1 000 events with the highest S/B ratios for each catalogue are selected for the unfolding. The number of chosen events should be small to have fewest background events in the sample as possible, but still have enough statistics for a reliable result.After several tests and trials a number of events of 1 000 was found to be the best compromise for this analysis.

1.1 Source catalogues

A catalog consists of several similar sources. This can be physical similarities, like the same accelerating process, as well as experimental similarities like being seen in a certain energy range from one experiment.

In these proceedings we describe, as an example, the analysis on the starburst galaxies catalogue.

- **Starburst galaxies:** This catalogue assembles galaxies with a high star formation rate (SFR). The high SFR results in a higher rate of super novae and supernova remnants (SNRs) which accelerate particles up to high energies (TeV). These particles are thought to create neutrinos in the interaction with dust clouds feeding the SFR. [2]

2 Unfolding analysis

For this analysis the upgoing events from the point source data sample of the 59-string configuration of IceCube was used [3]. It contains 43 339 events and has a 4.7% muon contamination according to Monte Carlo studies.

The unfolding is done by the software TRUEE [4]. To optimize the unfolding several configurations are evaluated on Monte Carlo data to determine the final settings.

2.1 Unfolding result

The unfolding technique gives as a result the energy distribution of the measured neutrinos. To verify the method on data, first we use a scrambled data set. By scrambling the events' right ascension in data it is possible to remove any influence of a signal in the sample. The result of this unfolding can be seen in figure 1. Although the last three bins have non-zero unfolded events which is used for limit calculations, the limit is compatible with zero to within one standard deviation.

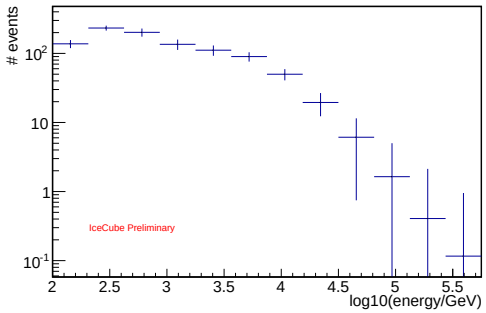


Figure 1: Unfolded result of the scrambled analyses. The x-axis shows the neutrino energy and on the y-axis the number of events are shown.

3 Sensitivity calculation

To calculate an upper limit for an excess above the expected background a profile likelihood method introduced by Rolke [5] is used. To get a robust background estimation, 2 000 different scrambled data sets were generated. The shape of these results for each bin follow a Gaussian distribution. Hence the model assuming the background Y is Gaussian distributed is used.

The sensitivity is calculated by using these background values as expected background as well as observed events with a certain confidence level. The calculated sensitivities for a 90% confidence level for an excess on top of the expected background can be seen in figure 2. Any neutrino flux, and not those following a unbroken power-law spectrum, exceeding the sensitivities in each energy bin will be excluded with a 90% confidence level.

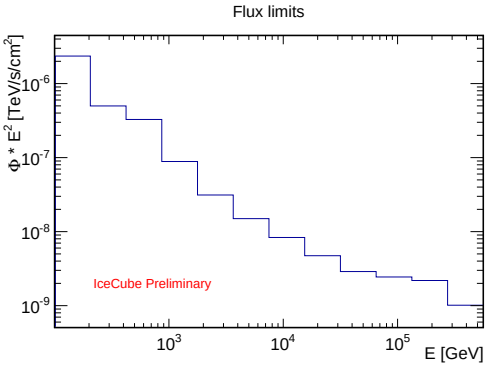


Figure 2: Unfolded sensitivities for the starburst catalogue. This result is calculated from the 90% upper limit for a pure background measurement, the effective area for the catalogue and the livetime of the detector.

4 Conclusion and outlook

This new method, combining a stacked analysis with an energy unfolding, yields sensitivities for different energies. These sensitivities are completely independent of an assumed model, as well as Monte Carlo simulation.

The limiting factor for the unfolding is the low number of events. With larger configurations of IceCube it is possible to increase the number of events from the regions of interest and improve the energy resolution. A better energy resolution in the high energy region would increase the sensitivity, due to the higher detection probability and the lower background rates.

References

[1] J. Braun, J. Dumm, F. De Palma *et al.*, Astroparticle Physics 29 (2008) 299 doi:10.1016/j.astropartphys.2008.02.007

[2] J. Dreyer *et al.*, (2009) 27 pages, arXiv:0901.1775

[3] IceCube Coll., paper 0550 these proceedings

[4] N. Milke, M. Doert *et al.*, Nuclear Instruments and Methods in Physics Research Section A 697 (2013) 133 doi:10.1016/j.nima.2012.08.105

[5] W. Rolke, A. M. López and J. Conrad, Nuclear Instruments and Methods in Physics Research Section A 551 (2005) 493 doi:10.1016/j.nima.2005.05.068

The MAGIC Monte Carlo chain and the use of Monte Carlos in a standard 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 of the MAGIC telescopes is given. Particular attention is paid to the application of Monte Carlo files in the MAGIC data analysis, which is also explained briefly.

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 this section the MC production chain for MAGIC in Dortmund is discribed. It is important that a big number of MC files are produced on the available clusters. So, the LiDO cluster (Linux cluster Dortmund) with 3328 CPUs and 215 TB storage and the PhiDo cluster (Physics cluster Dortmund) with 1200 CPUs and 200TB storage are used to produce and store the MC files. Before a MC file is ready, it has to pass through a chain of simulation programs.
At first *CORSIKA* [2] simulates the interaction of the primary particle with atoms and molecules of the atmosphere. Additionally the Cherenkov photons are simulated from their origin until they reach telescope level.
The next step is to simulate the mirrors of the telescope. This is done within the program *Reflector* [3]. The absorption of the Cherenkov photons on their way through the atmosphere to the mirrors and the reflection on the mirrors are simulated. The arrival time of the photons are calculated also.
The next step on the way of a Cherenkov photon is the arrival in the camera. *Camera* [1] simulates the electronics of the camera and the trigger of the photon.
Camera is the last program in the simulation chain. Afterwards the calibration chain starts and uses only programs of the Magic Analysis and Reconstruction Software (*MARS*) [4] like *Sorcerer*, *Star* and *SuperStar*.
Within *Sorcerer* the calibration and extraction of the simulated signal is executed.
Then the program *Star* performs the image cleaning and calculation of the Hillas parameter.
The last step in the automatic MC chain is the stereoscopic reconstruction of the shower done within *SuperStar*.
To ensure the passing through the MC chain, there are several scripts that check automatically if the files are produced correctly and none of the programs crashed during the calculation (see Figure 1). Besides there is a database in which all MC jobs and their state of production are enlisted.

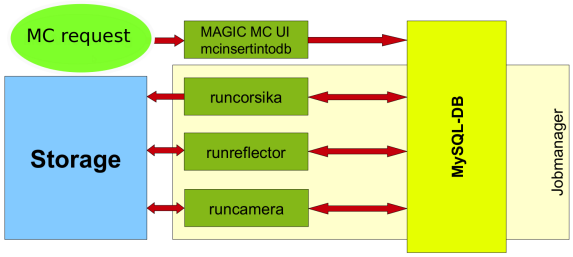


Figure 1: Schematic view of the MC production using runscripts and a MySQL database.

In the case of a MC request, the mcinsertintodb script inserts the request into the database. In the background there a script called jobmanager is always running that checks if there are new requests in the database. In that case, the runcorsika script

is called that starts the program Corsika. After the program has finished the produced Corsika files are stored and an entry in the database is made that a new Corsika file is available. Afterwards the jobmanager starts automatically the next program and so on.
In the last year several updates to this automatic production structure were made. A new Reflector as well as a new Camera program were installed which include the hardware changes made to the telescope in summer 2012. Furthermore the Corsika program had to be updated and a lot of changes to the inputcards of Sorcerer and Star were made. Now, everything is running stable and the mass production of gamma Monte Carlos for the MAGIC collaboration is performed in Dortmund and every collaboration member can access the data via grid to analyse real data.

3 Analysis

At the moment a study of the standard analysis of data is ongoing with help of Crab data provided during a software school in february 2013. The analysis of MAGIC data starts with downloading data from La Palma and downloading convenient MC data. Afterwards an analysis is performed with help of the *MARS* framework. This program package offers a lot of programs and scripts to perform the full analysis of MAGIC data, with the objective to get a lightcurve and a spectrum of a source. It also offers a program to calculate a sky map.
But at first the gamma/hadron separation and an energy estimation is performed. For this a Random Forest is trained within the program Coach and applied to the data within the program Melibea. Figure 2 shows that the energy reconstruction is working fine.

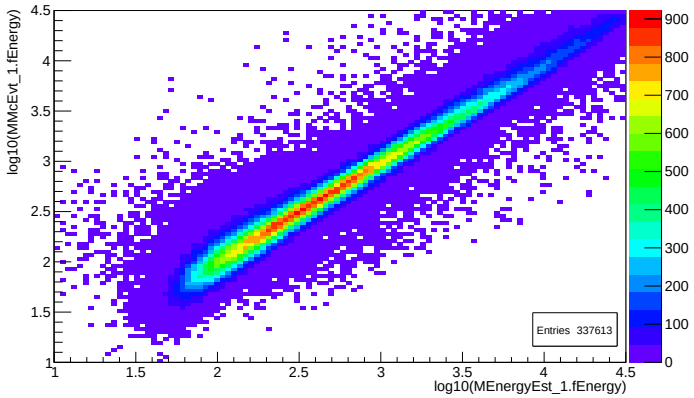


Figure 2: Comparison of the estimated energy computed by Melibea (MEnergyEst_1.fEnergy) to the true MC energy (MMcEvt_1.fEnergy).

In order to show this Melibea was applied to MC data with known energy.

4 Conclusion and outlook

This TexReport gave a short résumé about the MC production for the MAGIC experiment in Dortmund and pointed out the importance of Monte Carlo data to the analysis of real data. Afterwards the way to analyse data was shown and an example of the application of Monte Carlos in the energy estimation was explained.
The next step in this analysis chain is the unfolding of the spectrum. This will be performed with the software TRUEE and after that a comparison will be made to the new developed software DSEA [5], developed in this SFB876.

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.

[5] T. Ruhe, M. Schmitz, T. Voigt, and M. Wornowitzki. DSEA: A Data Mining Approach to Unfolding. *Proceedings of the ICRC 2013*, 2013.

Development of a Monte-Carlo simulation for lepton propagation

Jan-Hendrik Köhne
Experimentelle Physik 5
Technische Universität Dortmund
jan-hendrik.koehne@tu-dortmund.de

IceCube is a large scale neutrino detector located at the South Pole. The one cubic kilometer detector volume consists of the South Pole ice, which has excellent optical properties [1]. IceCube uses the physical effect, that neutrinos interacting with a media produce charged leptons such as muons, electrons and taus. These leptons propagate through the detector and emit Cherenkov light, which is detected by high sensitve photon sensors [3].
The complexity to analyse the data is, that leptons coming from the atmosphere produce a similar signal in the detector. The outcome of this is a huge amount of background which overlaps the neutrino signal. The ratio of signal to background is about one to a million.
To analyse the data and to find neutrino signals, Monte-Carlo simulations are essential.

1 Simulations in IceCube

The IceCube Monte-Carlo chain consists of several programs, each of which simulate a different part of the experiment. These programs can be classified into generators, propagators and hardware simulations.

Generators create the particles. In IceCube the program CORISKA [4] is used to simulate atmospheric leptons. To generate the neurinoflux through the earth the program NuGen is used.

Propagators take the generated particles and simulate their behaviour while propagting through the detector. The currently used propagation software is PROPOSAL

(**PR**opagator with **O**ptimal **P**recision and **O**ptimized **S**peed for **A**ll **L**eptons) the successor of MMC (Muon Monte Carlo) [2].

Hardware simulations describe the reaction of the different detector components such as photon sensors when a particle propagates through the detector.

1.1 PROPOSAL

As mentioned above PROPOSAL is the successor of MMC and currently the main propagation program in the IceCube Monte-Carlo chain. From physical point of view MMC was a good choice for simulating leptons and monopoles. It has been tested for several years in astroparticle physics for example in the AMANDA experiment which was the first neutrino detector at the South Pole. But several technical problems such as homogeneity of the Monte-Carlo chain (MMC is written in Java, everything else in C++) and maintainability made a revision of MMC necessary.

PROPOSAL provides the possibility to propagate leptons and monopols through any type of media. PROPOSAL takes the most important energy loss mechanisms into account:

- Ionisation
- Bremsstrahlung
- Electron positron pair production
- Photonuclear interaction

2 Status and Plans

The first release of PROPOSAL is completed and published [5]. Also a detailed documentation of the C++ code was included. The excellent agreement of PROPOSAL and MMC is shown in figure 1.

PROPOSAL is now the default lepton propagator in the IceCube Monte-Carlo chain. Studies to estimate the influence of different energy loss parametrizations are in progress.

The program structure of the first release of PROPOSAL is very similar to the Java program MMC. To get rid of the Java like structure PROPOSAL was designed in a complete new way to improve performance and the maintainability. A comparison of the needed computing time for the two releases to propagate a particle through ice is shown in figure 2.

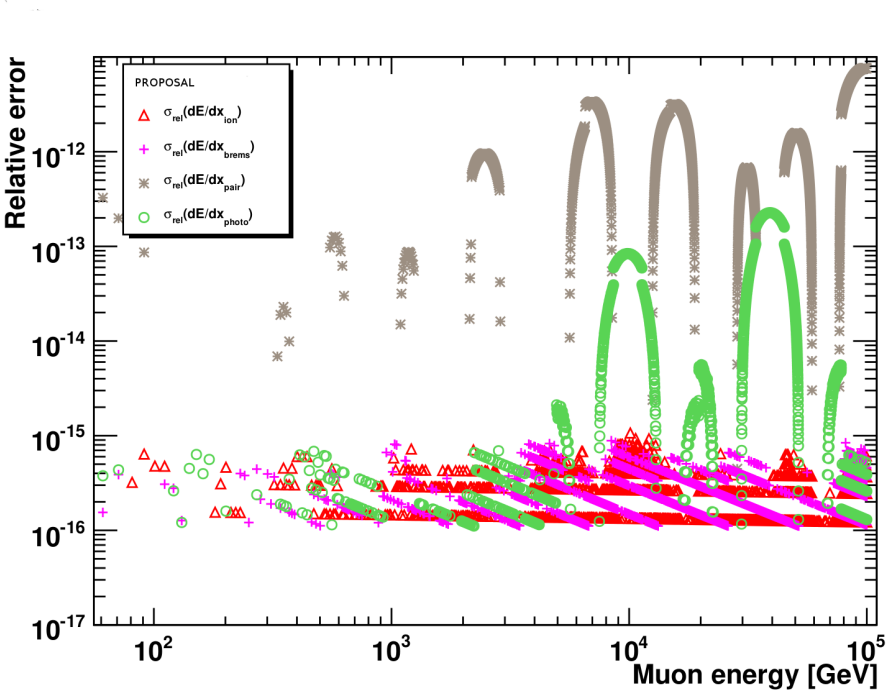


Figure 1: The relative error of PROPOSAL and MMC is shown using the example of the energy loss per distance. Due to a different double precision in Java and C++ the result of MMC and PROPOSAL are slightly different.

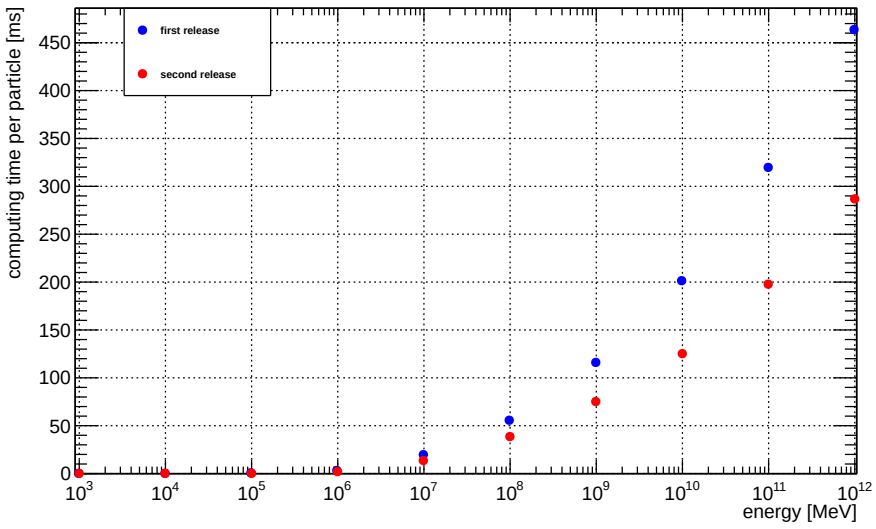


Figure 2: Comparison of the first and second release of PROPOSAL. Shown is needed computing time to propagate a particle through ice until they decay.

Because parallelization will speed up the simulation a lot and therefore save computing time and money, it is planed to implement PROPOSAL for GPUs in one of the next releases. Here the first tests were done by Tomasz Fuchs and are very promissing as shown in figure 3.

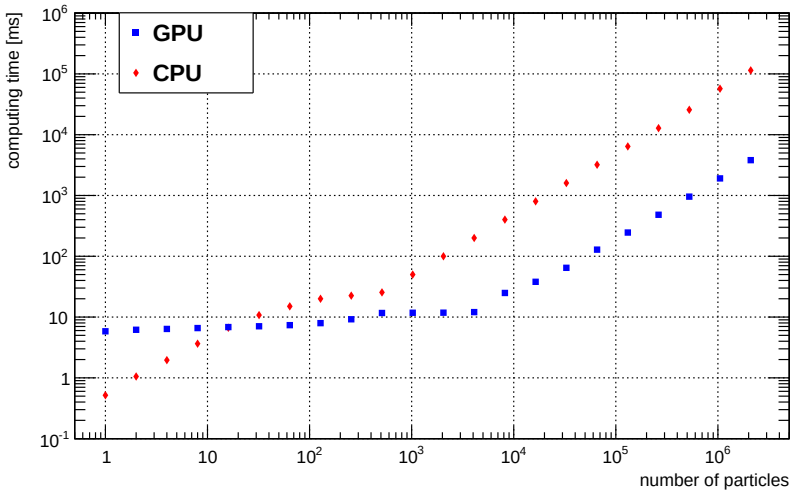


Figure 3: The needed computing time of the first step in the propagation of particles. A Comparison of CPUs and GPUs is shown. Parallelization can speed up the propagtion by a factor of 30.

References

[1] M. Ackermann, J. Ahrens, X. Bai, et al. Optical properties of deep glacial ice at the South Pole. *Journal of Geophysical Research (Atmospheres)*, 111:13203–+, July 2006.

[2] D. Chirkin and W. Rhode. Propagating leptons through matter with Muon Monte Carlo (MMC). *ArXiv High Energy Physics - Phenomenology e-prints*, July 2004.

[3] F. Halzen. IceCube Science. *Journal of Physics Conference Series*, 171(1):012014–+, June 2009.

[4] D. Heck, J. Knapp, J. N. Capdevielle, G. Schatz, and T. Thouw. CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers. Technical Report FZKA 6019, Forschungszentrum Karlsruhe, 1998.

[5] J.-H. Koehne, K. Frantzen, M. Schmitz, T. Fuchs, W. Rhode, D. Chirkin, and J. Becker Tjus. Proposal: A tool for propagation of charged leptons. *Computer Physics Communications*, 184(9):2070 – 2090, 2013.

Identification of unassociated gamma-ray sources of the 2FGL catalog

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

A fascinating type of sources emitting cosmic rays is the class of Active Galactic Nuclei (AGN), which are capable of producing the highest possible energies in the universe ever recorded. It is interesting that there are various types of AGN. These can be classified according to the spectral behavior in different energy ranges. With the help of modelling this behavior conclusions about the acceleration mechanisms in AGNs can be drawn.

Several catalogs of the different experiments - satellites and ground-based detectors - summarize the detected sources in the experiment's specific energy range. This work is focussed on the second *Fermi*-LAT catalog (2FGL) because it is based on a sky survey which is scanning the whole sky in the gamma-ray energy range. Given that the emission of gamma-rays is an essential indicator for AGNs, this catalog was chosen as starting point.

About one third of all listed sources in the 2FGL catalog are so-called unidentified gamma-ray sources because their source type could not be determined so far. In order to identify possible AGN candidates within this group, multivariate methods were applied in [4]. This report summarizes the first steps towards an identification of these AGN candidates.

Introduction

The Large Area Telescope on board of the *Fermi Gamma-ray Space Telescope*, the *Fermi*-LAT, is an instrument which scans the sky in the very high energy range from 100 MeV to 100 GeV. The second *Fermi*-LAT catalog (2FGL) [8] containing the observation results

of the 1873 gamma-ray emitting sources was published in 2012. 1297 of these are associated with a known source type, leaving 576 unassociated.

With the application of multivariate methods an approach towards classifying the yet unassociated sources was started. The associated sources were separated into two groups, namely "AGN" and "non-AGN". The properties of the labelled sources are used to train and test the performance of the two algorithms which were used. The random forest method and the neural networks were used separately inside the RapidMiner framework. In the end their combined confidence is taken to classify the unassociated sources. Every unassociated source with a confidence above 80% to be an AGN is classified as AGN candidate (see [4] for more information).

With the above explained method 231 of the 576 unassociated sources are classified as AGN candidates [4]. For these sources it is necessary to also consider their behavior in other energy ranges to obtain a Spectral Energy Distribution (SED). A SED shows the overall spectral behavior of a source and is essential for identifying the source type and determining the processes to produce this specific radiation.

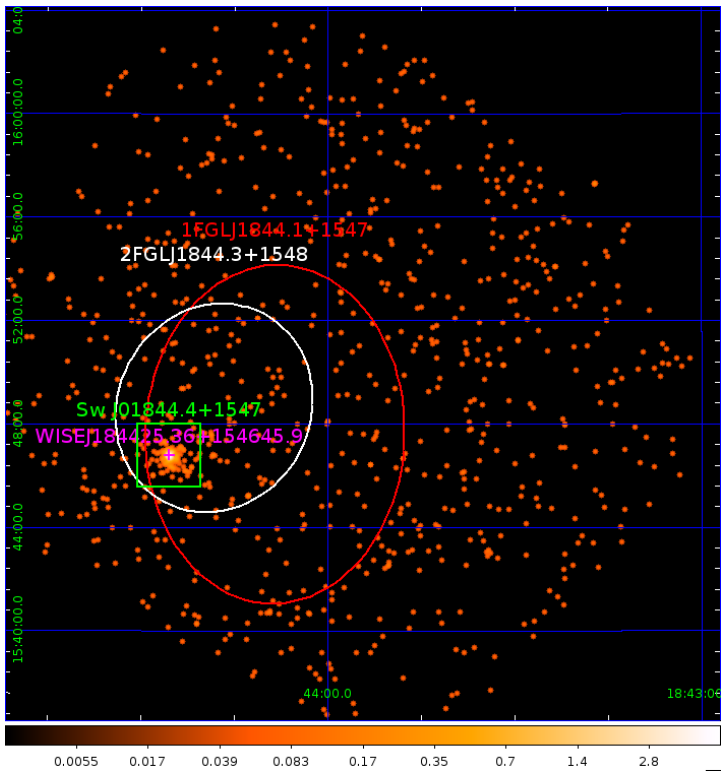
Procedure of source identification

Investigations of the spectral features in other energy ranges, e.g. X-ray, infrared, optical, and radio, are essential to build the SED. Therefore an archival search for these data was started. It is important to look for the correlations of the source positions detected by different telescopes to be confident that the radiation is related to the same source.

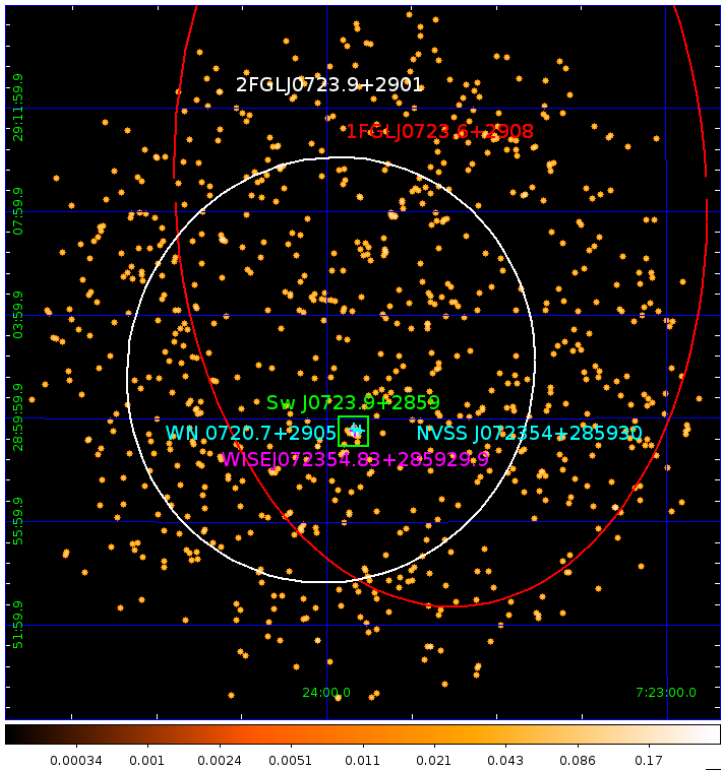
At first the images of the *Swift*-XRT, the X-ray Telescope on board of the *Swift* satellite, are analyzed. With the analysis software HEASoft [1] point sources are detected in the image. The box with the maximum signal-to-noise ratio is used to calculate the position and the intensity of the detected point sources [2].

In [5] the behavior in infrared energies of the unassociated gamma-ray sources of the 2FGL catalog was investigated. Therefore the data of *WISE*, the *Wide-field Infrared Survey Explorer* was analyzed. Blazars, a certain subclass of AGNs, have distinct features in the infrared energies. The features of classified AGNs which have an infrared counterpart were inspected for analogies in this energy range. Then the infrared features of the unassociated gamma-ray emitters were examined and sources with consistent features could be classified as AGN candidates. The same procedure has been carried out for radio energies with the observations of the Westerbork Synthesis Radio Telescope (WSRT) and the Very Large Array (VLA). The data of their sky surveys WENSS [9] and NVSS [7] were examined for the radio properties of blazars and then compared to the radio properties of the unidentified gamma-ray sources to find AGNs [6].

All available data for the sources 2FGL1844.3+1548 and 2FGLJ0723.9+2901 are shown in figures 1(a) and 1(b) as an example. The basic image are observations of *Swift*. The detected point sources are displayed with the above mentioned error box. On top of that the observations of *Fermi*, *WISE*, WSRT and VLA are plotted as well.



(a) Image of 2FGL1844.3+1548.



(b) Image of 2FGL0723.9+2901.

Figure 1: Images of the X-ray data of *Swift*-XRT [3] and the detected point source with error box (green) are displayed. Included are the gamma-ray observations of *Fermi*-LAT in red (1FGL, first *Fermi*-LAT catalog) and white (2FGL) with the ellipse containing 95% of the signal. The infrared observations of *WISE* in magenta are included in (a) and (b) and the radio observations of WSRT (WENSS) and VLA (NVSS) in cyan are included for (b).

Outlook

Further research to look for source position correlation with optical data is planned. Also the archival search in X-ray, infrared, and radio wavelengths will be extended to other telescopes.

It is planned to perform the same procedure described above to associate and identify more of the yet unassociated sources of the 2FGL catalog. Another interesting source group are pulsars which are so far represented by 108 of the 1873 sources [4]. To search e.g. for the few but very interesting binary systems an outlier detection might be taken into consideration.

References

[1] NASA's HEASARC: Software - HEASoft, http://heasarc.nasa.gov/lheasoft/, September 2013.

[2] NASA's HEASARC: Software - XIMAGE User's Guide, http://heasarc.nasa.gov/docs/software/lheasoft/xanadu/ximage/manual/node5.html, September 2013.

[3] Swift homepage, http://www.swift.ac.uk/, October 2013.

[4] M. Doert and M. Errando. High confidence AGN candidates among unidentified Fermi-LAT sources via statistical classification. *Proceedings of the 33rd International Cosmic Ray Conference, Rio de Janeiro*, June 2013.

[5] F. Massaro et al. Unveiling the nature of unidentified gamma-ray sources. II. Radio, infrared, and optical counterparts of the gamma-ray blazar candidates. *The Astrophysical Journal Supplement Series*, 206:13 (15pp), June 2013.

[6] F. Massaro et al. Unveiling the nature of unidentified gamma-ray sources. III. Gamma-ray blazar-like counterparts at low radio frequencies. *The Astrophysical Journal Supplement Series*, 207:4 (15pp), July 2013.

[7] J. J. Condon et al. The NRAO VLA Sky Survey. *The Astronomical Journal*, 115:1693–1716, May 1998.

[8] P. L. Nolan et al. Fermi Large Area Telescope Second Source Catalog. *The Astrophysical Journal Supplement Series*, 199:31 (46pp), April 2012.

[9] R. B. Rengelink et al. The Westerbork Northern Sky Survey (WENSS) - I. A 570 square degree Mini-Survey around the North Ecliptic Pole. *Astronomy and Astrophysics Supplement Series*, 124:259–280, August 1997.

DataMining for IceCube

Florian Scheriau
Experimentelle Physik 5
Technische Universität Dortmund
florian.scheriau@tu-dortmund.de

The IceCube neutrino telescope is located beneath the glacial ice at the South Pole. IceCube data shows a unfavourable signal to background ratio for atmospheric muon neutrinos. With methodes developed within the SFB 876 it was possible to create a sample of atmospheric neutrinos with very high purity and efficiency. With this sample it was possible to measuare the energy spectrum of atmospheric muon neutrinos up to 3.3 PeV for the first time.

In December 2010 the IceCube neutrino detector was completed with the deployment of the last of 86 strings. A schematic picture of the detector can be found in figure 1. The strings are cables melted in the glacial ice of the South Pole up to the depth of 2450 m. The spacing between the strings is 125 m. Between 1450 m and 2450 m 5160 Digital Optical Modules (DOMs) are mounted on the strings. This configuration makes IceCube with an instrumented volume of 1km^3 the worlds biggest neutrino telescope. DOMs measure the radiation of muons going through the detector. With the measured radiation over 2000 observables are calculated which describe every event in the detector. [6], [1], [2]

In the interaction between cosimc rays and the nuclei of the Earth's atmosphere extended air showers are produced. In these air showers pions and kaons are produced wich decay to neutrinos. These neutrinos can than be measured with IceCube. [8] The goal of this analysis is to separate the atmospheric neutrinos from the dominant background. This is possible by training a random forest on simulated signal and background. The pure atmospheric neutrino data can then be used unfold theenergy spectrum of the atmospheric neutrinos. [4], [5], [7]

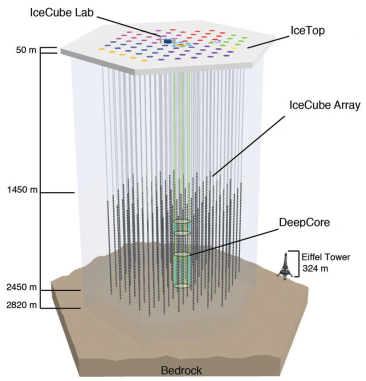


Figure 1: The schematic structure of the IceCube detector. The picture shows the 86 Strings in light grey and the 5160 DOMs in dark grey. [10]

In order to create a sample with high puruty and efficncy we have started by reducing the background with two very simple cuts wich eliminate hughe amouts of backgroud while keeping most of the wanted atmospheric neutrinos. The two cuts together reject over 80% of all background while keepig over 80% of the wanted signal. For the so generated data a set of 25 observables have been derived using the MRMR feature selection algorithm [3]. With this set of observables one can optimize a random forest. The random forest confidence distribution for data and simulation is shown in 2.

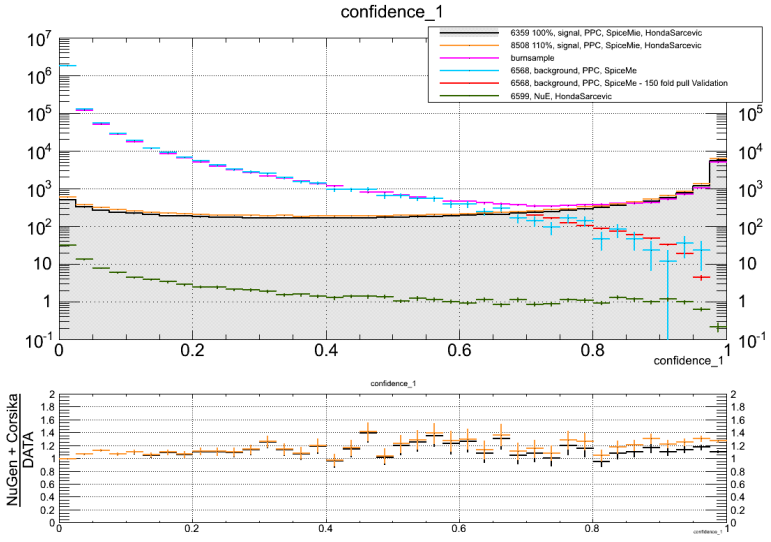


Figure 2: confidence distribution for the IC86 detector configuration

The ratio between data and simulations in figure 2 clearly shows that the simulation describes data well and that the model is capable to spearat between our wanted signal

and background.

With the so derived sample it was possible to unfold the atmospheric neutrino spectrum up to 3.3 PeV as shown in figure 3.

Due too the efficienct analys chain it was possible to adapt the analysis to the more recent IC86 detector configuration and get an separated dataset an a very short time scale. The resulting confience distitribution is shown in figure 4.

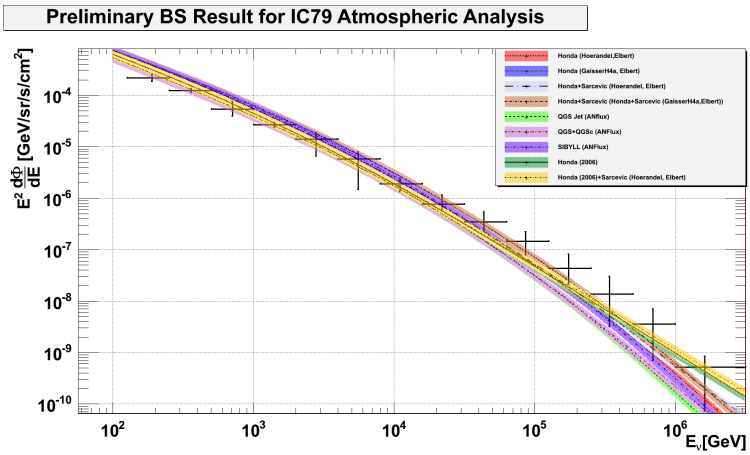


Figure 3: The mesuared atmospheric neutrino spectrum compared to theory

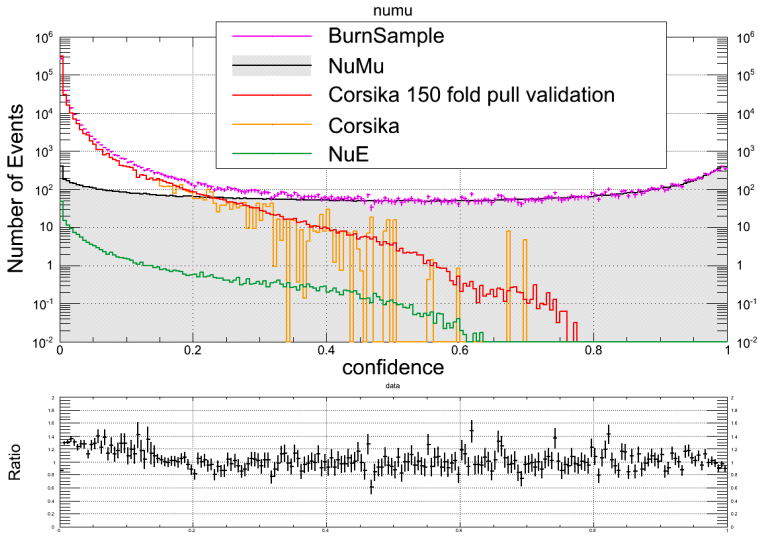


Figure 4: confidence distribution for the IC86 detector configuration

References

[1] R. Abbasi, Y. Abdou, T. Abu-Zayyad, J. Adams, J. A. Aguilar, M. Ahlers, K. Andeen, J. Auffenberg, X. Bai, M. Baker, and et al. Calibration and characterization of the IceCube photomultiplier tube. *Nuclear Instruments and Methods in Physics Research A*, 618:139–152, June 2010.

[2] R. Abbasi, Y. Abdou, T. Abu-Zayyad, J. Adams, J. A. Aguilar, M. Ahlers, K. Andeen, J. Auffenberg, X. Bai, M. Baker, and et al. Measurement of the atmospheric neutrino energy spectrum from 100 GeV to 400 TeV with IceCube. *Physical Review D*, 83(1):012001–+, January 2011.

[3] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185–205, April 2005.

[4] A. Gazizov and M. Kowalski. ANIS: High energy neutrino generator for neutrino telescopes. *Computer Physics Communications*, 172:203–213, November 2005.

[5] D. Heck, J. Knapp, J.N. Capdevielle, G. Schatz, and T. Thouw. A monte carlo code to simulate extensive air showers. Technical Report 6019, Forschungszentrum Karlsruhe GmbH, Karlsruhe, 1998.

[6] Henrik Johansson. *Searching for an Ultra High-Energy Diffuse Flux of Extraterrestrial Neutrinos with IceCube 40*. PhD thesis, Stockholm University, 2011. ISBN 978-91-7447-290-5.

[7] N. Milke, M. Doert, and W. Rhode. Solving inverse problems with truee: Examples in astroparticle physics, 2012.

[8] K. Nakamura et al. Review of particle physics. *J.Phys.G*, G37:075021, 2010.

[9] T. Ruhe, K. Morik, and Schowe. B. Data mining icecube, 2011.

[10] The IceCube Collaboration. completedarraynoamanda.jpg. `http://icecube.wisc.edu/gallery`, 2011.

Spectral Reconstruction of IceCube-79 Data

Martin Schmitz
Experimentelle Physik 5
Technische Universität Dortmund
martin.schmitz@tu-dortmund.de

The determination of the atmospheric neutrino flux spectrum is important for the neutrino astronomy as its distribution at high energies can shed light on the predicted flux of extragalactic neutrinos. IceCube is a cubic kilo meter large neutrino telescope located at the geographic South Pole and is well suited for the detection of high energy neutrinos. IceCube is not measuring energy of atmospheric neutrinos directly. Instead it is measuring the light of secondary particles produced by neutrino induced muons. Therefore the estimation of an energy distribution of neutrinos is a highly non-trivial task. This task is traditionally solved by unfolding algorithms with respect to regularization. In this report an unfolding of an atmospheric neutrino spectrum using TRUEE is presented. Afterwards this spectrum is compared to the result of the novel unfolding approach DSEA.

Atmospheric neutrinos are produced in interactions of cosmic rays with Earth's atmosphere. On the one Hand they are the dominant background in a search of astrophysical neutrino sources. On the other hand they are a interesting field of study in itself. The neutrinos are detected with the IceCube neutrino telescope located at the geographic south pole. In the used configuration it contains 79 strings and is able to detect neutrinos above energies of 10^6 GeV [3].
To analyse a neutrino spectrum a rejection of background is necessary. The ratio of signal to background before separation is worse than 10^{-3}. Thus dedicated algorithms are needed to solve the problem. With use of the RapidMiner framework and a Random Forest [4], a purity of $(99.5 \pm 0.3)\%$ can be achieved. After this separation a sample containing 6938 events is collected in 32.5 days of detector uptime. This sample is studied using the unfolding software TRUEE [8].

First, observables need to be chosen which are showing a good correlation to the neutrino's primary energy. The chosen observables are reconstructed energy [2], reconstructed track length and number of reconstructed direct photons in a given time window. TRUEE provides different methods to find solid unfolding parameters like regularization

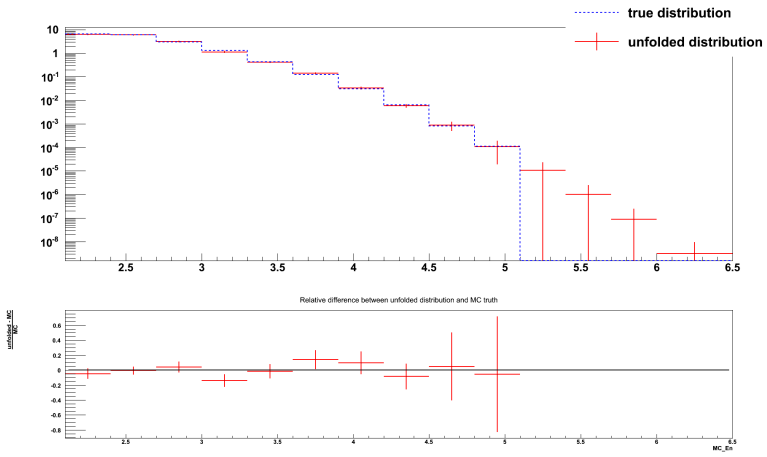


Figure 1: Test mode unfolding for IceCube data. Since this is a test for unfolding of 10% of the available data some of the bins are empty.

strength. One of them is the so called *test mode*. The available simulated events are divided into two parts. One for determining the response matrix and one to test the unfolding. Thus it is possible to compare the unfolding result with the truth which is known on simulated data. The result for this is depicted in figure 1.

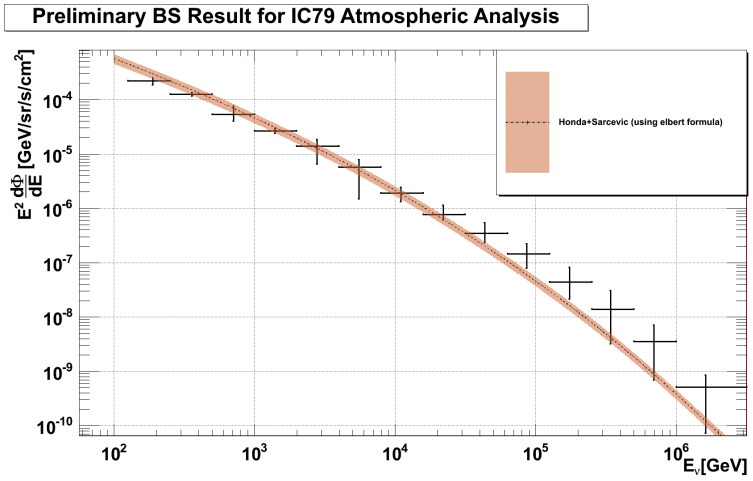


Figure 2: Unfolded flux of 10% of the IC79 data. The errorbars include different systematic errors introduced by the detection method. The presented reference flux is based on [7] and [5].

After some further tests this can be applied to the real data. The result is a spectrum

of neutrino energy. Because of limited acceptance this needs to be corrected. In figure 2 the acceptance corrected result is depicted with an E^2 scaling. For a comparison a theoretical flux extrapolated from [7] and [5] is shown.

DSEA [1] is a novel approach to unfold spectra. To estimate the spectrum of a sought-after variable x this variable is binned in a appropriate fashion. These bins serve as label in a multinomial classification problem. The only requirement of the classificator besides the ability to solve multi nominal classification problem is to calculate a confidence for each bin instead of just giving a prediction. The prediction is usually not well suited to construct a spectrum because of the ill-posedness of the problem. Instead the confidence distribution is summed up to estimate the spectrum.
This technique has some advantages compared to traditional unfolding algorithms. Besides the capability of using arbitrary numbers of observables, DSEA preserves individual event information during the unfolding process. Thus it is possible to analyse the spectra with respect to other observables. The requirement of those observables is, that they are measured with a high precision compared to the used binsize. Good examples for this in IceCube are detection time and zenith angle. DSEA was applied on the previously mentioned 10% of the IC79 data. To do so, a MRMR feature selection [9] was performed on the simulated data. The top 14 observables were selected and a forest of extremely randomized trees [6] was used to determine the spectrum. In figure 3 a comparison be-

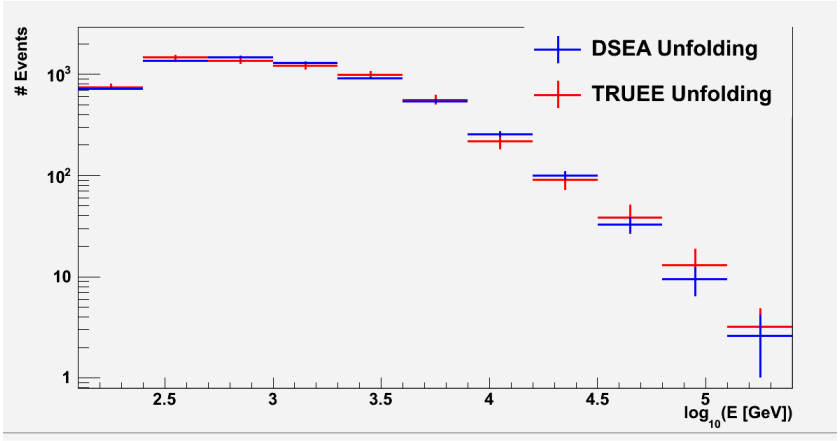


Figure 3:

tween the unfolded DSEA spectrum and the spectrum unfolded with TRUEE is shown. In general a good agreement can be seen.
In future DSEA will be used to unfold zenith and time depended neutrino fluxes measured by the full IceCube detector in the 86 string configuration.

References

[1] *DSEA: A Data Mining Approach to Unfold*, 2013.

[2] M. G. Aartsen et al. Energy Reconstruction Methods and Performance in the IceCube Neutrino Detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2013.

[3] M.G. Aartsen et al. First observation of PeV-energy neutrinos with IceCube. *Phys.Rev.Lett.*, 111:021103, 2013.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] Rikard Enberg, Mary Hall Reno, and Ina Sarcevic. Prompt neutrino fluxes from atmospheric charm. *Phys.Rev.*, D78:043005, 2008.

[6] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, April 2006.

[7] M. Honda, T. Kajita, K. Kasahara, S. Midorikawa, and T. Sanuki. Calculation of atmospheric neutrino flux using the interaction model calibrated with atmospheric muon data. *Phys. Rev. D*, 75:043006, Feb 2007.

[8] N. Milke, M. Doert, S. Klepser, D. Mazin, V. Blobel, et al. Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics. *Nucl.Instrum.Meth.*, A697:133–147, 2013.

[9] H. Long Peng and C. F. 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):1226–1238, 2005.

Development of a preprocessing Analysis Software for the Cherenkov Telescope 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. When a high energy astroparticle impacts the atmosphere a large amount of secondary particles is created. They form an air shower which emits Cherenkov light which is detected by the photosensors of the telescope. FACT, as the first of its kind, uses Geiger-mode Avalanche Photodiodes (G-APD) as photosensors, instead of the common Photo Multiplier Tubes. G-APDs promise more robustness, a lower bias voltage and higher photon detection efficiency at lower costs [5, 8].
To reduce the large amount of data to a reasonable level a preprocessing analysis software is being developed. First it applies some necessary calibration steps. Then the number of detected photons and their arrival times are extracted for each photosensor (pixel). An image cleaning algorithm is applied to the resulting image. The last step contains an image parameterization. The calculated image parameters are used for further analysis steps.

1 Introduction

FACT [1] is located on the Canary Island of La Palma on the mountain Roque de los Muchachos at an altitude of about 2200 m and records data since the 11th October 2011. The voltages in the 1440 pixels are measured with a sampling rate of 2 GHz. When the trigger criteria of the telescope is exceeded a time window of 150 ns is digitized. Therefore

the raw data of one single event contains $1440 \times 300 = 432000$ voltages. To analyse the raw data the preprocessing software PARFACT (**P**reprocessing **A**irshower for **R**apidMiner for **FACT**) is being developed [7].

2 PARFACT

The first step of the analysis is the so called DRS-Calibration. In FACT a data acquisition chip, the DRS4 chip [6], and an analog-digital-converter (ADC) are used to store and digitize the voltages in the pixels. Each combination of DRS4 chip and ADC has an individual offset and conversion factor. These differences between the pixels are compensated by applying the DRS-Calibration to the raw data. The required calibration constants are determined using previous calibration runs.
The calibrated data is shown in figure 1. To calculate the number of photons and their arrival times in the pixels, several features of the pulse in the timeline are calculated. The position of the maximum (blue line in figure 1), the integral around the maximum (green borders in figure 1) and the position of the prior rising edge (red line in figure 1). The integral depends linear on the number of photons registered in the pixel [3]. The arrival time of these photons is determined by the position of the rising edge.

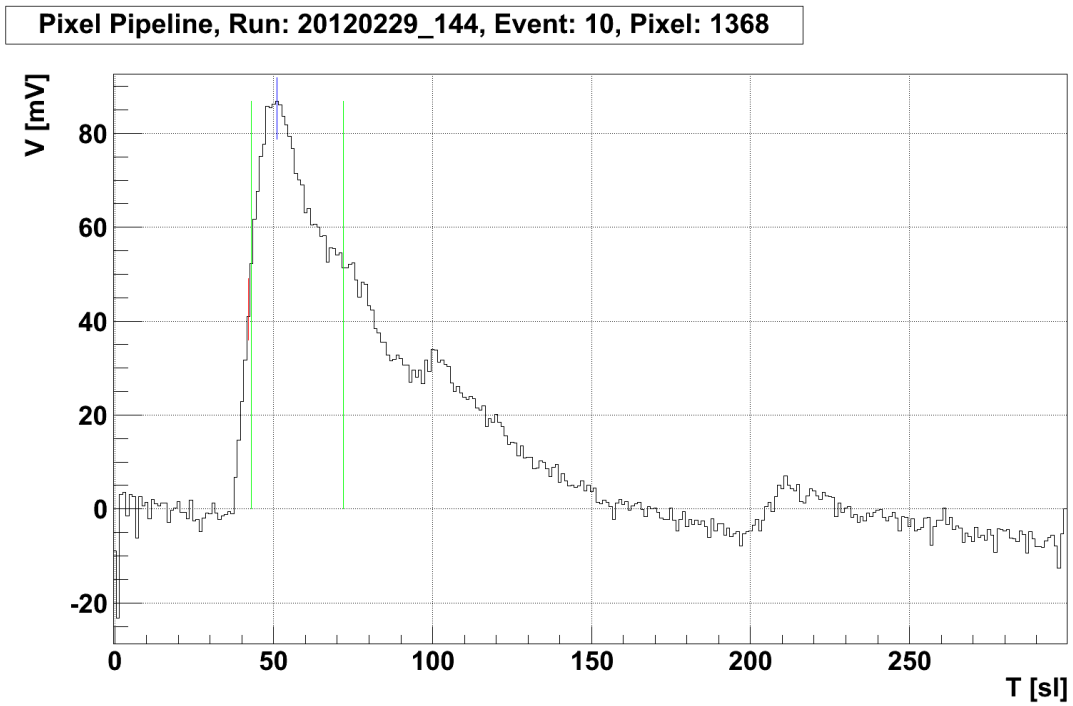


Figure 1: Pixel timeline after DRS-Calibration. A pulse in the timeline, induced by incoming photons can be seen. Several features of the pulse are marked.

The number of photons in all pixels yield to an image of the field of view from the telescope (see figure 2, left image). To extract the image of the triggered air shower an image cleaning algorithm is applied to the image. In PARFACT a so called Two-Level Cleaning is used. First a higher core threshold is applied to all pixels. On neighbors of remaining clusters (at least two pixels) another lower neighbor threshold is applied. At last pixels with an arrival time much higher or lower than the median of the remaining arrival times are also removed. The resulting image of the air shower is shown in figure 2, right image.

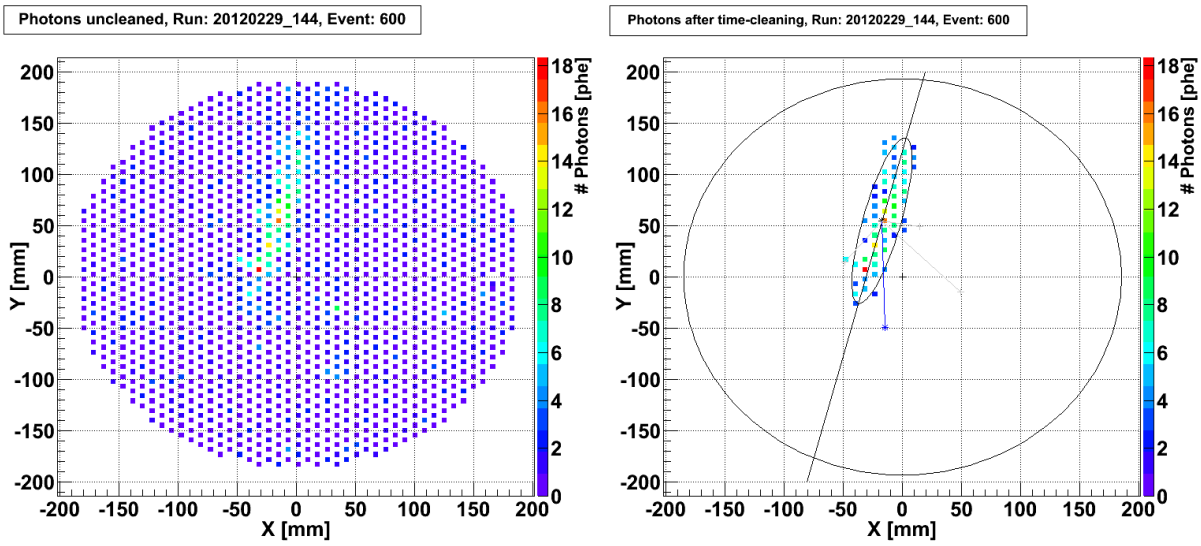


Figure 2: Image of the field of view of the telescope. A triggered air shower can be seen. Left: Uncleaned Event. Right: Event after application of the cleaning algorithm.

Then the image of the air shower is parametrized. Thereby the moments of the weigthed geometrical distributions of the shower pixels are used to determine an ellipse parametrization of the shower (as seen in figure 2, right image) [4]. In addition several other parameters are calculated, for example size, the number of photons of all shower pixels or the concentration, thus the maximum number of photons in all shower pixels divided by the size.
The calculated parameters are used for further analysis steps, thus the gamma hadron separation and the unfolding of the energy distribution of the primary particles.

3 Status and Outlook

The above mentioned analysis steps are already implemented in PARFACT and the resulting parameters are used for further analyses. The algorithms used in PARFACT are

included in the application fact-tools of the streams framework. The streams framework was developed by Christian Bockermann within the scope of the SFB 876, sub-project C1 [2].
It can be easily extended by adding additional operators with newly developed methods to the analysis chain. In cooperation with the sub-project B2 an improved cleaning algorithm using an active contour model is developed to improve the quality of the cleaning algorithm. Additional image parameters developed within the scope of the sub-project C3 might improve the quality of the gamma hadron separation and the unfolding of the energy distribution.
The streams framework also arises the possibility to analyse the data online in an embedded system.

References

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

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

[3] Jens Björn Buß. FACT - Signal Calibration, Gain Calibration and Development of a Single Photon Pulse Template for the FACT Camera. Diploma thesis, Technische Universität Dortmund, June 2013.

[4] A. M. Hillas. Cerenkov light images of EAS produced by primary gamma. In F. C. Jones, editor, *19th International Cosmic Ray Conference ICRC, San Diego, USA*, volume 3, pages 445 – 448. International Cosmic Ray Conference, August 1985.

[5] Thomas Krähenbühl et al. Geiger-mode Avalanche Photodiodes as Photodetectors in Cherenkov Astronomy. In *31st International Cosmic Ray Conference, Łódź, Poland*. International Cosmic Ray Conference, July 2009. Published online: http://icrc2009.uni.lodz.pl/proc/html.

[6] Stefan Ritt. Design and Performance of the 6 GHz Waveform Digitizing Chip DRS4. In *IEEE Nuclear Science Symposium Conference Record*, volume N11-8, pages 1512 – 1515. IEEE - Institute of Electrical and Electronics Engineers, 2008.

[7] 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.

[8] Quirin Weitzel et al. A Novel Camera Type for Very High Energy Gamma-Astronomy. In *31st International Cosmic Ray Conference, Łódź, Poland*. International Cosmic Ray Conference, July 2009. Published online: http://icrc2009.uni.lodz.pl/proc/html.

Signal-Background Separation of Monte Carlo Simulations and real data with RapidMiner for FACT

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 applied to a set of Monte Carlo Simulations and real data recorded with FACT using the data mining environment RapidMiner. In this report the results of the training and testing of the built model 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 consists of secondary relativistic charged particles. This shower emits a blueish light, the so-called Cherenkov light [5].
FACT is the first IACT which uses Geiger-mode Avalanche PhotoDiodes (G-APDs) instead of photomultipliers as photosensors to detect this light. It is located on the Canary Island La Palma at 2200 m a.s.l. and was commissioned for the first time on 11th October 2011 [3]. 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 [4], which is available in the data mining environment RapidMiner [2]. The model is trained on Monte Carlo simulations for FACT, which were produced by CORSIKA [1], and real hadronic background data, which were recorded in a sky region without an expected gamma source. For this purpose gamma Monte Carlo and real data proton showers were used and further processed by the analysis software PARFACT [6]. After data processing quality cuts were applied to each data set to filter out nonphysical events. For training and testing the RF model 19000 events were used. The Random Forest was

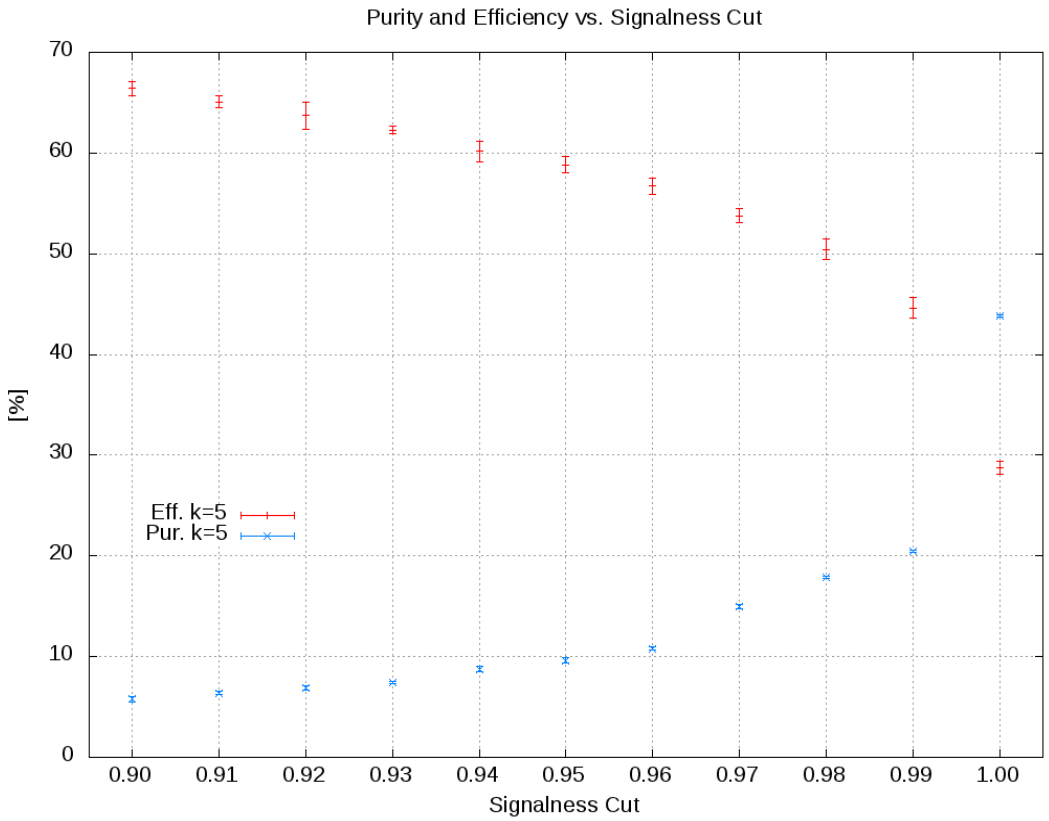


Figure 1: Displayed are the results of a trained Random Forest model on Monte Carlo simulation and real data with five randomly chosen attributes. Blue data points indicate the weighted purity increasing with a higher signalness cut, while red data points show the decreasing efficiency.

trained with parameters which describe the shower images and thus allow to distinguish between gamma showers and hadronic showers. For the RF 500 trees were built and five randomly chosen features taken out of a total amount of six parameters. Furthermore a signalness cut from S=0.9 to S=1.0 as well as a five fold cross validation were applied to the RF model to determine statistical mean and error values and to estimate the stability of the model. In Fig.1 the first results of the testing on the simulated and real data are

shown. The red data points show the efficiency against the signalness cut. Here the efficiency can be described as

$$E = \frac{N_{tp}}{N_S}$$

whereas N_{tp} is the amount of true positive classified events after the signalness cut and N_S the total amount of signal data. The blue data points show the purity weighted to a realistic signal to background ratio of 1:1000 and is

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}},$$

whereas N_{fp} is the amount of false positive classified events after the signalness cut. For a detailed description of the classification see [7]. While the purity is increasing with an increasing signalness cut, the efficiency is decreasing in the same time. The challenge is to find a signalness cut at which not too much data is cut away while the purity of the dataset is still high. One can find an efficiency of E=45 % - 28 % and a purity of P=20 % - 45 % between a signalness cut of S=0.99 and S=1.0 . The big range between the values shows that the signalness cut has to be trimmed. One possibility to decide which signalness cuts offers the best results is to determine a so-called quality factor Q.

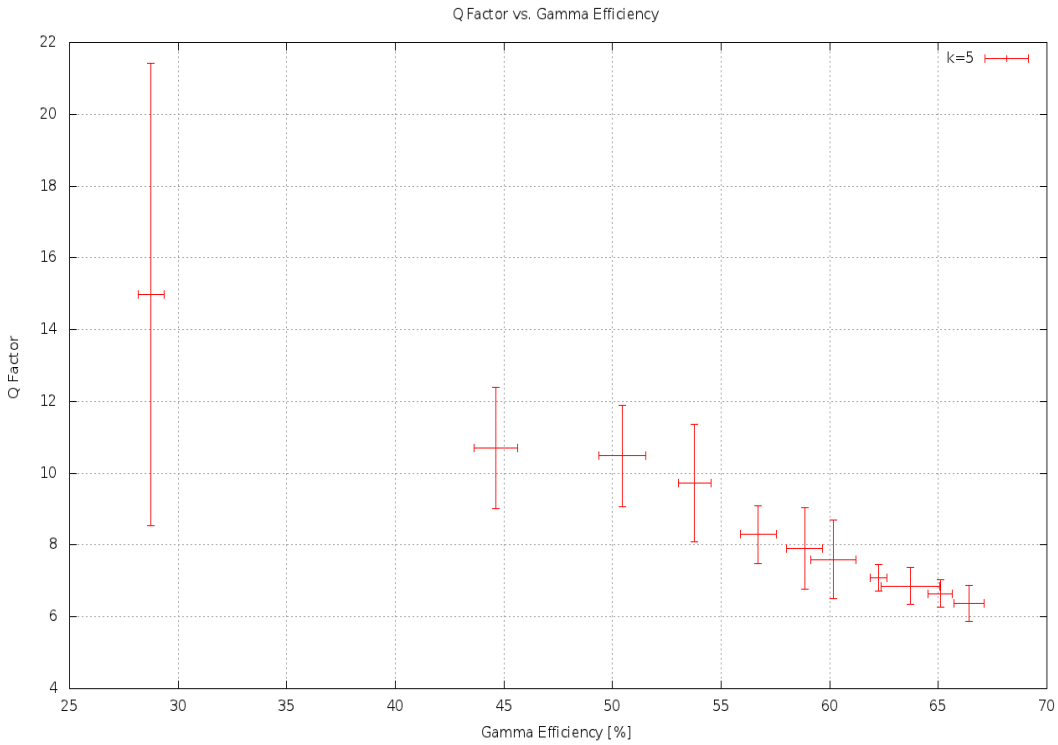


Figure 2: Displayed are the Q factors against gamma efficiencies for a RF model trained with five randomly chosen attributes.

It describes the ratio of the efficiency for gammas to the efficiency of hadronic showers and is

$$Q = \frac{E_G}{\sqrt{E_H}}.$$

In Fig.2 the Q factor is shown for each signalness cut against the gamma efficiency for the trained RF model. The highest quality of a cut would be achieved by a Q factor of Q= 15 at a gamma efficiency of E= 28 % where the supression of hadronic showers and the gamma efficiency are the highest. This leads to a signalness cut of S= 1.0.

The results show that a separation of hadronic and gamma showers is possible but yet still improvable. Due to the current investigations on the Monte Carlo simulations, a better separation power is expected.

References

[1] CORSIKA - An Air Shower Simulation Program
http://www-ik.fzk.de/corsika/.

[2] Rapid I Homepage
http://rapid-i.com/.

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

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

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

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

[7] Julia Thaele. *Gamma-Hadron Separation für das First G-APD Cherenkov Telescope (FACT): Eine Separationsanalyse mit MARS CheObs ed. und RapidMiner, Diploma thesis*. TU Dortmund, April 2012.

The R-Package RobPer

Anita Monika Thieler
Lehrstuhl für Statistik in den Biowissenschaften
Technische Universität Dortmund
anita.thieler@tu-dortmund.de

The R-Package `RobPer` provides functions to calculate periodograms of irregularly sampled time series with additional measurement accuracies and detect outstanding periodogram bars using outlier detection. This method is especially useful for analyzing light curve data, a typical data structure in astroparticle physics. A generator for artificial light curve data for simulation purposes comes with the package as well.

An important task in astroparticle physics is the detection of periodicities in irregularly sampled time series, called light curves. The classic Fourier periodogram cannot deal with irregular sampling and with the measurement accuracies, that are typically given for each observation of a light curve. Hence, methods to fit periodic functions using weighted regression were developed in the past for calculating periodograms.

We present the R-Package RobPer which allows to combine different periodic functions and regression techniques to calculate periodograms with the function `RobPer`, detect outstanding periodogram bars using `betaCvM` and generate artificial light curve data with `tsgen`. These three tasks and the refering functions will be sketched in the following. For more details on their usage, see the manual that comes with the package. For mathematical and implementational details about the methods as well as examples and references, see [6]. A preliminary version of the implementation was used in [5]. The Package is available on http://cran.r-project.org/web/packages/RobPer.

Calculating periodograms using `RobPer`

A periodogram of a time series `ts` calculated by the R-function `RobPer` in the homonymous package consists of periodogram bars that each equal a coefficient of determination.

Each coefficient of determination is the results of the fit of a periodic function to the time series with different period lengths (trial periods) assumed. The model for the periodic function and the regression method is the same for all periodogram bars.

As mentioned before, calculating a periodogram using fits of periodic functions is a common approach in astroparticle physics. Many of the periodograms proposed in the past are analogues of the proceeding implemented in `RobPer`. The measurement accuracies are taken into account using weighted regression by some methods. This is also possible in `RobPer` setting `weighting=TRUE`.

The shape of the periodic function to be fitted can be chosen setting the input parameter `model`. Possible functions are periodic step functions, fourier series up to third degree and periodic spline functions. It is also possible to fit two overlapping step functions separately and use the mean of both coefficients of determination as periodogram bar. Using least squares regression, the resulting periodogram is equal to the popular Phase Dispersion Minimization periodogram of [4].

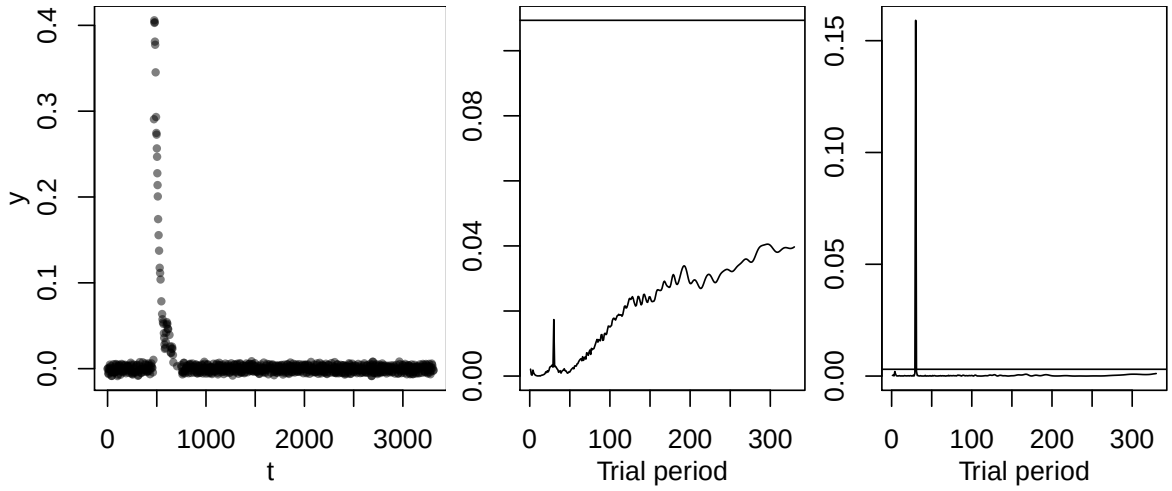
Other possible regression techniques in `RobPer` (controlled using input parameter `regression`) are least absolute deviation regression, least trimmed squares regression, M-regression, S-regression and τ-regression. To the best of our knowledge, the last two techniques mentioned have not been proposed for periodogram calculation before. Besides, robust regression has been mainly combined with the sine function in the past and most of the other `model-regression` combinations with robust regression techniques are new as well.

The regression techniques are implemented using provided R-functions (for example `rq` from the package quantreg for least absolute deviation regression), modifications of published R-Code (for example `FastTau` published with [3]) or own implementations (for example for M-regression, for which none of the available functions met the requirement of this special context).

Detecting outstanding periodogram bars using `betaCvM`

To detect valid periods, we propose to search for outliers in the periodogram instead of using fixed critical values that are only theoretically justified in case of least squares regression, independent periodogram bars and a null hypothesis allowing only normal white noise.

To detect outliers in our periodogram, we proceed the way proposed in [2]: Robustly fit a distribution to the q periodogram bars and call those bars outliers which lie over the $\sqrt[q]{1-\alpha}$-quantile of the distribution, α appropriately chosen. As distribution class, we choose the Beta distributions since the coefficient of determination fitted to a normal white noise sample is beta distributed, too. Our simulations indicate that this holds



(a) (b) (c)

Figure 1: Example for light curve detection (see Text).

at least approximately for the robust techniques above as well. To robustly fit the parameters of a beta distribution to the periodogram, we use Cramér-von-Mises distance minimization, which was successfully used in the past to robustly fit gamma distributions (see [1]) and in our experience works well also with beta distributions.

To fit the beta distribution to a periodogram, the function `betaCvM` from our package can be used. With the parameter estimates returned, the outlier search can be performed. Figure 1(a) shows real data from a peak observed with the BATSE telescope, combined with a sine fluctuation of small amplitude and a period of 30 days. Panel (b) shows the periodogram obtained using least squares regression and a sine fluctuation (so the irregular sampled equivalent to the Fourier Periodogram). The true period is visible, but not outstanding. The $\sqrt[9]{0.95}$ quantile of a fitted beta distribution (horizontal line) lies above all periodogram bars. Panel (c) shows the periodogram obtained when using M-regression with the Huber function (other robust techniques have similar periodograms) and here the true period is outstanding and lies above the respective beta quantile fitted (horizontal line).

Generating artificial light curves using `tsgen`

The function `tsgen` can be used to generate artificial light curves e.g. for simulation studies. The data generated exhibits the following properties:

Irregular periodic sampling An irregular, even periodically irregular sampling is quite typical in light curves due to periodic disturbances as moon light. Can be switched off in order to get unperiodic irregular or even regular sampling.

Periodic structure in observed values A periodic function underlying the observation values to be detected. Can be switched off in order to get pure noise series.

White noise and measurement accuracies The observed values contain white noise with changing standard deviations, which are returned as measurement accuracies.

Additional noise component An additional power-law or white noise component that does not depend on the measurement accuracies. Increasing the amount of this noise component in the overall noise means to lower the informativity of the measurement accuracies. Can be switched off.

Outliers in the measurement accuracies Some random measurement accuracies are replaced by smaller values in order to force the fitted line close to the respective observed value. Can be switched off.

Interval disturbances Observed values in a random time interval are replaced by values forming a high peak. This disturbance is implemented in order to model a similar effect that can be observed in real light curves. It turned out that using robust regression techniques is extremely useful when dealing with such random peaks in the light curve. Can be switched off.

References

[1] B. R. Clarke, P. L. McKinnon, and G. Riley. A fast robust method for fitting gamma distributions. *Statistical Papers*, 53(4):1001–1014, 2012.

[2] L. Davies and U. Gather. The identification of multiple outliers. *Journal of the American Statistical Association*, 88(423):782–792, 1993.

[3] M. Salibian-Barrera, G. Willems, and R. Zamar. The fast-τ estimator for regression. *Journal of Computational and Graphical Statistics*, 17(3):659–682, 2008.

[4] R. F. Stellingwerf. Period determination using phase dispersion minimization. *The Astrophysical Journal*, 224:953–960, 1978.

[5] A. M. Thieler, M. Backes, R. Fried, and W. Rhode. Periodicity detection in irregularly sampled light curves by robust regression and outlier detection. *Statistical Analysis and Data Mining*, 6(1):73–89, 2013.

[6] A. M. Thieler, R. Fried, and J. Rathjens. RobPer: An R Package to Calculate Periodograms for Light Curves Based On Robust Regression. Technical Report 2, Sonderforschungsbereich 876, TU Dortmund University, 2013. Available on http://sfb876.tu-dortmund.de/PublicPublicationFiles/thieler_etal_2013b.pdf.

Feature generation for classification in the FACT experiment based on distances for probability densities

Tobias Voigt
Statistik in den Biowissenschaften
Technische Universität Dortmund
voigt@statistik.tu-dortmund.de

The FACT telescope on the canary island of La Palma is an imaging Cherenkov telescope. Its purpose is to detect highly energetic gamma particles sent out by various astrophysical sources. Due to characteristics of the detection process not only gamma particles are recorded, but also other particles summarized as hadrons. For further analysis the gamma ray signal has to be separated from the hadronic background.
We construct new features for this classification by measuring the Hellinger distance between an observed Cherenkov light distribution and an idealized model distribution for the signal observations. We add these new features to the Hillas parameters and compare the results of classifications with and without the new features.

In VHE gamma-ray astronomy, so-called Hillas parameters are used as features for classification. These variables are based on moment analysis parameters introduced by Hillas (1985), which are based on fitting an ellipse to the shower image and using its parameters such as its length and width as features. The approach we are following in this paper is to extend the idea of fitting an ellipse to the shower image. Instead of fitting an ellipse we fit a bivariate, possibly elliptic-symmetric, density to the shower and use features calculated from the fitted distribution.

An obvious first extension of the idea of fitting an ellipse is the bivariate normal distribution, as it is the best known and most popular elliptical distribution. By fitting a normal distribution instead of an ellipse we can make use of additional information about the height of the shower inside and the tail behavior outside of the ellipse.

Variable	Beschreibung
normhell	$D_h(F_n, F_0)$ with F_0: cdf of bivariate normal
skewhell	$D_h(F_n, F_0)$ with F_0: cdf of bivariate skew-normal
alignnormhell	$D_h(F_n, F_0)$ with F_0: cdf of bivariate normal aligned to source
normkullobs	$D_k(F_n, F_0)$ with F_0: cdf of bivariate normal
normkulltheor	$D_k(F_0, F_n)$ with F_0: cdf of bivariate normal
normchi	Q_n of χ^2-Test for bivariate normal
varquot	Quotient of variances in a rotated shower picture (VarY/VarX)

Table 1: Newly constructed features for classification

A skewed extension of the multivariate normal is the multivariate skew-normal distribution (Azzalini & Dalla Valle, 1996). By fitting a skew-normal distribution we can thus incorporate the fact that signal showers are skewed in at least one direction.

We can also use further information when fitting a distribution to a shower. We know for example that signal showers are always aligned to the center of the camera. To incorporate this information we can fit for example a bivariate normal which is with one direction aligned to the center of the camera.

To fit one of the above densities to the shower images, one possibility is to use Maximum Likelihood estimators (MLEs) for the parameters of the distribution we want to fit. MLEs for the bivariate normal distribution are well known, but have to be adjusted because due to the telescope's imaging process we only have grouped data. For the skew-normal distribution we have numerical methods available for calculating Maximum Likelihood estimates, given by the R package sn (Azzalini, 2013).

We use the fitted distributions to construct new features for classification. The idea here is that we expect the fitted distribution to be close to the true distribution of signal events, so that the fit should be better for signal events than for background events. We thus want to measure the distance between the fitted and the empirical distribution of each of our observed events.

There are many distance measures for distributions, for example goodness of fit measures like the Chi-Squared distance (e.g. Greenwood & Nikulin, 1996) given by

$$Q_n = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i},$$

where m is the number of pixels, n_i the observed intensity in pixel i, p_i the probability of pixel i under the fitted distribution and n the total intensity cumulated over all pixels. There are also some distance measures for densities like the Kullback-Leibler divergence (Kullback & Leibler, 1951), which is in our discrete case given by

$$D_k(F_0, F_n) = \sum_{i=1}^{m} p_i \log\left(\frac{np_i}{n_i}\right).$$

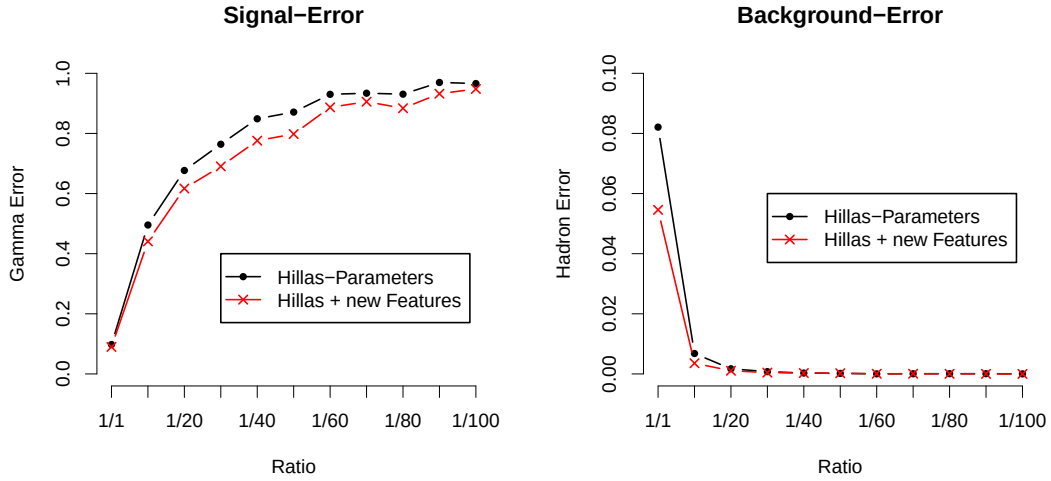


Figure 1: Error Rates depending on the signal-background ratio. Error of Background cannot be improved. Relevant gain for Signal.

A symmetric alternative to the Kullback-Leibler divergence is the Hellinger distance (Nikulin, 2001), which is also bounded by 0 and 1. For the discrete case it is given by

$$D_h(F,Q) = 1 - \sum_{i=1}^{m} \sqrt{p_i \frac{n_i}{n}}.$$

New features can be constructed by combining different distributions to be fitted to the shower images and distance measures. Because of limited space we do not look at all combinations of distributions and measures. An overview of the constructed features can be seen in Table 1.

The FACT data set we use for checking the usefulness of the proposals above consists of 13185 observations, 6478 of which are simulated signal events and the other 6707 are background observations. The dataset includes 14 standard Hillas-Parameters as well as the new features listed in Table 1.

From this dataset we draw subsamples with different signal to background ratios. As written above we expect a ratio of 1:1000 in real data, so we would like to include such unfavorable ratios in our study. However, with our relatively small sample size of about 13000 observations we instead look at ratios $1:1$, $1:10$, $1:20$, ... , $1:100$. We randomly draw 10 subsamples per ratio, train a random forest with the non-used observations and let it classify the sample. The number of background events in each subsample drawn is 1000 with a number of signal events corresponding to the signal-background ratio.

A part of the results is displayed in Figure 1, where the error rates of signal- and background-events can be seen, depending on the signal-background ratio. For the signal we see that the error is increasing, while for background it is decreasing. In background the error decreases very fast, so that it is almost 0 very quickly, and cannot be much improved by better features. So what is desirable is to improve the signal-error while maintaining the good background error. We see on the left side of Figure 1 that we accomplish this. Although with and without the new features the error increases as a function of the ratio, the error is smaller when we use the additional features.

We also had a look at the individual features to find out how important each of them is in the classifiaction. A Random Forest has a built in importance measure for the features used: The mean Gini decrease. It can be shown that some of our newly created features are among the most important of all features used. Others become more important with decreasing signal-to-background ratio.

Summarizing, we have introduced features for classification in the FACT experiment which are based on fitting a bivariate distribution to an image and measuring the distance of the fitted distribution to the observed one. We saw that we get better classification results than with the Hillas parameters alone. Additional physical information about signal showers could be used to construct more features to further improve the classification.

References

[1] Azzalini, A., and Dalla Valle, A.: The multivariate skew-normal distribution. Biometrika, 83, 4 (1996)

[2] Azzalini, A.: R package 'sn': The skew-normal and skew-t distributions (version 0.4-18). URL http://azzalini.stat.unipd.it/SN (2013)

[3] Greenwood, P.E. and Nikulin, M.S.: A Guide to Chi-Squared Testing. Wiley Series in Probability and Statistics, Wiley (1996)

[4] Hillas, A.M.: Cherenkov Light Images of EAS Produced by Primary Gamma. Proceedings of the 19th International Cosmic Ray Conference ICRC, 3, 445, San Diego (1985)

[5] Kullback, S., Leibler, R.A. On Information and Sufficiency. Annals of Mathematical Statistics, 22, 79-86 (1951)

[6] Nikulin, M.S.: Hellinger Distance. in Hazewinkel, M.: Encycolpedia of Mathematics, Springer (2001)

Two-sample Homogeneity Tests Based on Divergence Measures

Max Wornowizki
Statistik in den Biowissenschaften
Technische Universität Dortmund
wornowiz@tu-dortmund.de

In our work we consider f-divergences, density based distance-like measures between distributions [2]. Divergence measures do not focus on certain moments and therefore corresponding tests are capable of detecting any departures from the null hypothesis. Working in the one-dimensional setting with continuous random variables, we propose a new estimation technique involving kernel density estimation, spline smoothing and numerical integration. In addition, we make use of the permutation technique to tackle the two-sample homogeneity problem based on arbitrary divergence estimators. Both methods do not make any assumption on the underlying distributions and are thus widely applicable. They are compared to several non- and semi-parametrical competitors under different distributional assumptions in Monte Carlo experiments and are used to identify interesting attributes in a real data problem.

Given two distributions P and Q with probability density functions p respectively q such that $q(x) > 0$ implies $p(x) > 0$, the f-divergence from P to Q is defined by

$$D_f(P,Q) = \int f\left(\frac{p(y)}{q(y)}\right) dQ \; = E_Q\left(f\left(\frac{p\,(Y)}{q\,(Y)}\right)\right) \quad ,$$

where f is a convex function [2]. The measure attains its minimal value $f(1)$ if and only if $P = Q$ and for all common divergences $f(1) = 0$ holds. The class includes the widely used Kullback-Leibler divergence (D_{KL}), which is closely related to the AIC information criterion and maximum likelihood estimation, as well as the well-known squared Hellinger distance (D_H^2). In contrast to the unbounded D_{KL}, D_H^2 is bounded by 1.

Divergence measures, quite similar to Kolmogorov-Smirnov type statistics [3], reflect general discrepancies of the distributions. They take into account deviations in the mean, the scale, the skewness, the tail behaviour and any other characteristics of the distributions, and weight them implicitly according to the function f. Thus, methods based on divergences can reveal arbitrary dissimilarities between distributions.
The estimation of the measures often consists of two steps. At first, the density ratio function $r : \mathbb{R} \to \mathbb{R}$ with $r(x) = \frac{p(x)}{q(x)}$ is estimated by $\hat{r}$. Then, the divergence is estimated based on $\hat{r}$. Several approaches to both steps exist in the literature. For the density ratio estimation one can either estimate p and q separately by kernel density estimation (nonparametrical) or assume a model r_θ for the ratio r (semiparametrical). Thereby, the estimation of r boils down to the identification of a suitable parameter θ.
Given an estimator $\hat{r}$, an f-divergence can be easily estimated by

$$\hat{D}_f = \frac{1}{m} \sum_{j=1}^{m} f\left(\hat{r}(y_j)\right) \ ,$$

since the measure itself is nothing but the expectation $E_Q\left(f\left(r\left(Y\right)\right)\right)$.
We propose a numerical estimation procedure to estimate f-divergences, which considers the measures as the integral of $f\left(r(x)\right) \cdot q(x)$. At first, the densities p and q are estimated using kernel density estimation and $\hat{r}$ is set to $\frac{\hat{p}}{\hat{q}}$. Then, the function $f\left(\hat{r}(x)\right) \cdot \hat{q}(x)$ is smoothed by cubic splines [4] and integrated numerically.
The method is applied to simulated data from different distributions and sample sizes. It is compared to the non- and semiparametric alternatives sketched above and several others. Explicit representations of both divergence measures for the chosen distributions allow to assess the performance of the estimators in terms of the empirical mean squared error.
The results show that the Hellinger divergence is easier to estimate in comparison to the Kullback-Leibler divergence, possibly due to its boundedness on the interval (0,1). The numerical estimator shows a small mean squared error and is stable. It is outperformed by some semiparametric methods, but these suffer in case of a misspecified density ratio model r_θ.
To test $H_0 : P = Q$ using $f-$divergences without imposing distributional assumptions, we make use of the permutation technique. The method proceeds in the following way using an arbitrary divergence estimator: At first, b new sample pairs are generated from the original observations $x_1, \ldots, x_n, y_1, \ldots, y_m$ by randomly permuting the sample labels b times. Then, the divergence is estimated on each of the b sample pairs as well as on the original data. The permutation test rejects H_0 if the divergence estimate on the original data exceeds the empirical $(1-\alpha)-$quantile of all $b+1$ divergence estimates, where α is the predefined significance level.
The permutation procedure is applied for different divergence estimators, distributions, sample sizes and H_1-alternatives and compared to some parametric tests, an asymptotic divergence based test using a semiparametric divergence estimator [5] and the nonparametrical Kolmogorov-Smirnov [3] and Anderson-Darling test [6]. The results indicate

that the choice of the divergence measure does not affect the results as much as the estimation technique. The test based on the numerical estimator performs again stable and quite well among the class of permutation tests. Compared to the classical nonparametrical procedures, the permutation methods detect scale alternatives more often and location alternatives less often. For different scale and location for both distribution, the results depend on the situation. Comparing samples from Gaussian and t-distributions using the test procedures we observe that divergence based permutation tests detect departures from the Gaussian distribution more often then the nonparametrical methods. The asymptotic test procedure performs better than the permutation tests as long as the underlying assumptions are fulfilled, but breaks down in case of their violation. The parametric tests perform best for a suited alternative, but depend on the distributional assumptions. These conclusions seem to hold independently of the chosen sample size.
We also apply our method to identify interesting variables on the IceCube data. IceCube is a cubic kilometer large neutrino detector located at the geographic South Pole [1]. It consists of 86 strings with 60 digital optical modules each embedded in glacial ice between 1450 and 2450 m below the surface. Unfortunately, IceCube does not only detect atmospheric neutrinos. The measurements consist mostly of atmospheric muons, which are not of interest in the given context. Classification algorithms like Random Forests are obvious tools for the separation of the relevant from irrelevant observations. However, no correctly labeled real data is available to train the methods. Simulations have to be used instead at this step of the analysis. Since conclusions drawn from wrong simulations are useless or even worse, good agreement of simulated and observed data must be guaranteed. If the atmospheric muons are well simulated, the corresponding data set should resemble the real data, since the majority of the real data is based on atmospheric muons. Dissimilarities for particular variables may either be caused by a false simulation or by the atmospheric neutrinos in the observed sample. In both cases, the attributes are of great interest. We thus conduct the permutation procedure to test the equality of distributions for each attribute using observed data and a simulated atmospheric muon sample. The datasets consist of 1000 observations and 96 continuous variables. The permutation tests are applied using the numerical estimator for the squared Hellinger distance and 500 permutations.
The null hypothesis of equal distributions is rejected for 19 of the 96 variables. To illustrate the procedure, we consider the attribute SPEFit8Bayesian_SmoothAll, for which H_0 was rejected. Kernel density estimations for this attribute are shown in Figure 1 for both samples. The distributions seem to agree on the left part. However, the shape of the right parts differs somewhat showing more frequent observations than simulated values near 0.2.

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] S. M. Ali and S. D. Silvey (1966)
A General Class of Coefficients of Divergence of One Distribution from Another.
Journal of the Royal Statistical Society, Series B, 28(1), 131-142

[3] J. Durbin (1973)
Distribution theory for tests based on the sample distribution.
CBMS-NSF Regional Conference Series in Applied Mathematics 9, London School of Economics, London

[4] P. J. Green and B. W. Silverman (1994)
Nonparametric Regression and Generalized Linear Models:
A Roughness Penalty Approach.
CRC Monographs on Statistics & Applied Probability (Book 58), Chapman and Hall, New York

[5] T. Kanamori, T. Suzuki and M. Sugiyama (2012)
F-divergence estimation and two-sample homogeneity test
under semiparametric density-ratio models.
IEEE Transactions on Information Theory, 58(2), 708-720

[6] F. W. Scholz and M. A. Stevens (1987)
K-sample Anderson-Darling Tests.
Journal of the American Statistical Association, 82(399), 918-924

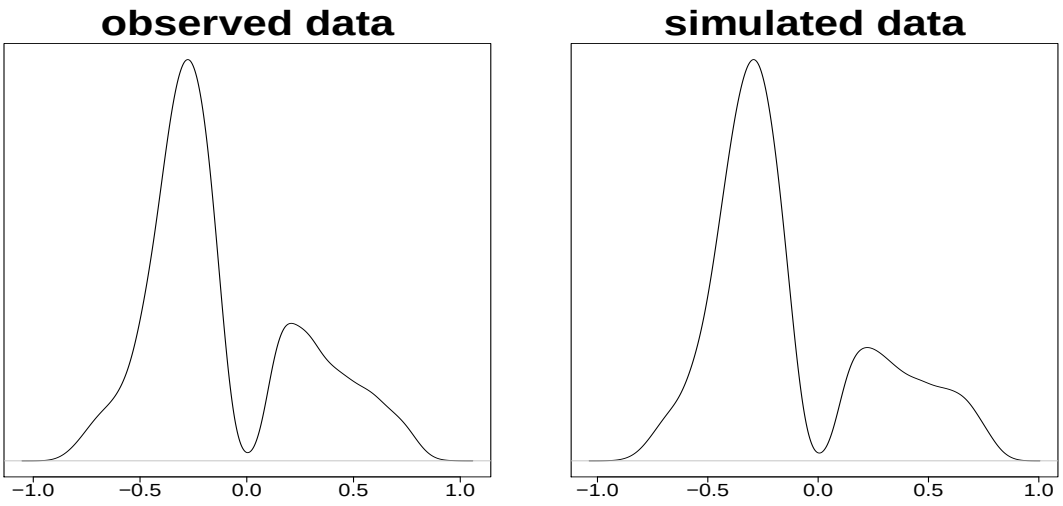


Figure 1: Kernel density estimations for the observed data and the simulated atmospheric moun data for the variable SPEFit8Bayesian_SmoothAll

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

Katja Ickstadt Christian Sohler

Empirical evaluation of the use of sketches for Bayesian regression

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

We use efficient sketches to reduce the number of observations in very large data sets. For both the original and the new data sets, Bayesian regression models are built. Even though the reduced data sets are much smaller in size, the results of the models stay virtually the same. The total run-time for reduced data sets is considerably less than for full data sets. The relative difference between the run-times increases even more for larger data sets.

Introduction

Regression analysis is an important tool in statistics. A classical linear regression model is given in equation (1):

$$Y = x\beta + \varepsilon. \tag{1}$$

$Y \in \mathbb{R}^n$ is a random variable containing the values of the response. n is the number of observations in the data set. $x \in \mathbb{R}^{n \times p}$ is a matrix containing the values of the p independent variables. To allow for measurement error, $\varepsilon \sim N(0, \sigma^2 I_n)$ is introduced. ε is an n-dimensional random vector. $\beta \in \mathbb{R}^p$ is the unknown parameter vector. In a classical setting, β is assumed to be unknown, but fixed. In a Bayesian setting, β is assumed to follow a distribution. A prior distribution for β can incorporate prior knowledge about the parameter. Absence of such knowledge can also be dealt with. The standard methods for analysing Bayesian models are Markov Chain Monte Carlo (MCMC) methods. MCMC methods sample candidate values from a proposal distribution and accept or reject the

candidates with a probability proportional to the posterior distribution. To calculate this probability the whole data set is employed. Consider a data set with $n \gg p$. On such large data sets the computational cost and the necessary memory to apply MCMC methods become prohibitive.

Efficient sketches

To address this problem, we suggest efficient random sketches based on the so-called *Johnson-Lindenstrauss-Transform* (JLT). Johnson and Lindenstrauss [2] showed that for every n-dimensional vector v there exists a random matrix $S \in \mathbb{R}^{k \times n}$ such that

$$(1 - \xi)\|v\|^2 \leq \|Sv\|^2 \leq (1 + \xi)\|v\|^2. \tag{2}$$

Applying equation (2) to Bayesian regression leads to minimising $\|Sx\beta - SY\|$ with respect to β. Our new regression problem only consists of k observations, with $k \in \Omega(p \ln(p/\xi)/\xi^2)$. Notably, k does not depend on n, which makes this approach very suitable for Big Data.

Munteanu [3] has described our approach to applying efficient sketches. For theoretical details, please see there. This technical report deals with the results of experiments we conducted to check how well the approach does in practice.

Simulations

For the simulations, we use the statistical software package R, version 2.15.1 [4] and the R-package `rstan`, version 1.0 [5]. We simulate data sets, which contain $p \in \{50, 100, 200\}$ variables and $n \in \{50\,000, 100\,000, 500\,000, 1\,000\,000\}$ observations. Y is modelled to depend on some or all of the variables. All assumptions of Bayesian linear regression are fulfilled.

A Bayesian regression model is built for the full data set. An additional model is built for the sketched version of the data set. The time it takes to calculate the sketch is also included in the run-time.

`rstan` employs the no-U-turn sampler, an enhancement of Hamiltonian Monte Carlo [1]. This algorithm was first published in 2012 and in general is faster than alternatives like WinBUGS. `rstan` is especially useful if hierarchical models are considered.

We used 4 chains with 10 000 MCMC iterations and a burn-in of 5 000 iterations each. In total 5 000 observations were kept for each of the chains. If convergence had not been reached by then, the number of observations was increased. This, however, only occurred in few cases.

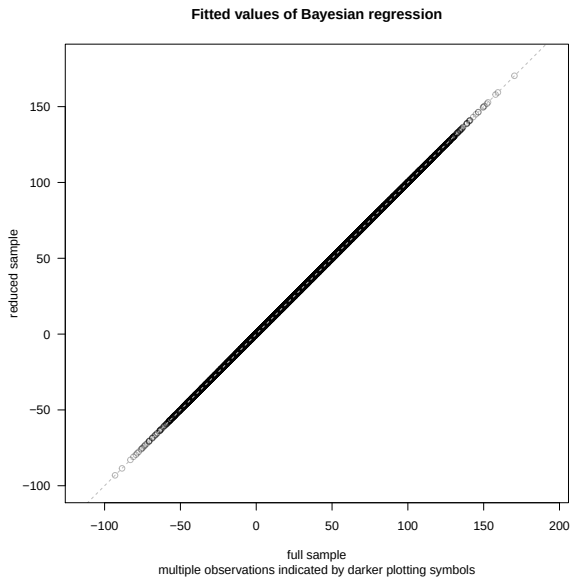


Figure 1: Fitted values of corresponding Bayesian regression models

Results

Generally speaking, the models, which were obtained using the full sample and their reduced counterparts are very similar. Figure 1 contains a scatterplot with the fitted values of a model based on a full sample on the x-axis and the fitted values of a model based on a reduced sample on the y-axis. It is based on a comparatively small data set with $n = 50\,000$ and $p = 50$. ξ was chosed to be 0.1, which results in a reduced data set with $k = 16\,384$ observations and $p = 50$ variables. The fitted values of both models are almost exactly the same. This situation is very typical for our results, in general we have found only very minor deviations in the fitted values. The same also applies to the posterior distributions of the components of beta.

Figure 2 shows the necessary run-time to obtain the results. For the reduced sample, this includes both the time needed to calculate the sketch as well as the run-time of the MCMC sampler. The run-time of the sketching function is negligible in comparison, however. If n increases, the calculation of the sketch matrix needs a longer time, but as the size of the reduced sample does not increase, the run-time of the MCMC sampler does not change substantially. In total this means that the difference in relative run-times increases even more as n increases.

Taking [3] into account, we can conclude that both theory and practice of this method show only very small differences between models based on the full sample and models based on a reduced sample. The statistical results and their interpretation remain the same.

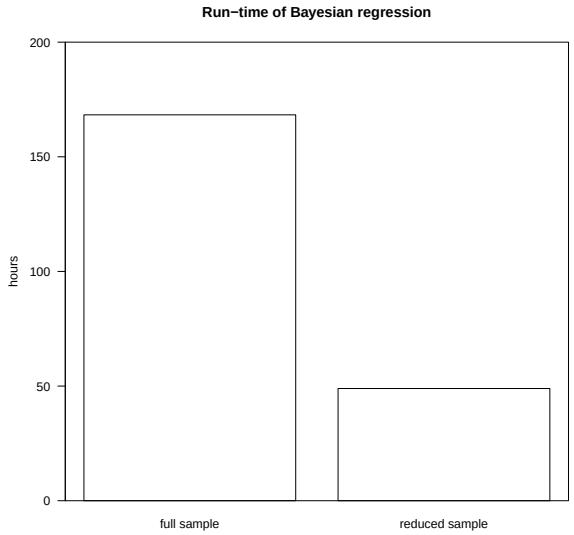


Figure 2: Run-time of corresponding Bayesian regression models

Please note that a sensible model is required in order to find the sketch matrix S. If the initial full model is wrong, the performance of the reduced sample model decreases. Careful model selection and model checking is necessary.

References

[1] Matthew Douglas Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 2012.

[2] William Buhmann Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[3] Alexander Munteanu. Efficient sketches for bayesian regression. Technical report, Technische Universität Dortmund, 2012.

[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] Stan Development Team. Stan: A c++ library for probability and sampling, version 1.0, 2012.

Towards dimensionality reduction for Bayesian regression analysis

Alexander Munteanu
Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
alexander.munteanu@tu-dortmund.de

This report deals with a dimensionality reduction technique based on principle component analysis. While for many applications it is the best choice to keep the largest principle components of the data as a low dimensional approximate representation, this is not true for regression analysis. Our main contribution is to derive selection rules for the best choice of a subset consisting of k principle components, when the objective is to minimize regression error. The results can be further generalized to the Bayesian treatment of regression problems.

Introduction

In *Big Data* analysis we are faced with the problem of dealing with huge amounts of observations. Data reduction techniques like sampling, coreset constructions and random projection algorithms have been proposed to address these difficulties in streaming as well as in distributed models of computation. While their time, space and communication complexities mostly have a very small polylogarithmic or even constant dependencies on the number of observations, the number of attributes (i.e., the dimension of the points) often still dominates these quantities. In this report we develop a dimensionality reduction technique based on the *best* choice of k principle components of the input data. While for many applications the notion of *best* is only based on loosing as little variance as possible of the input data points, this is not desirable for regression analysis since the variance of points spanning a subspace does not provide any information on the norm of an orthogonal projection to that subspace. In Bayesian regression analysis in turn, we are

interested in preserving as much information on the variance of a multivariate distribution as well as its mean. Our analysis shows that keeping the large variances of the induced distribution corresponds to keeping the *smallest* variances of the data. This in turn may still interfere with keeping most of the information on the mean value because most of the norm of the corresponding projection may lie in components of large variance.

Principle component regression

As a dimensionality reduction approach for linear regression, students textbooks like [3] propose a method called principle component regression. The main idea of this technique is to perform a principle component analysis (PCA) on the normalized data to identify the main directions in which the data is spread. The principle components form an orthogonal basis for the space spanned by the data points. For the sake of presentation the components are assumed to be arranged in decreasing order of the variance they explain, i.e. the *first* principle component carries the highest amount of norm while the *smallest* component explains the least amount of variance in the data.

Now the regression is performed on the principle components instead of the actual data. If we keep all the components, then the regression is equivalent to the regression on the data up to affine transformations. But remember, we are interested in reducing the dimension of the data from d to $k < d$. The actual dimensionality reduction is achieved by only regressing on the first k components instead of keeping all of them. Note that the major amount of variance in the data is kept but any correlation with the dependent variable is ignored by this simple selection rule.

Geppert [1] and Jabs [5] reported preliminary results which confirm that principle component regression can be adapted to the Bayesian treatment of linear regression problems and actually lead to a considerable dimensionality reduction and faster convergence of Markov chain Monte Carlo algorithms. The major drawback of regressing on principle components turned out to be that the model fit was a little worse than for the slightly slower ℓ_1 penalized LASSO-Regression. For more information on principle component regression and the LASSO consider [7] and [8] respectively.

Improving the selection rules

As repeatedly noted and demonstrated in [2, 6] and [4] the small components can be as important or even more important to keep for linear regression as the large components. Hawkins also gives decision rules which favor the small components [4]. Otherwise, cautionary notes on the use of the top components are given based on actual data examples.

In the following we theoretically develop decision rules on how to choose exactly k out of d principle components to ensure the best model fit among all k-element subsets of the components. As in [3] we will assume the data is normalized.

Let $A \in \mathbb{R}^{n \times d}$ be a matrix consisting of n data points from d-dimensional space and let $b \in \mathbb{R}^n$ be a vector of target variables corresponding to the n data points. Then linear regression is the problem of finding $\text{argmin}_{x \in \mathbb{R}^d} ||Ax - b||_2^2$. That is, we seek to find the point in the column space of our data that minimizes the squared Euclidean distance to the target vector b. It is a well-known fact that the minimizer corresponds to the orthogonal projection of b to the column space of A. Let $A = U\Sigma V^T$ be the singular value decomposition of A. Then the columns of V correspond to the principle components, the squared diagonal elements of Σ are the variances along these components and we can characterize the optimal solution to our optimization problem by $x_{opt} = V\Sigma^{-1}U^T b$. Note that by the Pythagorean theorem we can measure the *additional* loss for any x by $||x - x_{opt}||_2^2$. For this sake we define index sets $I \subseteq [d]$ corresponding to any fixed choice of a subset of principle components and its complement $\bar{I} = [d] \setminus I$.

Now let I be an arbitrary but fixed choice of k principle components and regress only on these. This leads to $y_{opt} = V\Sigma_I^{-1}U^T b$ where Σ_I is is derived from Σ but all the diagonal entries with indices in $\bar{I}$ are set to 0. Note that the resulting matrix is not invertible but it suffices to work with the pseudo-inverse, where only non-zero entries are inverted and the zeros are kept. Thus

$$\begin{aligned} ||y_{opt} - x_{opt}||_2^2 &= ||V\Sigma_I^{-1}U^T b - V\Sigma^{-1}U^T b||_2^2 \\ &= ||\Sigma_I^{-1}U^T b - \Sigma^{-1}U^T b||_2^2 \\ &= ||(\Sigma_I^{-1} - \Sigma^{-1})U^T b||_2^2 \\ &= \sum_{i \in \bar{I}} \frac{(U^T b)_i^2}{\sigma_i^2} \end{aligned}$$

is the exact loss of regressing only on the components indexed by I instead of using the original high-dimensional data. Since the principle components are orthogonal and therefore the terms are independent of each other, this sum can be easily minimized by choosing I to consist of the k largest components according to the scores $s_i = \frac{(U^T b)_i^2}{\sigma_i^2}$. Note that these scores incorporate the norm of b contained in the principle components of the data as well as the *inverse* variance of the data.

Outlook

It is straight-forward to extend the above derivation to the Bayesian treatment of linear regression. The resulting scores have an additional additive term to deal with the loss in variance when leaving out single components. Therefore the scores take the trade-off into

account which occurs between keeping the projected norm and keeping as much variance as possible. Another interesting aspect is that the variance of the posterior distribution (which we would like to preserve) is *inversely* related to the variance of the data points. Therefore the components of *small* variance are more important when their amount of projected norm is of comparable order. This was previously indicated an motivated in [4] for standard linear regression.

It would be interesting to conduct an empirical evaluation following the work of Geppert and Jabs [1, 5] and to compare the performance of PCA-based Bayesian regression using our adjusted selection rule to the slightly slower LASSO-Regression. The evaluation of the model fit should improve and possibly outperform the LASSO models due to the more appropriate choice of the principle components to keep, which we have developed theoretically in this report.

References

[1] L. Geppert. Combining Dimensionality Reduction Techniques and Bayesian Regression. In Katharina Morik and Wolfgang Rhode, editors, *Technical report for Collaborative Research Center SFB 876 - Graduate School*, number 2, pages 212–215, September 2012.

[2] Ali S. Hadi and Robert F. Ling. Some cautionary notes on the use of principal components regression. *The American Statistician*, 52(1):pp. 15–19, 1998.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003.

[4] Douglas M. Hawkins. On the investigation of alternative regressions by principal component analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3):pp. 275–286, 1973.

[5] Verena Jabs. Vergleich von Methoden zur Dimensionsreduktion unter Berücksichtigung der Rechenzeit und des Speicherbedarfs. Bachelorarbeit, TU Dortmund, 2012.

[6] Ian T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):pp. 300–303, 1982.

[7] Ian T. Jolliffe. *Principal component analysis*. Springer series in statistics. Springer, New York, 1986.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, 58(1):267–288, 1996.