technische universität
dortmund

SFB 876

Technical report for Collaborative Research Center SFB 876

Providing Information by Resource-Constrained Data Analysis

October 2011

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 Felix Jungermann 2

1.2 Nico Piatkowski 6

1.3 Jochen Streicher 10

2 Subproject A2 **15**

2.1 Sylvie Temme 16

2.2 Andreas Thom 20

3 Subproject A3 **25**

3.1 Michel Lang 26

3.2 Sascha Plazar 30

3.3 Helena Kotthaus 34

4 Subproject A4 **39**

4.1 Matthias Meier 40

4.2 Markus Putzke 44

4.3 Björn Dusza 48

4.4 Christoph Borchert 52

5 Subproject A5 **57**

5.1 Patrick Krümpelmann 58

5.2 Marcel Preuß 62

5.3 Cornelia Tadros 66

6 Subproject B1 **71**

6.1 Dominik Kopczynski 72

6.2 Martin Marcel 76

6.3 Kathrin Rupp 80

6.4 Anne-Christin Hauschild 84

7 Subproject B2 **89**

7.1 Dominic Siedhoff 90

7.2 Pascal Libuschewski 94

7.3 Constantin Timm 98

8 Subproject B3 **103**

8.1 Fabian Bohnen 104

8.2 Benedikt Konrad 108

8.3 Marco Stolpe 112

8.4 Christian Bockermann 116

9 Subproject B4 **121**

9.1 Christoph Ide 122

9.2 Brian Niehöfer 126

9.3 Timo Knaup 130

9.4 Daniel Weber 134

10 Subproject C1 **139**

10.1 Kathrin Fielitz 140

10.2 Melanie Heilmann 144

11 Subproject C3 **149**

11.1 Tobias Voigt 150

11.2 Anita Monika Thieler 154

11.3 Jan-Hendrik Köhne 158

11.4 Dominik Neise 162

11.5 Michael Backes 166

11.6 Nikola Strah 170

11.7 Fabian Clevermann 174

11.8 Malwina Uellenbeck 178

11.9 Natalie Milke 182

11.10Tim Ruhe 186

12 Subproject C4 **191**

12.1 Leo Geppert 192

12.2 Chris Schwiegelshohn 196

12.3 Alexander Munteanu 200

SFB 876 Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Olaf Spinczyk

Information Extraction in RapidMiner

Felix Jungermann
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
felix.jungermann@tu-dortmund.de

This paper describes the *Information Extraction Plugin*[1] [3] which allows the use of *Information Extraction* mechanisms in *RapidMiner*[2].

1 Introduction

RapidMiner – which is shortly presented in the following – works with a certain data structure for storing datasets. This data structure has to be respected by our extension leading to particular requirements which are presented in this Section, too. The process of a particular data mining task can be separated in 4 distinct parts in *RapidMiner*. Certain requirements of these parts for *Information Extraction* are presented in Section 2.

RapidMiner is an open source data mining framework. It offers many operators which can be plugged together into a process. The major function of a process is the analysis of the data which is retrieved at the beginning of the process. The framework offers a well designed graphical user interface that allows to connect operators with each other in the process view. Each *RapidMiner*-process can be split into four distinct phases:
Retrieve: During the *Retrieve* phase the data which is used for later processing is loaded from specific datasources.
Preprocessing: The retrieved data has to be prepared or enriched in the *Preprocessing* phase.
Modelling: The prepared data is used in the *Modelling* phase to extract or to create models which can be used for analysing former unknown data.
Evaluation: The expected or real performance of the created models is evaluated during

[1] `http://www-ai.cs.tu-dortmund.de/SOFTWARE/IEPLUGIN`
[2] `http://www.RapidMiner.com`

the *Evaluation* phase.
These phases nearly stay the same for every *RapidMiner* process. Therefore, we will define the particular phases and the corresponding specialties using the *Information Extraction Plugin* in Section 2.

The data structure in *RapidMiner* is comparable to a spreadsheet in which the lines represent examples and the columns represent attributes. It is remarkable that for many data mining tasks the examples are handled indepedently. It follows that the analysis of a particular example *i* just depends on the attributes of example *i* instead of depending on other examples. The structure of documents and texts should be respected for *Information Extraction* tasks. Conditional random fields (CRF) [6], for instance, process all the tokens of a particular sentence at the same time, and the tokens of the certain sentence condition each other. We used the datastructure of *RapidMiner* and we developed mechanisms to respect the circumstances of *Information Extraction* tasks.

2 Information Extraction Plugin

Like for every data mining task the process for *Information Extraction* tasks also can be split in four phases. These phases and the according operators to be used in each phase are described in this section.

Retrieve
The process to retrieve datasets into *RapidMiner* is remarkable for *Information Extraction* issues. We present two approaches. The resulting dataset contains an additional special attribute (*batch*-attribute) which groups the single examples. In addition to the grouping the sequential ordering of the examples is respected by many operators of the plugin.
Retrieve via Document The *Text Mining extension* of *RapidMiner* already offers the possibility to retrieve document structured data. For *Information Extraction* purposes we would like to tokenize the document and to preserve the order of such tokens, therefore, we implemented tokenizers which are able to process example sets extracted from documents. The application of these tokenizers result in a spreadsheet containing the tokens in the particular order as they have been found in the document. Each word-token of a particular sentence, for instance, contains the number of the sentence, whereas each sentence-token of a document contains the number of that document. The *Tokenizer*-class can be easily extended to create own tokenizers.
Retrieve via File Datasets like the one for the CoNLL 2003 shared task on named entity recognition (NER)[3] which is well-known in the NER community are already presented as tokenized documents. Comparable to csv-files the datasets contain a token each line whereas the tokens additionally contain features which are also stored in the particular line. The main difference to ordinary csv-files is the fact that sentences are split by empty

[3] http://www.cnts.ua.ac.be/conll2003/ner/

lines. Although the *Read CSV*-operator of *RapidMiner* can be adjusted to neglect such lines, we developed an own retrieve operator which respects the empty lines to distinguish between distinct sentences.

Preprocessing

Preprocessing is a very important point in *Information Extraction*. In contrast to traditional datamining tasks the data is not given by examples already containing different attributes extracted from a database, for example. The original data just contains tokens consisting of nothing but the token and the contextual tokens itself. The tokens have to be enriched by attributes to get a more general representation. The first type of processing we present is to enrich tokens for NER. The other type of processing is for enriching relation candidates for relation extraction.

Named Entity Recognition It is very important to enrich tokens for NER by internal and external information. We developed an abstract class for preprocessing operators that allows to focus on tokens before or after the current token and on the current token itself. Using this class allows accessing contextual tokens in a relative way. Each token is processed and the abstract class accesses a number of tokens before and after the current token. The number of tokens to access before and after the current token are parameters that can be adjusted by the user. The most simple way of preprocessing would be to enrich the current token by the values of the surrounding tokens. If a token is at the beginning or at the end of a sentence some of the contextual attributes will contain *null*-values because the tokens of one particular sentence only contain informational units from that sentence. In addition to the relative contextual tokens to be taken into account another interesting parameter to set for preprocessing operators is the *length*. Some created attributes like *prefixes* or *suffixes*, for instance, have a specific length which has to be adjusted by this parameter. Another important point for NER datasets is the manual labeling of documents and texts. We implemented an annotator operator which allows to display the dataset as a textual document allowing the user on the one hand to create new labels and on the other hand to use those labels for annotating the tokens. After having used that operator the dataset contains a label attribute carrying the annotations and a formerly defined default value if no annotation is given for a particular token.

Relation Extraction For relation extraction we developed particular preprocessing operators working especially for relational learning purposes. In addition to the flat features developed by [7], relation extraction heavily relies on structural information like parse trees, for instance. We developed parsing operators to first of all create tree-structured attributes. Additionally, we implemented pruning methods for the creation of more condensed tree structures. We developed a generic form of attribute which allows the storage of every type of *Java*-object. This generic object-attribute can be used to work with tree structures in *RapidMiner* [4]. Like for nominal values, the object attribute is storing a mapping which maps numerical values to particular objects.

Modelling and Evaluation

We implemented or embedded many learning techniques for *Information Extraction* in *RapidMiner*. The particular learning methods are CRFs [6], Tree Kernel SVMs [2], Tree Kernel Perceptrons [1] and Tree Kernel Naïve Bayes Classifier [5]. The learning methods already available in *RapidMiner* of course can be used, too. Although the operators of *RapidMiner* for evaluation can be used for many *Information Extraction* tasks, some *Information Extraction* tasks need specific evaluation. We implemented a certain `PerformanceEvaluator` especially for NER tasks.

Summary

We presented the *Information Extraction Plugin* in this work. The plugin is an extension to the well-known open source framework *RapidMiner*. Traditional datamining tasks can be split into four particular phases in *RapidMiner*. These phases are apparent for *Information Extraction* tasks, too. Due to the spreadsheet datastructure internally used by *RapidMiner* we developed a representation of datasets for *Information Extraction* based on that datastructure. In addition, we implemented several operators helpful and needed for *Information Extraction* purposes.

References

[1] Fabio Aiolli et al. Efficient kernel-based learning for trees. In *Computational Intelligence and Data Mining (CIDM). IEEE Symposium on*, pages 308 –315, 2007.

[2] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Proceedings of Neural Information Processing Systems, NIPS 2001*, 2001.

[3] Felix Jungermann. *Documentation of the Information Extraction Plugin for RapidMiner*, August 2011.

[4] Felix Jungermann. Handling tree-structured values in rapidminer. In *Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011)*, 2011.

[5] Felix Jungermann. Tree kernel usage in naive bayes classifiers. In *Proceedings of the LWA 2011*, 2011.

[6] John Lafferty et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th ICML*, pages 282–289, 2001.

[7] GuoDong Zhou et al. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 427–434, 2005.

Parallel Algorithms for GPU accelerated Probabilistic Inference

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

This report introduces and compares two approaches for accelerating the inference in graphical models by using GPUs as parallel processing units.

Introduction

Data which is extracted from ubiquitous devices like smartphones, medical devices or sensor networks is likely to contain an inherent structure. Those structures may be represented with graphs which encode independence assumptions within the data. However, performing inference on those models with an complex structure on a large amount of data is computational expensive and nearly intractable on mobile devices or casual workstations. To overcome this issue, we take advantage of recent developments in computer hardware, namely programmable Graphics Processing Units (GPUs) to accelerate the inference process. In this report, we investigate the parallelization of *Loopy Belief Propagation* (LBP) which corresponds to the application of ordinary Belief Propagation (BP) [5] to graphs with loops. LBP is a broadly used algorithm for probabilistic inference in graphical models like markov random fields (MRFs) or conditional random fields (CRFs). For the sake of generality, we will use the term LBP, even if the underlying graphical model contains no loops.

Parallel LBP for GPUs

A GPU may be used as a parallel processing unit beside the usual central processing unit. Here, parallel processing means the concurrent execution of several instances of one and the same procedure. To perform such computations, a GPU is composed of several so called *Multiprocessors* (MPs). Each of those MPs consists of a fixed number P of *Cores* which perform the actual arithmetic or logic computations. Cores of different MPs may execute different instructions while all cores within an MP have to execute the very same instruction simultaneously. However, the operands of all cores within an MP may differ, what they usually do. Each MP is equipped with a small, low-latency *shared memory* of size S and is furthermore connected to a large high-latency *global memory*.

An instance of a concurrently executed procedure is called *thread*. The threads are partitioned into equally sized sets, called *thread-blocks* [4]. When a parallel procedure is executed, the thread-blocks are automatically assigned to an MP with free resources and the corresponding threads to this MP's cores. A GPU may manage substantially more threads than real cores are available, since their computations are used to hide global memory's latency. To achieve the maximum throughput, there always have to exist some threads which perform computations on the cores while other threads are waiting for their memory requests. Design patterns for parallel algorithms may help to utilize all cores of a GPU. *Parallel reduction* is a pattern for the generic parallelization of functions satisfying the associative property. We apply the parallel reduction algorithm as proposed by Harris [1], which is known to achieve the maximum throughput on GPUs.

We will now introduce the necessary notation and the basic concept of LBP. Following [2], we focus on graphical models which are represented as factor graphs. A factor graph $G = (V, E, F)$ consists of a set of *variable nodes* V, a set of *factor nodes* F and an undirected edge set $E = F \times V$. The variable nodes are indentified with discrete random variables. Their joint realization is denoted with $\mathbf{v}$. Let $\mathbf{v}_U$ be the joint realization $\mathbf{v}$ restricted to nodes in an aribtrary set $U \subseteq V$. For notational convenience, let v describe the node v as well as the singleton set $\{v\}$. In the following, we assume that the variable nodes are partitioned into two distinct sets X and Y with domains $\mathcal{X}$ and $\mathcal{Y}$.The nodes $f \in F$ correspond to positive functions $f : \mathcal{Y}^{|\Delta(f)|} \times \mathcal{X}^{|\tilde{\Delta}(f)|} \to \mathbb{R}^+$, where $\Delta : F \to 2^Y$ assigns to a factor node it's adjacent nodes[1] in Y and $\tilde{\Delta}$ those in X, respectively. The computation of conditional marginal probabilities $p(\mathbf{y}_{\Delta(f)}|\mathbf{x})$ with LBP consists in repeatedly computing Eq. 1 and 2.

$$m_{f\to v}(\mathbf{y}_v|\mathbf{x}^{(i)}) = \sum_{\mathbf{y'}_{\Delta(f)-v}\in\mathcal{Y}^{|\Delta(f)|-1}} f(\mathbf{y}_v, \mathbf{y'}_{\Delta(f)-v}|\mathbf{x}^{(i)}) \prod_{u\in\Delta(f)-v} m_{u\to f}(\mathbf{y'}_u|\mathbf{x}^{(i)}) \tag{1}$$

$$m_{v\to f}(\mathbf{y}_v|\mathbf{x}^{(i)}) = \prod_{g\in\Delta^{-1}(v)-f} m_{g\to v}(\mathbf{y}_v|\mathbf{x}^{(i)}) \tag{2}$$

The *factor messages* (Eq. 1) have to be computed for all possible combinations of factor node $f \in F$, neighbor $v \in \Delta(f)$, realization $\mathbf{y}_v \in \mathcal{Y}$ and instance $\mathbf{x}^{(i)}, 1 \leq i \leq b$. Although the number of *variable messages* (Eq. 2) is similar to the number of factor messages, the time complexity $\mathcal{O}(\max_{v\in V} |\Delta^{-1}(v)|)$ for computing such a message from v to f is rather low if compared to the time complexity $\mathcal{O}(\max_{f\in F} |\Delta(f)||\mathcal{Y}|^{|\Delta(f)|-1})$ for computing a factor message from f to v. The constant b denotes the number of given realizations $\mathbf{x}^{(i)} \in \mathcal{X}^{|X|}$, i.e. the number of parallel processed training instances or the number of instances whose most probable realization of Y should be predicted. If the

[1] 2^U denotes the power set of a set $U \subseteq V$.

underlying graph contains no loops, LBP will converge after a number of iterations which only depends on the number of edges. Otherwise, the algorithmen is not guaranteed to converge and is therefore terminated after a fixed number I of iterations. In both cases, the (approximated) single node marginals may be computed with $p(\mathbf{y}_v|\mathbf{x}^{(i)}) \propto m_{f\to v}(\mathbf{y}_v|\mathbf{x}^{(i)})m_{v\to f}(\mathbf{y}_v|\mathbf{x}^{(i)})$. We now present two appraoches for parallel LBP.

Data-Parallel LBP To compute the factor messages in a data-parallel manner, $|F| \times \max_{f\in F}|\Delta(f)| \times |\mathcal{Y}|$ concurrent blocks are instantiated, each with b threads. In this setup, thread i in block (f, v, y) computes message $m_{f\to v}(y|\mathbf{x}^{(i)})$. The time complexity per thread is $\mathcal{O}(\max_{f\in F}|\Delta(f)||\mathcal{Y}|^{|\Delta(f)|-1})$. Afterwards, the variable messages are computed likewise with $|V| \times \max_{v\in V}|\Delta^{-1}(v)| \times |\mathcal{Y}|$ concurrent blocks.

Thread-Cooperative LBP Considering the summation in Eq. 1, one may observe that all outgoing messages of one factor node have several terms in common. This observation leads to the Thread-Cooperative version of LBP. In this approach, one block for each pair of factor node f and given instance $\mathbf{x}^{(i)}$ is launched. The number T of threads per block is a free parameter of the algorithm. Each thread computes $|\Delta(f)| \times |\mathcal{Y}|$ partial outgoing messages which are eventually merged. To do so, let $J(f, \mathbf{x})$ be the set $\{j(f, \mathbf{x}, \mathbf{y}) = f(\mathbf{y}|\mathbf{x})\prod_{v\in\Delta(f)} m_{v\to f}(\mathbf{y}_v|\mathbf{x}) : \forall \mathbf{y} \in \mathcal{Y}^{|\Delta(f)|}\}$. A thread $t, 1 \le t \le T$ in block $(f, \mathbf{x}^{(i)})$ computes $|J(f, \mathbf{x}^{(i)})|/T$ elements of $J(f, \mathbf{x}^{(i)})$. Once such an $j(f, \mathbf{x}^{(i)}, \mathbf{y})$ is computed, the partial outgoing messages of thread t may be updated with Eq. 3.

$$m^{(t)}_{f\to v}(\mathbf{y}_v|\mathbf{x}^{(i)}) \quad += \quad \frac{j(f, \mathbf{x}^{(i)}, \mathbf{y})}{m_{v\to f}(\mathbf{y}_v|\mathbf{x}^{(i)})}, \forall v \in \Delta(f) \tag{3}$$

When all $|J(f, \mathbf{x}^{(i)})|$ have been processed cooperatively by all threads in a block, the partial messages are merged by parallel reduction. This results in a time complexity of $\mathcal{O}(\max_{f\in F}|\Delta(f)||\mathcal{Y}|^{|\Delta(f)|}T^{-1} + |\Delta(f)||\mathcal{Y}|\log T)$ per thread. Here, the left term is an outcome of computing a subset of $J(f, \mathbf{x}^{(i)})$ and the right term is introduced by the final reduction and may be intrepreted as parallelization overhead. Finally, one may recognize the possibility to launch $T \times R$ threads per block instead of T, in order to perform the updates in Eq. 3 concurrently for all neighbors. This introduces the factor R^{-1} to the left term of the time complexity. The just described algorithm is called *Thread-Cooperative LBP*. We now give a comparison of the LBP variants described above as well as the Thread-Cooperative Forward-Backward (FB) algorithm, which is a specialized LBP variant for Linear-Chain structures. The messages are propagated sequentially from the left-most to the right-most node in the graph and back, which reduces the number of blocks from $|F| \times b$ to b. With the exception of FB which is exact, we approximated the marginals with $I = \sqrt{|F|}$. Thread-Cooperative LBP was configured with $T = 16$ and $R = 2$. The values where determined heuristically. The optimal values for T and R, in terms of runtime, heavily depend on the actual GPU and the actual factor graph. They may be obtained by a grid search, constrained on solutions which satisfy $((T \cdot R) \bmod P) = 0$ and $B \cdot T \cdot \max_{f\in F}|\Delta(f)| \cdot |\mathcal{Y}| \le S$, where B is the desired number of

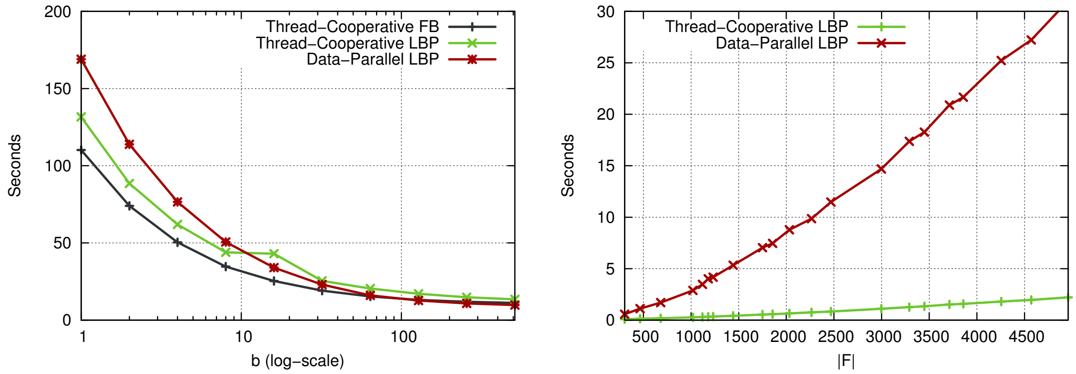


Figure 1: The left plot shows how the runtime evolves on a fixed $|F|$ and increasing b. The right plot shows how runtime evolves for fixed $b = 1$ and increasing $|F|$.

concurrently scheduled blocks per MP. Figure 1, shows the results on the whole CoNLL-2000 data set [6] (left) and single instances of Chen Yanover's PROTEIN data set [7]. While all approaches scale well when the number of concurrently processed instances increases, only the Thread-Cooperative variant scales with an increasing number of factor nodes. The second plot looks similar for different fixed values of b. This example with $b = 1$ should show, that the Data-Parallel LBP is not able to distribute the workload of one single instance equally among all Cores of a GPU. The difference in runtime becomes higher with increasing complexity of one single instance. Future experiments will involve synthetic data sets to evaluate how the Thread-Cooperative LBP performs when either $|\mathcal{Y}|$ or $\max_{f \in F} |\Delta(f)|$ is increased. It is also planned to integrate our approach into GraphLab [3] to have a broader platform for comparison with other inference methods.

References

[1] Mark Harris. Optimizing Parallel Reduction in CUDA. NVIDIA Corporation, 2008.

[2] Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. on Infor. Theory*, 47(2):498–519, 2001.

[3] Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M. Hellerstein. Graphlab: A new parallel framework for machine learning. In *Conference on Uncertainty in Arti. Intell. (UAI)*, California, July 2010.

[4] NVIDIA Corporation. *CUDA Programming Guide 4.0*. June 2011.

[5] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[6] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 shared task: chunking. NJ, USA, 2000. Assoc. for Comp. Linguistics.

[7] Chen Yanover, Ora Schueler-Furman, and Yair Weiss. Minimizing and learning energy functions for Side-Chain prediction. Berlin, 2007. Springer.

Dynamic Adaption of Ubiquitous System Software for Resource Efficiency

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

System software for classic embedded systems is optimized for minimal resource consumption by statically tailoring it to the application a priori known application demands. Ubiquitous Systems however, while being subject to the same resource constraints, have to deal with dynamically changing context-dependent requirements. My research therefore focuses on the dynamic adaption of resource management strategies to changing application requirements, using machine learning and dynamic aspect-oriented programming.

1 Safe Dynamic AOP for Kernel Data Acquisition

Applying machine learning to infer the current application context in order to adapt resource management strategies may require information about events in the operating system kernel, which has to be *instrumented* for this purpose (i.e., augmented by code which extracts and collects context data). Since manual instrumentation is tedious, instrumentation languages like *dtrace* [1] or *SystemTap*[1] are used. These are reminiscent of aspect-oriented programming (AOP) [3], providing a *pointcut language* for specifying sets of *join points* (interesting points in the control flow) that can be instrumented with the data collection code. These are, however, special-purpose languages. Real AOP languages integrate well with the programming language they instrument, while also exhibiting a much more powerful pointcut language.

[1] http://sourceware.org/systemtap/

domain	*issue*	*check*
content	assignments to context variables	language
	assignments to dereferenced pointers, including arrays	static
	calls to non-trusted functions and calls to function pointers	
	around()-advice with a path not executing exactly one tjp->proceed()	
	dereferencing of pointers referencing unmapped locations (triggers an exception)	dynamic
	member calls on objects without valid VMT pointer	
	endless recursion	
time	endless loops or recursions	

Table 1: Hazards to instrumented code in AspectC++

An important issue of instrumentation, is to ensure that the instrumentation does not interfere with the instrumented code, neither in the *content* nor in the *time* domain.

content Instrumentation must not directly alter the kernel's state. While this can be achieved by language mechanisms in type-safe languages, typical languages used for kernel development, like C and C++, allow direct manipulation of memory by the use of pointers, effectively circumventing the language-enforced access restrictions. Inside the kernel, this can have effects ranging from crash to permanent damage.

time While instrumentation is not possible without altering the timing behavior, user-perceivable effects on performance and particularly endless loops are not acceptable.

In dtrace, the instrumentation code runs in in a virtual machine (VM), which enforces these restrictions at runtime. Furthermore, the VM does not allow forward jumps in the compiled instrumentation code, which effectively prevents loops and particularly endless loops. SystemTap scripts are written in a C-like language, restricted to simple data types and compiled to a kernel module. The only data structures are managed associative arrays. Pointers can be dereferenced, but no writing is allowed to dereferenced pointers. Dynamic checks are inserted during the compilation process, like checking if dereferenced pointers are valid and counting/checking loop iterations and call stack depth.

Aspect languages for C and C++ (like AspectC++ [4]) allow violations in both domains (see Table 1). Therefore I intend to extend AspectC++ by a safe dialect, using a combination of language restrictions and dynamic checks. While *language*-checkable hazards require only a small change in the aspect weaver, the *static*ally checkable issues require an analysis of the instrumentation code. For the *dynamic* checks, code has to be generated (e.g., a counter for loops and nested function calls), which, however, means additional overhead. Techniques for control flow and data flow analysis can further help to avoid most of the dynamic checks. A trusted library replaces missing language features (e.g., with a safe collection type that can be used instead of arrays).

The next step based on this are more fine-grained and configurable access restrictions in order to 1.) have better control over what kind of data can be collected and 2.) allow controlled modification of the kernel's data structures in order to adapt operating system strategies. This fine-grained control should be applicable to other aspect languages.

2 Adaption Experiments

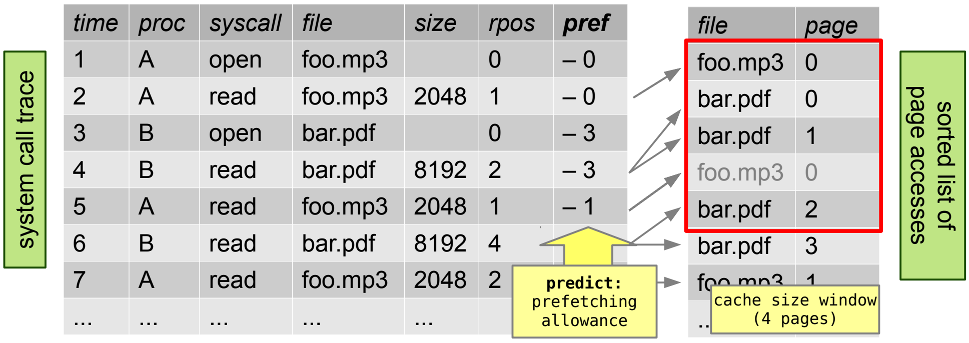
Adaption Experiments help to understand what kind of system strategies and decisions are amenable to context-dependent optimization, the potential they offer for resource savings, and which kind of data is interesting for this purpose.

One of the first experiments in project A1 of the SFB 876 was the improvement of file system *read-ahead*. For this purpose, we traced several *system calls* related to file system and process management. Every system call opening a file was labeled with the subsequent read behavior (sequentially, fully sequentially, random or not at all), which was then used as input to supervised machine learning. [2] The strategy was evaluated in a system simulator, which triggered prefetching a file as a whole if it was predicted to be fully read sequentially. Although the model's prediction accuracy was good, the strategy itself turned out to be less effective than the Linux's original strategy. Since most processes do not read a file at once, even if they eventually read the file completely sequentially, pages aged and where replaced by LRU before they would be accessed. Also, the costs for mispredictions where disproportionately high. Linux however, uses an *adaptive strategy* that incorporates a the actual reading pace regarding a certain file.

An alternative strategy that I am currently evaluating also prefetches aggressively but takes the aforementioned reading pace into account. Furthermore, it allows prefetching every time a file is accessed, not only when it is opened. For that purpose, the trace output is extended by the *prefetching allowance* for every system call that read from a file. The prefetching allowance is a set of consecutive pages, which are guaranteed to 1.) be accessed before they are replaced and 2.) not lead to replacement of a page which would have been accessed before all pages in the allowance were read.

The prefetching allowance is calculated by considering the set of pages currently *ideally* being in the cache. For an assumed cache size of *n* pages, it consists of the *n* pages with the smallest *forward distance* (i.e., the temporal distance between the current time and the next time the page is accessed). The trace output (left table in Figure 1, featuring 2 concurrent *proc*esses) is therefore transformed into a list of page accesses (right table), ordered by their respective forward distance. The current cache set always consists of the first *n* entries of the access list, plus one for each duplicate access to a page.

The trace output is now processed for every call which opens, reads or seeks in a file: First, its prefetching allowance size is calculated by counting all and only sequential page accesses (e.g., an access after the next seek or a positional read does not count). The set of pages now *allowed* for prefetching now starts at the next page to be read (*rpos*) plus the prefetching allowance size. The page accesses satisfying the read request (preceding



system call trace
sorted list of page accesses
predict: prefetching allowance
cache size window (4 pages)

time	proc	syscall	file	size	rpos	pref
1	A	open	foo.mp3		0	– 0
2	A	read	foo.mp3	2048	1	– 0
3	B	open	bar.pdf		0	– 3
4	B	read	bar.pdf	8192	2	– 3
5	A	read	foo.mp3	2048	1	– 1
6	B	read	bar.pdf	8192	4	
7	A	read	foo.mp3	2048	2	
...	...	...	...	...		

file	page
foo.mp3	0
bar.pdf	0
bar.pdf	1
foo.mp3	0
bar.pdf	2
bar.pdf	3
foo.mp3	1
..	

Figure 1: Calculation of prefetching allowance.

rpos) are then deleted from the access list. For the implementation in an operating system, I plan is to infer the prefetching allowance using machine learning.

However, the prefetching allowance alone does not automatically yield a good strategy for prefetching. Always prefetching everything which is inside the allowance would lead to several small file accesses. However, accessing a bigger number of sequential blocks at once is generally more energy-efficient. Therefore, prefetching only takes place when the ratio between the number of pages which are prefetchable, but not yet prefetched and the number of pages, which are prefetched, but not yet read, exceeds a certain threshold. Altering this threshold means trading off responsivity (lower probability of I/O waits) and energy efficiency (lower number of I/O requests). Although the strategy has not been integrated into an operating system yet, an evaluation in a full-system simulator with a threshold of 50% showed that it actually does not lead to unnecessary page accesses, which were the reason the first approach performed that poor.

References

[1] B.M. Cantrill, M.W. Shapiro, and A.H. Leventhal. Dynamic instrumentation of production systems. In *Proc. of USENIX ATC*, pages 2–2. USENIX Assoc., 2004.

[2] P. Fricke, F. Jungermann, K. Morik, N. Piatkowski, O. Spinczyk, M. Stolpe, and J. Streicher. Towards adjusting mobile devices to user's behaviour. *Springer LNCS/LNAI 6904: Mining and Modeling of Ubiquitous Data in Social Media*, 2011.

[3] G. Kiczales, J. Lamping, A. Mendhekar, C. Maeda, C. Lopes, J.M. Loingtier, and J. Irwin. Aspect-oriented programming. *ECOOP'97*, pages 220–242, 1997.

[4] O. Spinczyk, A. Gal, and W. Schröder-Preikschat. Aspectc++: an aspect-oriented extension to the c++ programming language. In *Proc. of 40th Int. Conf. on Tools Pacific: Objects for internet, mobile and embedded applications*, pages 53–60. Australian Computer Society, Inc., 2002.

Subproject A2
Algorithmic aspects of learning methods in embedded systems

Christian Sohler Jan Vahrenhold

Cache-oblivious construction of semi-separated pair decompositions

Sylvie Temme
Lehrstuhl 11 - Algorithm Engineering
Technische Universität Dortmund
Sylvie.Temme@tu-dortmund.de

Given a set P of n points in $\mathbb{R}^d$ and a parameter $0 < \varepsilon < 1$, we show how to construct a $\frac{1}{\varepsilon}$-semi-separated pair decomposition ($\frac{1}{\varepsilon}$-SSPD) for P using an expected number of $\mathcal{O}\left(\text{sort}\left(\frac{n}{\varepsilon^d}\right)\log n\right)$ memory transfers in the cache-oblivious model of computation; here, sort (x) denotes the complexity of sorting x elements. The algorithm is based on a randomized recursive RAM-model algorithm of Abam and Har-Peled [1].

Many problems can be modeled by interpreting the data objects as a set P of n points in $\mathbb{R}^d$. If it is reasonable for such problems to consider approximate distances between the points, a so-called well-separated pair decomposition (WSPD) can be used as a compact representation of the pairwise distances. A WSPD for P is a set of pairs of subsets of P, which represents implicitly a partition of all pairs of different points $p, q \in P$ and in which each pair of point sets is "well-separated" in the sense that the distance between the two point sets is "large" compared to the maximum diameter of the two point sets. Thus, for two well-separated sets P_1 and P_2, the distance between two fixed points $p_1 \in P_1$ and $p_2 \in P_2$ approximates the distance between any two points $q_1 \in P_1$ and $q_2 \in P_2$ up to a precision depending on the input parameter ε.

The notion of WSPD was introduced by Callahan and Kosaraju [3]. They also presented an algorithm for constructing a WSPD with a linear number of pairs in $\mathcal{O}\left(n \log n + \frac{n}{\varepsilon^d}\right)$ time. There are several applications of WSPDs, for example constructing t-spanners, constructing approximate minimum spanning trees or computing the nearest neighbor for each point of a given point set – see, e.g., the book by Narasimhan and Smid [5].

For certain tasks it is reasonable to use semi-separated pair decompositions (SSPD) instead of WSPDs. In an SSPD the pairs are "semi-separated" in the sense that the distance between the two point sets is "large" compared to the minimum (instead of maximum) diameter of the two point sets [6]. With this definition, well-separated pairs of point sets are also semi-separated, but not necessarily vice versa.

Abam and Har-Peled [1] presented a randomized recursive construction of an SSPD for P with the following properties: Every point in P participates with high probabilty in $\mathcal{O}\left(\frac{1}{\varepsilon^d}\log n\right)$ pairs of the SSPD, the number of pairs in the SSPD is $\mathcal{O}\left(\frac{n}{\varepsilon^d}\right)$ and the expected weight of the SSPD (i.e., the total size of all point sets in the SSPD) is $\mathcal{O}\left(\frac{n}{\varepsilon^d}\log n\right)$. Under the assumption that the floor-function can be evaluated in constant time, the expected construction time is in $\mathcal{O}\left(\frac{n}{\varepsilon^d}\log n\right)$.

The I/O-model introduced by Aggarwal and Vitter [2] is the standard model for analyzing algorithms for hierarchical memory. In this model the memory system is modeled as a two-level hierarchy which consists of a small fast memory of size M and a slower large memory of size at least n. The data is transfered in blocks of size B. The efficiency of an algorithm is measured in the number of such memory transfers.

The cache-oblivious model introduced by Frigo *et al.* [4] generalizes this two-level model to a multilevel memory hierarchy. Algorithms are designed as in the RAM-model, i.e., without any knowledge of (the existence of) the parameters M and B, but analyzed as in the I/O-model. Thus, if an algorithm is optimal, it is optimal on all levels of the hierarchy since it does not depend on any particular values of M and B.

We modified the algorithm of Abam and Har-Peled, so that the algorithm is efficient in the sense of the cache-oblivious model.

The basic idea of the algorithm of Abam and Har-Peled is to partition the input point set (of each recursive call) into three particular point sets that lead to an easier construction of parts of the desired SSPD. Two of the point sets are contained in balls of bounded diameter; for these sets, a WSPD (which is also an SSPD by definition) is computed by constructing and traversing a (partial) quadtree. Also, we can directly compute SSPD pairs for one of these sets and the third set. In the remaining cases, an SSPD is computed recursively. In a postprocessing step, the number of pairs of the SSPD is reduced by merging "matching" pairs. To find matching pairs, a hashing data structure is used. To manage and to merge the matching pairs, the algorithm relies on managing multiple linked lists.

In the cache-oblivious model, we face the following challenges: The original algorithm, which simultaneously constructs and traverses the quadtree, does not exhibit spatial coherence in the memory access pattern. In addition, tree traversal has a sorting lower bound in the cache-oblivious model. Moreover, while there are algorithmic components for hashing and maintaining multiple lists in the RAM-model (and also in the I/O-model), we are not aware of matching algorithmic building blocks that are efficient in the cache-oblivious model.

We performed the following modifications to reach the temporal and spatial coherence needed in the cache-oblivious model: Instead of simultaneously constructing the quadtree and computing the WSPD-pairs as in the original algorithm, our algorithm first fully constructs the (partial) quadtree and then computes a WSPD for the partition induced by the nodes of the tree. To this end, we first need to (conservatively) estimate the depth of the quadtree. Both the construction of the tree and the computation of the WSPD then are performed level-by-level: the construction is done bottom-up, while the computation is done top-down. At the beginning of the construction of the quadtree the points in P (which will be contained in the leaves of the quadtree) are sorted according to the so-called $\mathcal{Q}$-order. Afterwards, the quadtree can be constructed easily by simultaneously scanning each level and writing the next level. During the computation of the WPSD we rely on cache-oblivious sorting and scanning to efficiently extract the data relevant for computing the WSPD pairs corresponding to the current level of the tree.

In the postprocessing step, in which the number of pairs is reduced by merging "matching" pairs, the original algorithm uses hashing to find the matching pairs. Since a cache-oblivious counterpart is not available, we work in feature space where we can find the matching pairs using cache-oblivious batched orthogonal range queries. Using the same approach, we can also efficiently address the problem of merging the lists of matching pairs.

With this modifications the algorithm requires $\mathcal{O}\left(\text{sort}\left(\frac{n}{\varepsilon^d}\right)\log n\right)$ memory transfers in the cache-oblivious model; thus, we have obtained the first efficient algorithm for constructing an $\frac{1}{\varepsilon}$-SSPD in the cache-oblivious model.

References

[1] Mohammad Ali Abam and Sariel Har-Peled. New constructions of SSPDs and their applications. In *Proceedings of the 2010 Annual Symposium on Computational Geometry*, SoCG '10, pages 192–200, New York, NY, USA, 2010. ACM.

[2] Alok Aggarwal and Jeffrey S. Vitter. The input/output complexity of sorting and related problems. *Commun. ACM*, 31:1116–1127, September 1988.

[3] Paul B. Callahan and S. Rao Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. *J. ACM*, 42:67–90, January 1995. A preliminary version appeared in: Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing (1992), 546–556, ACM.

[4] Matteo Frigo, Charles E. Leiserson, Harald Prokop, and Sridhar Ramachandran. Cache-oblivious algorithms. In *Proceedings of the 40th Annual Symposium on Foun-*

dations of Computer Science, FOCS '99, pages 285–298, Washington, DC, USA, 1999. IEEE Computer Society.

[5] Giri Narasimhan and Michiel Smid. *Geometric Spanner Networks*. Cambridge University Press, New York, NY, USA, 2007.

[6] Kasturi R. Varadarajan. A divide-and-conquer algorithm for min-cost perfect matching in the plane. In *Proceedings of the 39th Annual Symposium on Foundations of Computer Science*, FOCS '98, pages 320–331, Washington, DC, USA, 1998. IEEE Computer Society.

Spatio-temporal and progressive data analysis for massive datasets

Andreas Thom
Lehrstuhl 11 - Algorithm Engineering
Technische Universität Dortmund
andreas.thom@tu-dortmund.de

In this project, we work towards a (semi-)automated geometric approach to exploring astronomical data based on the *Sloan Digital Sky Survey* (SDSS). In the context of a close collaboration with the department of Physics and Astronomy of the Ruhr-University Bochum, we develop progressive algorithms and data-structures supporting astronomical exploration. Our first method is based on the principle of connected component analysis and can identify regions of high density based on photometric data (or any arithmetical expression of photometric features). As part of the ongoing work, we have developed a prototypical implementation of our approach that can be used by the domain experts for exploration and verification of the results. The first results obtained by this seems to indicate that this new approach can lead to new astronomical awareness due to the (semi-)automated detection of structures of interest in the multidimensional data space.
In a related, yet orthogonal project, we also work on clustering of soccer players' trajectories. We aim at developing different methods for spatio-temporal clustering that are suited to identify similar sub-trajectories with respect to predetermined characteristics or frequently occurring motion patterns of the soccer player.

The (semi-)automated analysis of data sets has become an increasingly important issue for researchers in astronomy [1]. This is especially the case for massive data sets obtained from the (SDSS) (data volume comprises $49,6$ TB in its current release) which is said to be "one of the most ambitious and influential surveys in the history of astronomy" [6]. Developing solutions for astronomical problems requires a focus on developing progressive analysis methods that can cope with the challenges brought by large-scale multidimensional data. Handling such data efficiently requires the usage of large amounts of main

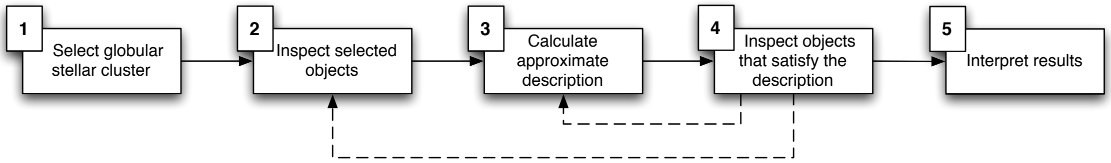
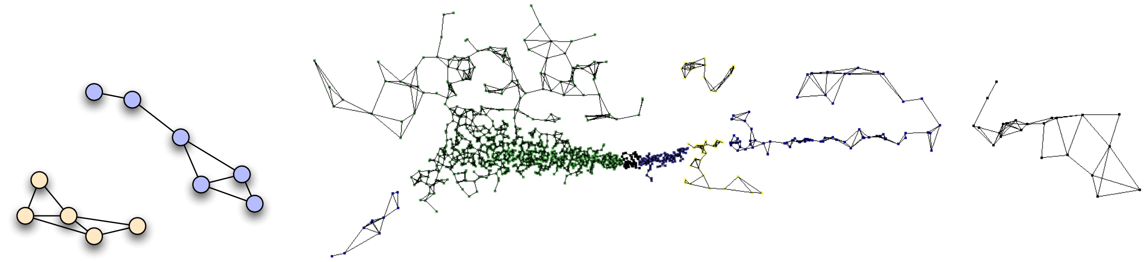

1
Select globular stellar cluster
2
Inspect selected objects
3
Calculate approximate description
4
Inspect objects that satisfy the description
5
Interpret results

Figure 1: Workflow for detecting *Streams*. The shown sequence of steps 2-4 is valid for multiple feature spaces and is supported by feedback loops (dashed lines).

memory, which is fast and expensive, or databases access to which is slower. An algorithm is called progressive if the quality of the result concerning a given quality-measure is strictly monotonically increasing with the progress of the algorithm. Such progressive algorithms, see, e.g., [5], are predominantly developed in the field of databases. In addition, modeling astronomical questions requires a domain knowledge of the physical backgrounds of the problems which needs to be deeply integrated into the analysis method so that the already mentioned cooperation with the Ruhr-University Bochum is absolutely essential. To verify our results and methods, we thus have developed a prototypical implementation of our approach which allows domain experts to test the developed methods and perform a visual inspection of the results.

The first astronomical task we deal with is to find *Streams* induced by globular clusters. *Streams* arise during the the growth of galaxies by gravity interaction between massive objects (e.g., the Milky Way) and smaller objects (e.g., dwarf galaxies) which are cracking apart. Our procedure to detect and characterize *Streams* is schematically shown in Figure 1. The first step is the selection of a globular cluster based on the area of high density in the two-dimensional feature space (*ra*,*dec*). As a running example, we select the objects of the *Palomar5* globular cluster from the SDSS which is the basis for our analysis. The goal is identify and geometrically described high-density regions of interest in astrophysically relevant feature (sub-)spaces. We model regions of high density and/or special interest using an undirected graph on the stellar objects in question. We compute connected components based on the symmetric adjacency with respect to the k-nearest neighbors of each object. A simple example of this method is given in Figure 2a with $k = 4$. The edges show the symmetric adjacency of the connected objects and the data set is split into two connected components. Figure 2b shows a visualization of the connected components of *Palomar5* which induce a *Stream* already described in the literature. The resulting connected components in a feature space related to the *Hertzsprung-Russell-diagram*, which is commonly used for data analysis in astronomy, have been verified by the astronomers to identify important structures and regions, e.g., the break of the main sequence.

For each connected component, different geometric representations can be calculated. For the purpose of enabling a progressive data exploration, the geometric representation is first roughly approximated and in the following iterations of the workflow (see Figure 3) aligned to the real shape of the cluster in a stepwise fashion. This representation is then utilized to select astronomical objects form the SDSS which fit in the geometric repre-



(a) Connected components with $k = 4$.

(b) Connected components of *Palomar5* concerning photometric features $psfMag_g - psfMag_r$,$psfMag_r$, commonly used for data analysis in astronomy.

Figure 2: Connected components concerning the symmetric adjacency with $k = 4$ and a visualization of the result with $k = 7$ on selection of 1504 objects of *Palomar5*.

sentation in the specific feature space. One can see that the selectivity of queries from Figure 3 (a) to (d) is improved but at the costs of a higher representation complexity. This approach reflects the main idea of a progressive algorithm. The first rough approximation can be computed quickly and leads to fast but less accurate results. One can improve the quality of the results by investing more computation time for calculating more detailed representations, which cause higher costs with regard to database queries and selections. This two counterbalancing measures will be one important object of investigation for our future work.

To reduce the complexity of our calculations and facilitate the visual examination of the results, we rely on a method to build a reasonable sub-sample of the data set which preserve the original density distribution of the underlying point set. To achieve this we make use of a method originally located in the field of surface reconstruction [3]. The next steps of our work focus on the efficient calculation of the geometric representations and the progressive refinement of their selectivity.

The second field of work deals with the analysis of soccer players' trajectories. Here, the goal is to finding characteristic motion patterns of one player or its main area of movement. To address the problem, we have developed two approaches. The first variant segments the whole trajectory into sub-trajectories defined by time or length that can overlap or are chosen disjunctively. Then, each of these sub-trajectories is simplified with the *Min-ϵ*-trajectory simplification algorithm which constructs an approximate curve regarding a given size m which consists of at most m line segments with minimum error. Afterwards the resulting simplified sub-trajectory is mapped to a point in $2m$-dimensional space with d being the number of vertices of each sub-trajectory. An alternative simplification method yields a simplification of arbitrary size but with a given approximation bound [2]. These constructed feature points are clustered using, e.g., the density based clustering algorithm DBSCAN [4] to group similar trajectories. In a second approach, our algorithm construct feature vectors of the sub-trajectories based upon the minimum enclosing box, the median of the trajectory, or other features that represent characteristics of the shape of a trajectory. Again we can first simplify the trajectory and then break

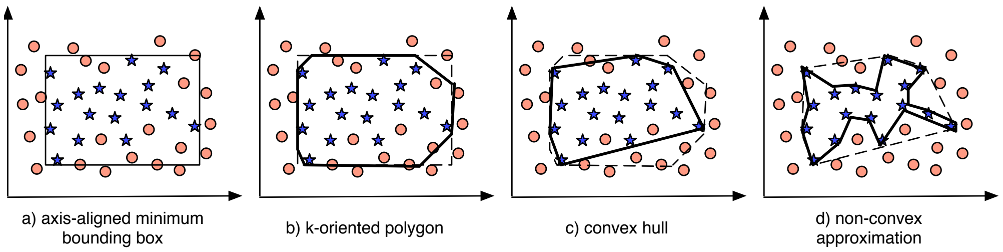

a) axis-aligned minimum bounding box
b) k-oriented polygon
c) convex hull
d) non-convex approximation

Figure 3: Sequential arrangement of nested approximations. The objects of a connected component in feature space are marked by '*star*'. Other objects are marked by '∘'. In (b)-(d) the approximation of the previous step is denoted by the dashed lines.

it into sub-trajectories of prescribed length. Afterwards we continue as discussed in the first variant. This project is conducted in collaboration with Dr. Joachim Gudmundsson, principal researcher at National ICT Australia and the University of Sydney.

References

[1] Nicholas M. Ball and Robert J. Brunner. Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(7):1049–1106, 2010.

[2] Ovidiu Daescu and Ningfang Mi. Polygonal path approximation: A query based approach. In *Proceedings of the 14th International Symposium on Algorithms and Computation*, Lecture Notes in Computer Science, pages 36–46, 2003.

[3] Daniel Dumitriu, Stefan Funke, Martin Kutz, and Nikola Milosavljevic. How much geometry it takes to reconstruct a 2-manifold in $\mathbb{R}^3$. *ACM Journal of Experimental Algorithmics*, 14, May 2009, Article 2.2, 17 pages.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[5] Dimitris Papadias, Yufei Tao, Greg Fu, and Bernhard Seeger. Progressive skyline computation in database systems. *ACM Transactions on Database Systems*, 30(1):41–82, March 2005.

[6] Sloan Digital Sky Survey (SDSS). `http://www.sdss.org`, 2011.

Subproject A3
Methods for Efficient Resource Utilization in Machine Learning Algorithms

Jörg Rahnenführer   Peter Marwedel

Multivariate survival models with high dimensional genomic covariates

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

An important application of high dimensional gene expression measurements is the prediction of survival times and the interpretation of the variables. When the response variables are censored survival times, an appropriate hazard framework is required. The largest problem in this context is the typically large number of genes compared to the number of observations. Thus we apply feature selection procedures in order to generate predictive models for future patients. This approach targets to identify models with high prediction accuracy and at the same time low model complexity. Implementations of new statistical methods mostly appear first as an addon package for R [10], the lingua franca of statistical computing. The high-dimensional feature space combined with the need for resampling methods induce high computational costs. Furthermore the analysis of subgroups of samples increases complexity. Therefore fast and parallelized algorithms are essential in this domain. Because the analysis of multiple high dimensional datasets is a short term goal the consideration of computational costs becomes even more important.

In the following we assume that we have a sample size of n patients and p explanatory variables $X = (X_1, \ldots, X_p)$. The explanatory variables are either clinical covariates (age, tumor status, etc.) or gene expression values. We model the hazard function $h(t)$ using the Cox Proportional Hazard model [4]

$$h(t \mid x) = h_0(t) \exp(\beta' x),$$

where $h_0(\cdot)$ is an arbitrary baseline hazard function and $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of regression coefficients. In the classical setting with $n > p$, the regression coefficients are estimated by maximizing the log partial likelihood

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta' x_i - \log \left(\sum_{j \in R(t_i)} \exp\left(\beta' x_j\right) \right) \right] . \qquad (1)$$

For patient i, this expression contains the possibly censored failure time t_i, the (non-censoring-)indicator δ_i (equal to 1 if t_i is a true survival time and to 0 if it is censored) and the vector of covariate values x_i. Further, $R(t_i)$ is the risk set at time t_i; this is the set of all patients who have neither failed yet nor been censored. The value of $\beta' x_i$ is called prognostic index or risk score of patient i.

In a high dimensional feature space with $n \ll p$, there exists no distinct solution to (1). Moreover, with more than 20 000 covariates, collinearity complicates numerical optimization attempts.

One approach is to use regularization. This includes optimization using the Lasso [11]

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left(ll\left(\beta\right) - \lambda \sum_{k=1}^{p} |\beta_k| \right) .$$

Here, $ll(\cdot)$ denotes the log partial likelihood. Ridge regression [8] is very similar, using a quadratic instead of an absolute penalty:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left(ll\left(\beta\right) - \lambda \sum_{k=1}^{p} \left(\beta_k\right)^2 \right) .$$

The tuning parameter λ can in both Lasso and Ridge regression be estimated simultaneously using the K-fold cross validated partial likelihood [12]

$$\mathrm{CVPL}(\lambda) = \sum_{k=1}^{K} \left[ll(\hat{\beta}_{(-k)}(\lambda)) - ll_{(-k)}(\hat{\beta}_{(-k)}(\lambda)) \right] ,$$

where the subscript $(-k)$ indicates calculation or estimation on all samples except of those in the k-th fold. In Lasso regression most of the components of β get shrunken to exactly zero resulting in sparse and interpretable models. Ridge regression on the other hand has typically better prediction performance [3] but does not provide automatic selection of important covariates. The simultaneous use of both penalty terms is called elastic net [14].

An alternative approach is the component wise likelihood-based boosting [1]. In each iteration the component of β which improves the fit most, gets updated utilizing penalized partial likelihoods. In the subsequent step the updated β is used as a model offset. The number of boosting steps M is also selected by cross-validation.

Greedy search algorithms such as forward search and univariate search usually get outperformed in terms of prediction performance and model accuracy [2, 3]. Furthermore they neglect completely the correlations between covariates [7]. However, the simplicity of these methods implies good interpretation possibilities.

The evaluation of all methods depends on a preceding split of the patients into training and test datasets. The Brier Score [6] evaluates the differences of predicted and observed survival times. The concordance index [9, 13] judges the ranking of predicted survival times. Another performance measure is based on the risk score: the goodness of fit of a cox model on the test data with the risk score as single covariate points to the performance [2]. In order to reduce the effect of the random split into training and test data, the steps of splitting, model fitting and evaluation must be repeated several times, whereat each model fit consumes a not negligible amount of computational time. New technologies such as Exon or SNP arrays raise the number of covariates further to 10^6 covariates.

We applied these techniques to several cancer types in the recent past, publications on the findings are currently under review. A future goal of our research will be based on the concurrent analysis of multiple datasets. As more and more freely accessible and suitable datasets emerge the models are not limited anymore to be learned and evaluated on single datasets. It is now rather possible to evaluate the portability of models or assess datasets in regard to their generalizing properties. First results, utilizing only clinical covariates, draw a clear distinction between datasets in terms of generalization and also uncovered models specialized on specific cohorts.

For our analysis we use the programming language R [10]. In combination with the BioConductor repository [5], R provides unchallenged the most comprehensive framework for high dimensional survival analysis. Unfortunately, R was not designed with high dimensional data in mind. But the nested cross-validation, the demand of specialized models for patient-subgroups as well as the concurrent analysis of multiple datasets makes gene expression measurement analysis a high performance computing task.

References

[1] H. Binder and M. Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC bioinformatics*, 9:14, January 2008.

[2] Hege M. Bøvelstad, S. Nygård, and Ornulf Borgan. Survival prediction from clinicogenomic models–a comparative study. *BMC bioinformatics*, 10:413, January 2009.

[3] Hege M Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjaerde. Predicting survival from microarray data – a comparative study. *Bioinformatics (Oxford, England)*, 23(16):2080–7, August 2007.

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

[5] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[6] T. Gerds and M. Schumacher. Efron-type measures of prediction error for survival analysis. *Biometrics*, 63(4):1283–7, December 2007.

[7] Isabelle Guyon and Andre Elisseefi. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(7-8):1157–1182, October 2003.

[8] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[9] H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1):50–60, 1947.

[10] R Development Core Team. R: A Language and Environment for Statistical Computing, 2011.

[11] R Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4):385–95, March 1997.

[12] P J Verweij and H C Van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–14, December 1993.

[13] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[14] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005.

Optimizations for Timing Constraint Embedded Systems

Sascha Plazar
Computer Science 12
TU Dortmund University
D - 44221 Dortmund, Germany
sascha.plazar@tu-dortmund.de

The worst-case execution time (*WCET*) is a key parameter in the domain of real-time systems since its knowledge is crucial to verify if timing critical systems meet their deadlines. Today's embedded system applications are written in a high-level language and the code generated by compiler out of it has to be verified w.r.t. such timing constraints. If the application does not meet certain deadlines, the code has to be tuned, compiled and optimized again in another cycle of the design flow.

Hence, automatic compiler-based WCET optimization becomes a challenging research area. For this purpose, we propose a compiler framework for the development of novel WCET-driven optimizations. In this paper, the WCET-aware C Compiler framework *WCC* is introduced. Based on this framework, various low-level memory based optimizations as the *WCET-driven memory content selection* or *WCET-driven branch prediction aware code positioning* have been developed. Their functionality and achievable WCET reductions are presented as well.

Embedded systems are also often hard real-time systems for which the knowledge of an upper bound of the execution time is mandatory. This bound is called worst-case execution time (WCET) and is a key parameter for the development of tailored hardware platforms. Static analysis techniques have been developed to allow a safe estimation of the execution time of a program (WCET$_{est}$) since the real WCET can not be measured.

In the established design flow of embedded applications, programs are written in high-level languages for which timing constraints have to be verified. Usually, the code is

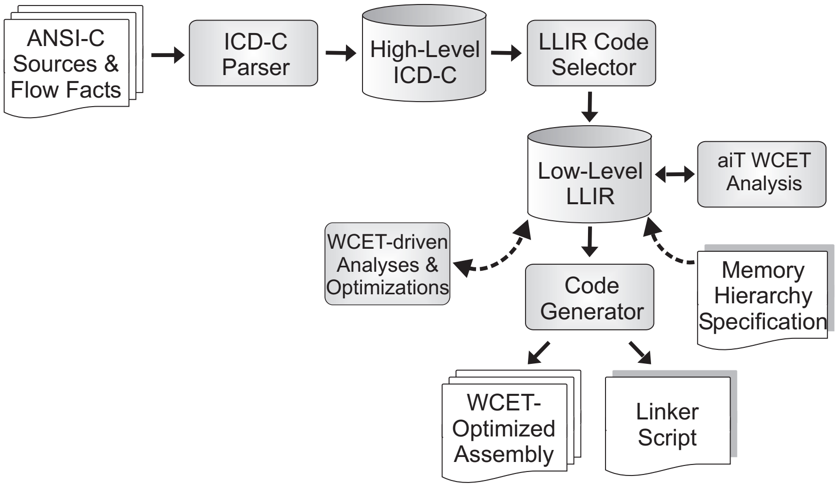


Figure 1: Workflow of the WCET-aware C compiler WCC

therefore manually fed into a sophisticated timing analyzer which determines its WCET. If it turns out that the program misses given deadlines, the programmer has to tune the program and start a new compilation/analysis cycle.

To relieve a developer from the burden of this error prone and time consuming optimization cycle, automatic compiler-based WCET optimizations can be used. WCET-driven optimizations need support of an underlying compiler to collect WCET data and to perform possible memory layout modifications. We propose the WCET-aware C compiler framework, called *WCC* [3], which is intended to assist the development of various high- and low-level WCET-driven optimizations. It is a compiler targeted at Infineon's TriCore TC1796 processor coupling AbsInt's static WCET analyzer *aiT* [1] which provides WCET$_{est}$ data that is imported into the compiler backend and made accessible for optimizations.

Figure 1 depicts WCC's internal structure. One or more files of a program are read in the form of ANSI-C source files with user annotations for loop bounds and recursion depths, called *flow facts*. These source files are parsed and transformed into the high-level intermediate representation (*IR*) called *ICD-C* [6]. At this level, the compiler frontend provides several standard compiler optimizations focusing on ACET minimization.

In the next step, the *LLIR Code Selector* translates the high-level IR into a low-level IR called *ICD-LLIR* [2]. Again, several standard compiler optimization can be performed – now on this TC1796-specific low-level IR.

To enable WCET-aware optimizations, *aiT* is employed to perform static WCET analyses on the low-level IR. Mandatory information about loop bounds and recursion depths is supplied by flow fact annotations. These flow facts are automatically translated from the high-level IR to the low-level IR and are always kept valid and consistent during each optimization and transformation step of the compiler.

Optimizations exploiting memory hierarchies as well as the static WCET analyzer require detailed information about available memories, their sizes and access times. For this purpose, WCC integrates a detailed memory hierarchy specification available at *ICD-LLIR* level. Finally, WCC emits WCET-optimized assembly files and its own linker script in order to generate the optimized binary.

In the following, the two low-level memory based WCET optimizations called *memory content selection* and *branch prediction aware code positioning* are presented.

WCET-driven Memory Content Selection Caches are widely used to bridge the increasingly growing gap between processor and memory performance. They store copies of frequently used parts of the slow main memory for faster access. But if, however, the code of a program is not suitably arranged in the address space and the memory accesses are random or widely spread over the address space, the performance can be also decreased by the usage of a cache.

Intelligent allocation of beneficial functions to cached memory areas and unfavorable functions to non-cached memory areas can ensure that functions whose WCET highly profits from a cache are not evicted from the cache by functions with a low benefit. This can lead to a faster execution and a decreased WCET due to a dramatically decreased number of cache misses.

We developed a novel WCET-driven cache-aware memory content selection algorithm [5] to decide which functions should be placed in a cached memory area in order to improve the worst-case I-cache performance. The proposed algorithm uses a greedy approach and evaluates the impact of executing a function from a cached memory area on the WCET_{est}. Changes on the worst-case execution path are taken into account in order to optimize along this path which allows an effective minimization of the WCET_{est}.

Applying this technique, we were able to achieve a WCET_{est} decrease of up to 20% and ensure that the performance of the optimized program is never worse than the original. On average, WCET_{est} reductions between 4% and 8% were achieved for cache sizes ranging from 5-20% of the overall code size.

WCET-aware Code Positioning In the past decades, embedded system designers moved from simple, predictable system designs towards complex systems equipped with caches, branch prediction units and speculative execution. This step was necessary in order to fulfill increasing requirements on computational power. We developed two novel WCET-driven code positioning algorithms [4] to rearrange the order of basic blocks of a function. They aim at improving the performance of static branch predictors and at avoiding unnecessary unconditional jumps. The first algorithm employs a genetic approach which starts with a random population. By exploiting the evolutionary techniques crossover and mutation, offspring individuals are generated which desirably converge to the optimal solution w. r. t. the WCET of a program.

Usually, evolutionary strategies can be implemented with small effort and often without understanding the mechanism behind the optimization problem. We employ an Evolutionary algorithm (*EA*) to understand the mechanisms behind the optimization problem and to explore the possible optimization potential of code positioning techniques before developing complex algorithms.

The second optimization is such a “complex” algorithm. A more promising order of basic blocks is determined based on an integer-linear programming (*ILP*) approach. The ILP explicitly models the WCEP as well as the impact on the branch prediction and thereby avoids repetitive WCET analyses. For the first time, both branch penalty costs and the amount of executed unconditional jumps are modeled in an ILP.

Applying these techniques, we were able to achieve a WCET decrease of up to 24.7%. On average, WCET reductions of 8.9% were achieved for the *EA* whereas the ILP achieves average WCET reductions of 6.7% were achieved.

References

[1] AbsInt Angewandte Informatik GmbH. Worst-Case Execution Time Analyzer aiT for TriCore. 2011. `http://www.absint.com/ait`.

[2] J Eckart and Robert Pyka. ICD-LLIR Low-Level Intermediate Representation. `http://www.icd.de/es/icd-llir`, 2011. Informatik Centrum Dortmund.

[3] Heiko Falk and Paul Lokuciejewski. A compiler framework for the reduction of worst-case execution times. *Journal on Real-Time Systems*, 46(2), October 2010.

[4] Sascha Plazar, Jan Kleinsorge, Heiko Falk, and Peter Marwedel. Wcet-driven branch prediction aware code positioning. In *Proceedings of the International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES)*, Taipei, Taiwan, 2011. [to appear].

[5] Sascha Plazar, Paul Lokuciejewski, and Peter Marwedel. Wcet-driven cache-aware memory content selection. In *Proceedings of the 13th IEEE International Symposium on Object/Component/Service-oriented Real-time Distributed Computing (ISORC)*, pages 107–114, Carmona / Spain, may 2010.

[6] Robert Pyka and Jörg Eckart. ICD-C Compiler Framework. `http://www.icd.de/es/icd-c`, 2011. Informatik Centrum Dortmund.

Optimization Potential for Runtimes of R Programs

Helena Kotthaus
Computer Science 12
TU Dortmund University
D - 44221 Dortmund, Germany
helena.kotthaus@tu-dortmund.de

The GNU R programming language is a de facto standard in the domain of statistical computing. R is a language with functional characteristics and a dynamic type system. These characteristics support the development of statistical algorithms and analysis at a high-level of abstraction. Like many dynamic languages, R programs are processed by an interpreter. Such an interpretation leads to an unacceptably slow execution of computation-heavy R programs. Our goal is to optimize the execution runtimes of R programs. Therefore, we want to develop a compiler toolchain as employed for imperative languages, such as ANSI C or Java. To reach the best possible optimization potential for runtimes of R programs, it is necessary to examine different compiler strategies. This report discusses those possible compiler strategies.

Due to its open-source nature, R has become very popular among statisticians. The *comprehensive R archive network* (CRAN) [5] contains a huge amount of R-packages which support the rapid development of statistical algorithms and analysis. Especially in the domain of bioinformatics R has become invaluable for analysis and evaluation of genomic data. For this purpose, *Bioconductor* [2], as an open-source software framework written in R, supports free R-packages that contain algorithms which are unrivalled up to now.

Although R is common in the domain of statistics, a big disadvantage lies in the performance penalty especially in the case of computation-heavy programs. R-programs are processed by an interpreter. This interpretation is responsible for performance loss, because every single program line has to be evaluated separately. R is an originally functional programming language with a dynamic type system. In contrast to imperative languages, like ANSI C, which could be translated directly into machine code, a sophisticated compiler is not available for the R language.

Due to the performance problems of the R language, there are already approaches trying to optimise the runtime of R programs. Two of the most popular approaches are the *RCC project* [7], based on a master thesis from the Rice University, and the *byte code compiler* developed by Luke Tierney [10].

The RCC project tries to translate a given R program into ANSI C. It uses the intermediate representation produced by the parser of the original R-interpreter, which is written in C. Since the project only transforms the R program into function calls to the original interpreter API, the performance speedups are fairly low. So, an optimized R program is just up to four times faster than the original program. The byte code compiler developed by Luke Tierney makes a slightly different approach. Here, byte code is generated for a stack-based virtual machine with a simple set of instructions. The compiler itself is written in R and part of the latest R distribution. The resulting byte code needs also to be interpreted and the compiler supports only a few optimizations. Thus, there is quite room for improvement.

In addition, there are approaches which try to reengineer the original R-interpreter by implementing it in different languages. The *CXXR* project [9] refactors the R-interpreter into C++ to support the development of experimental interpreter versions. There is also a project called *Renjin* [1], representing a JVM-based R-interpreter written in Java.

To obtain efficient code, R developers tend to manually export computation-heavy parts to faster running languages like ANSI C. Therefore, the R programming language supports special interfaces. This strategy requires not only knowledge of an additional programming language, it is also not applicable in every case, especially for complex algorithms which analyse and evaluate genomic data.

The above mentioned approaches do not represent complete solutions, because only a few parts of R programs are optimized and thus, the achieved runtime improvements are still unsatisfying. A possible solution for this problem is the development of a compiler toolchain as employed for imperative languages, such as ANSI C or Java. To achieve this goal, it is necessary to examine different compiler strategies.

One possible compiler strategy is to translate R into ANSI C. This *C-based approach* is proposed in [8]. The idea is, to divide the compiler toolchain into four phases. In the first phase source level optimization are executed on the intermediate representation produced by the R-parser. Phase two transforms the optimized R code to C code. In the third phase the generated C code can be optimized by applying C source level optimizations. In the last phase, the optimized C code is translated into machine code by a standard compiler.

The main challenge in this approach is the translation from R to C code. R has dynamic features, which are hard to translate into an imperative language like C. Especially, the dynamic type system as well as dynamically growing arrays and the garbage collection are quite hard obstacles to deal with. A dynamic type system allows a fast development without the need of declaring types for variables. Thus, types of variables may change at runtime. In contrast, C has a static type system where types of variables are known at compile time. In addition, R uses arrays which can be dynamically adjusted in size at runtime. In C this feature is not natively supported.

Another challenge is R's garbage collection, which is needed to manage the heap space. To translate this feature in C, complex data flow analyses are required. Due to the mentioned dynamic features of R it is difficult to perform an efficient and full translation from R to ANSI C.

In terms of runtime optimizations for languages with dynamic features, like MATLAB or Python, *VM-based* approaches including *Just-In-Time*-based (JIT) solutions have been applied [4]. Such approaches concentrate more on the dynamic features of a language and could be very promising for the optimization of R runtimes. One of the most famous virtual machine is the *Java Virtual Machine* (JVM). It offers sophisticated optimizations and includes an advanced JIT-compiler. Therefore, a compiler strategy based on the JVM could support a better solution for the optimization of R programs compared to the C-based approach.

There are several possibilities to run R on the JVM. One possibility is to reengineer the R-interpreter in the Java programming language, like Renjin [1] does. However this approach does not lead to optimized runtimes, because the R language still needs to be interpreted. A more promising approach is to directly generate Java byte code from an R program. There are already other languages, which are very similar to R, targeting the JVM by using this approach. One of those languages is Scheme. With the *kawa language framework* [3] it is possible to compile Scheme into Java byte code. Kawa is a framework for dynamic languages and supports the compilation of dynamic languages into Java byte code. Thus, kawa could be a possible approach for translating R into Java byte code.

A compiler strategy which targets the JVM has several advantages. First of all, the new JVM 7 has now support for dynamically typed languages. A key part of this support is a new byte code called *invokedynamic*, which can be used to link a dynamic function call to a fast static function call. In addition to this the included JIT-compiler makes it possible to translate hotspots directly into machine code. Further, the JVM supports sophisticated optimizations. Especially the array bounds-check optimization could be very helpful, since R uses arrays which can be dynamically adjusted in size at runtime. Compared to the C-based compiler strategy, proposed in [8], the JVM-based approach has not only a better support for dynamic features, it has also a sophisticated garbage collection, which is capable of running parallel.

In order to demonstrate the advantages of translating R into Java byte code and thus, targeting the JVM, we exemplarily translated parts of R code into Java code. For this purpose we evaluated the same hotspot as in [8] of the R package *rda* for *Regularized Discriminant Analysis*. In [8] a single for-loop of rda's `apply()` function was manually translated into C and compiled with the GCC compiler. This leads to a speedup by a factor of 90. Translating the same part in Java, the speedup of the rda's `apply()` function is quite similar. We also compared runtimes of different benchmarks from [6] written in Java and C. The result shows, that Java is on average 1.7 times slower compared to the equivalent C program. Even if Java programs are not as fast as C programs, the JVM-based compiler strategy supports the dynamic features of R. This should lead to a higher optimization potential for runtimes of R programs compared to a C-based strategy.

References

[1] Bertram, A.: Renjin: JVM-based Interpreter for the R Language for Statistical Computing. `http://code.google.com/p/renjin`, 2011.

[2] Bioconductor Core Team: `http://www.bioconductor.org`, 2011.

[3] Bothner, P.: Kawa: compiling dynamic languages to the Java VM In *Proceedings of the annual conference on USENIX Annual Technical Conference*, pp. 41-41, 1998.

[4] Chevalier-Boisvert, M., Hendren, L. and Verbrugge, C.: Optimizing Matlab through Just-In-Time Specialization. In *Proceedings of the 19th International Conference on Compiler Construction*, pp. 46-65, 2010.

[5] Department of Statistics and Mathematics, WU Wien: The Comprehensive R Network. `http://www.cran.r-project.org`, 2011.

[6] Fulgham B.: The Computer Language Benchmarks Game `http://shootout.` alioth.debian.org, 2011.

[7] Garvin, J.: RCC: A Compiler for the R Language for Statistical Computing. Rice University, http://scholarship.rice.edu/handle/1911/17678, 2004.

[8] Plazar, S., Marwedel, P. and Rahnenführer, J.: Optimizing Execution Runtimes of R Programs. In *Book of Abstracts of International Symposium on Business and Industrial Statistics*, pp. 81-82, 2010.

[9] Runnalls, A.: CXXR: Refactorising R into C++. `http://www.cs.kent.ac.uk /projects/cxxr`, 2011.

[10] Tierney, L.: Compiling R: A Preliminary Report. In *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, 2001.

Subproject A4
Resource efficient and distributed platforms for integrative data analysis

Peter Marwedel   Olaf Spinczyk   Christian Wietfeld

Measurement methods and prediction of resource consumption

Matthias Meier
Fakultät für Informatik, Lehrstuhl 12
Technische Universität Dortmund
matthias2.meier@tu-dortmund.de

The most limited resource on a portable embedded device is usually the power supply. A number of solutions and methods are already explored to reduce the energy consumption. For example voltage/frequency scaling or processor sleep modes. I am working on energy models for low-power embedded devices to reduce its power consumption. However, to create such an energy model from scratch it is necessary to measure the power consumption of the device in detail by the use of accurate and reliable measurment methods.

A further key aspect of the SFB 876 project A4 are resource models. In order to examine resource models for configurable multiprocessor systems, I used our LavA [2, 3] framework, which facilitates the development of application-specific multiprocessor systems-on-chip (MPSoCs). LavA hides low-level hardware details by the use of a meta-model developed with the Eclipse Modeling Framework [1]. The output of the LavA framework is a tailored MPSoC written in VHDL[1].

1 Measurement methods for low-power devices

The first work package of the project A4 deals with the study of measurement methods. An elementary part of this package is the development of a suitable method to measure the power consumption of an embedded device. The common way to measure the power consumption of a device is the use of a shunt[2]. The shunt is placed in series with the

[1] Very High Speed Integrated Circuit Hardware Description Language

[2] Low-ohmic and precise resistor

power supply and the device under test. In order to determine the flowing current it is necessary to measure the voltage drop across the shunt. Due to the proportional relationship between the voltage drop and the flowing current, the power consumption of the device under test can easily be calculated by the use of Ohm's Law.

For our first steps we have decided to use the Texas Instruments eZ430-Chronos [4] watch, which is a low-power embedded sensor node. This device consumes only 10µA when the MSP430 CPU of the watch is in sleep mode and all peripheral devices are turned off. When the CPU is active and the radio is turned on the watch needs roughly 20mA. However, the accurate measurement of low currents in the range from 10µA to 20mA is very challenging. Especially when it is necessary to log the power consumption of a device over a period of time with the full dynamic current range from 10µA to 20mA. To measure the full range, it is necessary to change the shunt during the measurement, since otherwise the voltage drop across the shunt is to high and causes a voltage collapse, leading to a reset of the watch. Commercial measurement equipment is available, but not able to capture current changes with high resolution accurately. In order to avoid these problems I placed a Schottky diode parallel to the shunt into our measuring arrangement. This reduces the voltage drop across the shunt to a maximum of 0.3V and ensures that the watch does not a restart within the complete potential measurement range. Furthermore, this results in an exponential characteristic curve with a fine resolution in the range from 1µA to 2mA and a rather coarse resolution in the range from 2mA to 20mA. For measuring the voltage drop I use a digital storage oscilloscope that records the results with 4GSa/s.

Another problem is that electrical interferences from the 230V/50Hz alternating current mains have a negative influence on the measurement, even with the use of a high-quality laboratory power supply. To circumvent this problem, I decide to use a battery power supply in combination with a voltage regulator.

2 Prediction of resource consumption for MPSoCs

In parallel I have continued to work on the prediction of resource consumption for configurable MPSoCs. The basic resource of an FPGA[3] is a look-up table (LUT). It is basically used to implement almost any Boolean function of the designed hardware. Another limited resource of an FPGA is the memory which is provided by Block RAMs. For an efficient calculation of the resource consumption on an FPGA, resource models are defined. However, these models strongly depend on the FPGA vendor and the FPGA itself, and are therefore only useable for a small family of similar FPGA types. Our models support two FPGA families from Xilinx, which are used very frequently—the Spartan-3E and Virtex-5 series. The Spartan-3E family is a low-cost series with 4-input

[3] Field Programmable Gate Array

LUTs, whereas the Virtex-5 series offers high-end FPGAs with more resources and a 6-input LUTs design. Furthermore, it should be taken into account that the required resources of the specified hardware design depend on the used synthesis software and its configuration. For the resource models the Xilinx ISE Design Suite 10.1 with default settings is used.

Due to the vast number of potential hardware configurations it is only possible to measure a selection of multiprocessor systems to identify trends in the resource consumption. For the resource models presented in this technical report, I focus on the LUT consumption and the occupied Block RAMs. The required measurement results can easily be extracted from the ISE Design Suite. It provides a hierarchical, itemized list of the resource consumption for each hardware component in the design. The resource estimations for the MPSoC and SoC components are the most difficult of all, because they implement virtually no logic, but act as interconnects for all other components, such as processors or peripherals. However, to take these resources into account for our estimation, only a simple approximation for the MPSoC and SoC components is used.

The functions below exemplarily show the calculations of the LUTs for the IPC[4], CAN[5] and UART peripheral devices. These cost functions of the hardware components are integrated into our configuration process for a fast calculation of the resources at configuration-time. Like the other attributes of the MPSoC, used to configure the components of the multiprocessor system, also the cost functions are annotated to each device in the meta-model. The calculation of the total costs for each resource is realized by the *Xtend* language of the Eclipse Modeling Framework. The cost calculation is also embedded into the plausibility check and reports an error in the case of a mismatch between the chosen FPGA and the calculated resources. This enables the developer at a very early stage of the design process to change the multiprocessor system or to replace the FPGA with a more suitable one without the use of time-consuming synthesis software.

$$LUT4_{IPC} = (\frac{Connection.Width}{8} \times 90) + 230$$

$$LUT4_{CAN} = 933 + (47 \times Filters) + (90 \times Buffer)$$

$$LUT4_{UART} = \begin{cases} 82 & , Baud \leq 38400 \\ 68 & , Baud > 38400 \end{cases}$$

[4] Device for inter-processor communication

[5] Controller-area network device (basically used in the automotive sector to connect control units)

Processor	Peripherals	Prediction	Synthesis	Difference
2 MB-Lite with 8KB BRAM and Barrel Shifter	1 UART 57600 Baud 1 CAN with Filter 2 IPC Ctrl. 32 Bit	7843	7793	+0.69 %
4 MB-Lite with 8KB BRAM and Barrel Shifter	4 CAN with Filter 4 IPC Ctrl. 32 Bit	17,496	17,577	-0.46 %
15 ZPU with 8KB BRAM	9 UART 57600 Baud 21 IPC Ctrl. 32 Bit	23,886	24,755	-3.51 %

Table 1: Ressource estimation for Spartan-3E 1600 (4-input LUTs)

The key question is how accurate such resource models can predict the resource consumption for MPSoCs on an FPGA. Table 1 shows the predicted and the real LUT consumption for three configured multiprocessor systems on a Spartan-3E 1600 FPGA. This FPGA offers 29,504 LUTs in total for the hardware design. The difference between the estimation and the real consumption is quiet low with a maximum of 3.51 percent. Even with a LUT usage of 81 percent for the third system with 15 ZPUs, the LUT prediction is still close to the results of the synthesis software. Only when the LUTs are almost exhausted, the prediction of the LUT consumption differs from the synthesis results due to aggressive optimizations performed by the synthesis software. The calculation of the memory usage is quiet simple and the predicted results correspond accurately to the synthesis results.

References

[1] Eclipse Modeling Project.
`http://www.eclipse.org/modeling`.

[2] Matthias Meier, David Austin, Horst Schirmeier, and Olaf Spinczyk. TMPL: A hardware transactional memory product line. In *Proceedings of the Workshop on Multiprocessor Systems on (Programmable) Chips (MPSoC 2011)*, pages 539–546, Istanbul, Turkey, July 2011. IEEE Computer Society Press.

[3] Matthias Meier, Michael Engel, Matthias Steinkamp, and Olaf Spinczyk. LavA: An open platform for rapid prototyping of MPSoCs. In *Proceedings of the 20th International Conference on Field Programmable Logic and Applications (FPL '10)*, pages 452–457, Milano, Italy, 2010. IEEE Computer Society Press.

[4] Texas Instruments. eZ430-Chronos.
`http://www.ti.com/chronos`.

Energy-Efficient OFDMA Systems

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

Due to continuously increasing data rates of Orthogonal Frequency Division Multiple Access (OFDMA) based systems, like Long Term Evolution (LTE), power consumption becomes a key challenge for today's mobile devices. Based on the lack of dynamic power consumption models, we present results of an analytical model capable to capture the power consumption of different components within smart phones. Moreover, we introduce an approach for OFDMA based systems, which can reduce the interference between LTE femto- and macrocells by random frequency hopping. A reduced level of interference results in lower power consumptions under the constraint of a constant Quality of Service (QoS) requirement.

1 Power Consumption of Smart Phones

In order to analyze power consumptions of smart phones, numerical results of a semi-empirical model, which was published in [1], are presented in this section. As the model divides the overall power consumption into six individual parts, the results enable to identify which components consume the most power and hence which are the preferable ones to optimize.

According to [1], the power consumption of smart phones can be captured by an **analytical state dependent model**, where each state is linked to a specific power

$$P = (\beta_{uh} f_h + \beta_{ul} f_l)\,\mathrm{u} + \beta_{CPU}\mathrm{CPU}_{on} + \beta_{br}\mathrm{br} + \beta_{Gon}\mathrm{GPS}_{on} + \beta_{Gsl}\mathrm{GPS}_{sl} + \beta_{WiFi_l}\mathrm{WiFi}_l + \beta_{WiFi_h}\mathrm{WiFi}_h + \beta_{Audio}\mathrm{Audio} + \beta_{3G_{idle}}\mathrm{3G}_{idle} + \beta_{3G_{FACH}}\mathrm{3G}_{FACH} + \beta_{3G_{DCH}}\mathrm{3G}_{DCH}\;. \tag{1}$$

The power consumption of the CPU is modeled by the first two terms, the consumption of the LCD display by the third term, the consumption of the GPS by the fourth and fifth term, the consumption of the WiFi module by the six and seventh term, the consumption of the audio module by the eighth term and the cellular power consumption by the rest of the terms. Numerical values and clarifications for all parameters of this model can be found in [1] and are matched to an *HTC Dream*.

In order to analyze **numerical results**, we have implemented the model in (1) on an *HTC Incredible S*. Fig. 1 shows how the power is distributed among the different components, if all components are active and a maximum brightness of the display is used. According to [1], the CPU works at 385 MHz and 246 MHz, WiFi uses 802.11g, GPS applies the *LocationUpdate* method and cellular communication utilizes T-Mobile UMTS 3G network. It can be seen that more than 50% of the overall power is consumed for the RF components, i.e. WiFi, 3G and GPS. Hence, it is very important to reduce the power consumption of the communication modules.

A detailed view of the cellular and WiFi power is depicted in Fig. 2. The power consumptions depends on different states of the components. The cellular power is divided into the *off* state, the *idle* state where the interface only receives messages, the *Cell FACH* state where random/forward access common channels are active and the *Cell DCH* state where a dedicated channel for communication is set up. In an analogous way, the WiFi power is partitioned into the *off* state, the *low* state, and the *high* state. The transition from *low* to *high* state and vice versa is given by a hysteresis with 15 and 8 transmitted or received packets per second respectively.

Future work will include power measurements of other mobile devices in order to find new sets of parameters, as the validity of the model in (1) is limited to an *HTC Dream*.

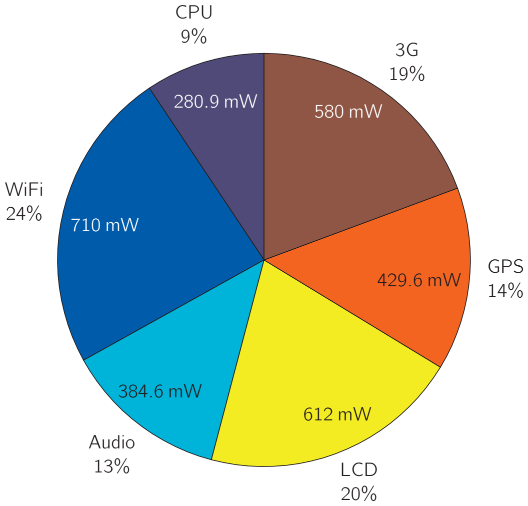


Figure 1: Distribution of overall power

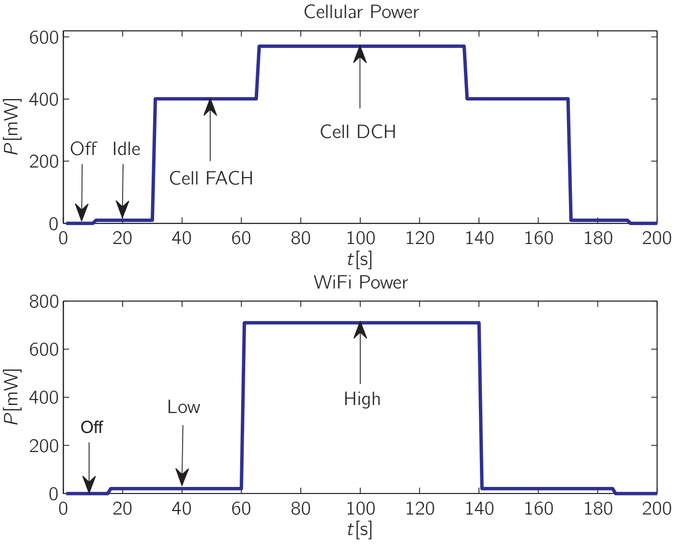


Figure 2: Power of 3G and WiFi

2 Reducing Power Consumption by Random Frequency Hopping

As wireless traffic of cellular and WiFi networks is constantly growing, operators are introducing femtocells to reduce excessive traffic in macrocells. In this way, capacity bottlenecks can be compensated by providing short-range hot-spots. As a result, higher signal qualities, higher throughputs and lower power consumptions can be achieved. Due to the fact that femtocells operate in the same bands as macrocells, interference between them occurs.

In classical resource planning of OFDMA systems, user subbands are changed according to deterministic time-frequency hopping patterns which are orthogonal to each other. Applying such centralized planning to femtocells is not feasible, as it would require a high administrative overhead. Therefore, we propose to replace centralized predetermined patterns by dynamical hopping patterns which are chosen randomly by the femtocell users [2]. In this way, the femtocell is able to integrate itself into the frequency bands of the macrocell with limited interference. A reduced level of interference results in a higher QoS level. Conversely, this means a reduced power consumption while preserving the same QoS requirements.

In order to evaluate the random frequency hopping gain compared to systems without resource planning, we determine an **analytical model** for the SINR (Signal-to-Interference-and-Noise Ratio) as well as the BER (Bit Error Rate). Beginning with an expression for an OFDMA signal in the time domain [3], we can determine an exact equation for the power of the interference signal after the baseband processing in the receiver and hence an analytical relationship for the SINR

$$\mathrm{SINR} = \frac{1}{\frac{M(r_u/r_i)^{\gamma}}{D_n D_m N_m} \sum\limits_{s=0}^{N_m-1} \sum\limits_{l=-\infty}^{\infty} \sum\limits_{d=0}^{N_n-1} E\left\{\mathrm{Re}^2\left\{\beta\right\}\right\} + \frac{1}{\mathrm{SNR}}} \,. \tag{2}$$

Here r_u denotes the distance between the access point and the analyzed system, r_i the distance to the interferer and M the number of considered interferers. All parameters can be found in [2]. Using inverse Laplace transform techniques, the BER per subcarrier results as

$$\mathrm{BER}_s = \frac{1}{2\pi \mathrm{j}} \int\limits_{c-\mathrm{j}\infty}^{c+\mathrm{j}\infty} \mathrm{e}^{\sigma_n^2 z^2/2} E\left\{\left[\int\limits_{-\infty}^{\infty} \prod_{l=-\infty}^{\infty} \prod_{d=0}^{N_n-1} \cosh\left(zA_i\sqrt{Q_n Q_m}\mathrm{Re}(\beta)\right) f(f_x)\mathrm{d}f_x\right]^M\right\} \times \frac{E\left\{\mathrm{e}^{-zA_u}\right\}}{z}\mathrm{d}z \,,$$

which can be solved numerically by Gauss-Chebychev quadrature. Herein $f(f_x)$ is the probability density function of the random frequency hopping, A_u the fading amplitude of the analyzed system and A_i the fading amplitude of the interference signal.

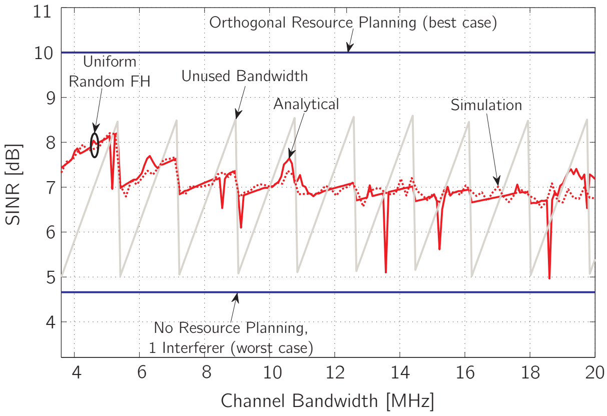


Figure 3: SINR vs. Hopping Bandwidth

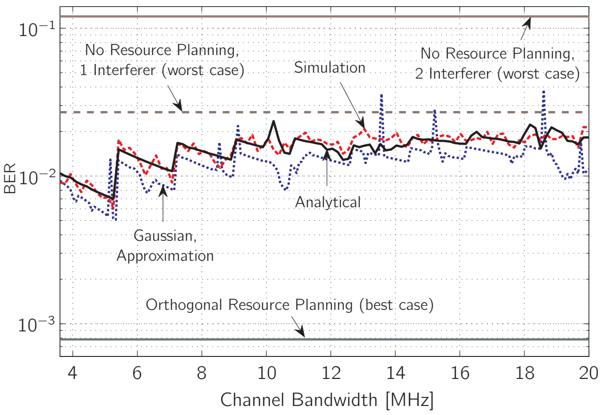


Figure 4: BER vs. Hopping Bandwidth

Finally, we present **results for the SINR and BER** of the proposed models. Assuming typical values of LTE and an allocation of 10 resource blocks, we obtain Fig. 3 for the SINR and Fig. 4 for the BER as a function of the channel bandwidth. Uniform random frequency hopping is used, where the carrier frequencies are chosen with equal probability from the channel bandwidth. In order to quantify the achieved gain, the results are compared to the performance of systems with centralized and orthogonal resource planning and to systems without resource planning and random hopping. Both the SINR and BER show a characteristic according to the unused bandwidth, which is the difference between the channel bandwidth and occupied bandwidth of all users in the cells. The higher the unused bandwidth, the higher the SINR and the lower the BER. All results are verified by simulations implemented in MATLAB. Moreover, an approximation can be made for the BER, if we assume that the interference signal follows a Gaussian distribution, cf. Fig. 4.

Future work will include the analysis for more complex scenarios with frequency-selective channel models and the impact of different service types for the users.

References

[1] L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R. Dick, Z. Mao and L. Yang. Accurate Online Power Estiamation and Automatic Battery Behaviour Based Power Model Generation for Smartphones. *IEEE/ACM/IFIP CODES+ISSS, Scottsdale, Arizona*, Oct. 2010.

[2] M. Putzke and C. Wietfeld. Self-Organizing OFDMA Systems by Random Frequency Hopping. *Accepted for 4th. IFIP Wireless Days, Niagara Falls, Ontario*, Oct. 2011.

[3] C. Snow, L. Lampe and R. Schober. Impact of WiMAX Interference on MB-OFDM UWB Systems: Analysis and Mitigation. *IEEE Transactions on communications, Vol. 57, No. 9*, Sep. 2009.

Modeling the Energy Efficiency of 4G Wireless Transmission

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

In this report a measurement based energy model for IEEE 802.16e conform mobile WiMAX devices is presented. For this purpose, extensive measurements of the energy consumption of a mobile WiMAX device have been performed for different system parameterizations and different data sizes. From the results of the measurements, analytical models have been derived, which allow for the calculation of the energy that has to be spent for successfully transmitted bit.

1 Motivation

The operational time for which a mobile communication device can be used prior it has to be recharged is one of the most important performance parameter for the consumers of new smart phones [1]. This is the reason why energy efficiency is one of the most important design targets if novel smart phones based on LTE or mobile WiMAX have to be developed. However, before extensive simulations on the energy consumption of new systems, protocols or algorithms can be performed, detailed energy models are mandatory as a basis for the performance evaluation. One of the most critical performance indicators in this context is the energy that has to be spent for the successful transmission of 1 bit. In the following a measurement based, analytical energy model addressing this emerging topic is presented.

2 Measurement Setup

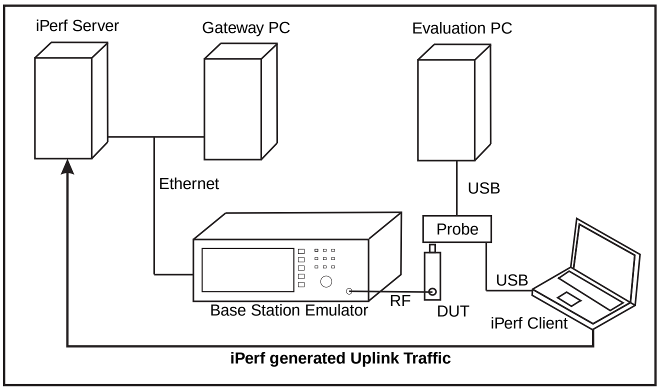
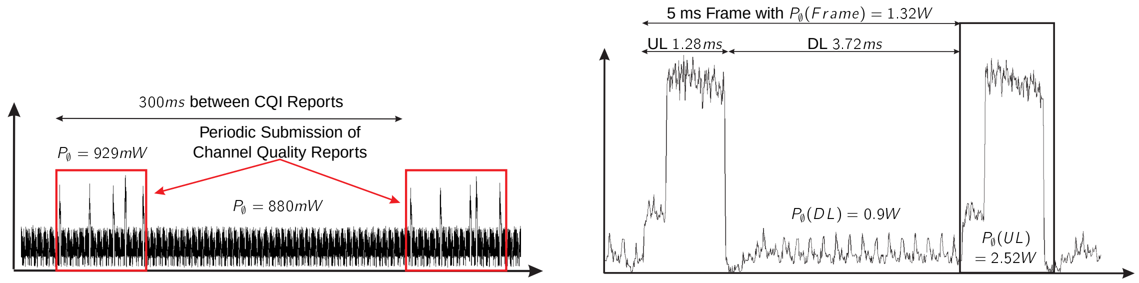


Figure 1: Measurement Setup for Detailed Power Measurements

The measurement setup used for the determination of the USB sticks energy consumption can be seen in Fig. 1. The actual mobile WiMAX link between the Base Station Emulator (BSE) and the Device Under Test (DUT) is realized my means of an RF cable to avoid external influences on the radio link. The BSE acts as a base station and creates a standard conform radio cell towards the DUT. In the context of energy consumption measurements, the main role of the BSE is forwarding of the uplink data to the server as well as the submission of power control MAC messages in the downlink. The iPerf server is connected to the BSE via Ethernet. Beside this, the BSE used for the investigations makes use of an additional gateway PC for end-to-end application testing. The actual power measurement is performed by a measurement probe that is placed between the DUT and the client PC. Therefore the energy consumed by the DUT has to pass the probe, where it is measured in terms of electric current and voltage. From this, the current power consumption is calculated at a frequency of 100 kHz and transmitted to an evaluation PC via USB. Here the data is stored for a detailed evaluation in terms of for example the determination of the consumed energy in a certain period of time. From this it is for example possible to very precisely measure the energy that is consumed for the transmission of single symbols, bursts and frames as it can be seen in Fig. 2.

3 Evaluation of Measurements and Derivation of a Power Consumption Model

A general overview on the course of the current power over time can be seen in Fig. 2. Fig. 2(a) illustrates the power consumption over time for a Mobile WiMAX device that is connected to a base station, but does not transmit or receive any data. Therefore a continuous average power consumption of 880 mW is needed, while every 300 ms a channel measurement report is submitted in the uplink. For this transmission period, the average power increases to 929 mW. Fig. 2(b) shows a zoom to the power measurements for one TDD (Time Division Duplex) frame. One can clearly see the different parts of the



(a) Idle State with Channel Quality Transmission

(b) Power Distribution for WiMAX Frame (UL/DL = 35/12)

Figure 2: Mobile WiMAX Power Consumption for Different System States

burst for the uplink and the downlink where the ratio corresponds to the defined TDD downlink to uplink ratio of 35/12. While the overall average power for the submission of one burst is 1.32 W only 0.9 W are consumed in during the reception phase, while 2.52 W are needed for the transmission of data in the uplink (assuming an uplink Tx power of -6 dBm as reported to the BS).

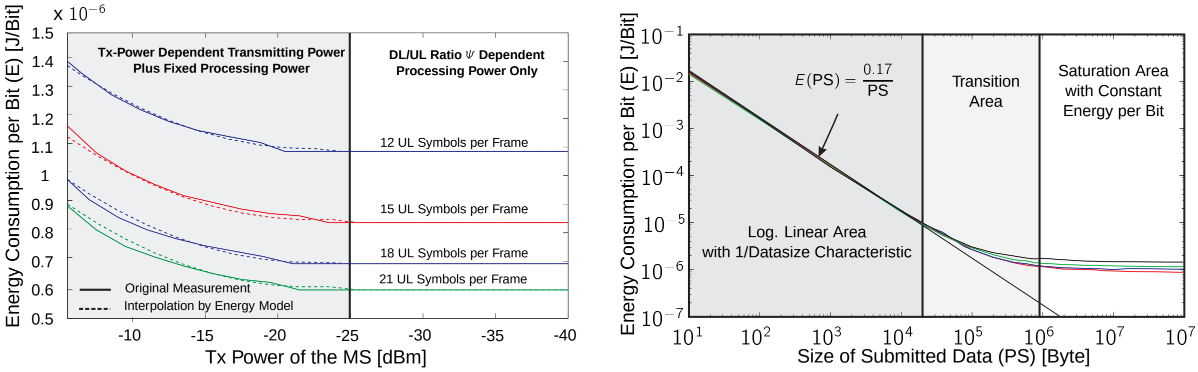
In the following section the dependencies of the energy efficiency from different system parameters Tx-Power and packet size are derived by means of measurements and analytical fitting: The influence of the uplink Tx-Power on the energy needed for the successful submission of one bit was determined by transmitting an iPerf generated data stream at the maximum possible data rate for the different DL to UL ratios Ψ. Afterwards the consumed energy for the transmission of one bit was calculated by

$$\frac{E}{bit} = \frac{\frac{1}{T} \cdot \int_{t=0}^{T} P(t)dt}{\frac{1}{T} \cdot \int_{t=0}^{T} DR(t)dt} \tag{1}$$

for a defined measurement period of 30 seconds. The results of the measurements can be seen from the solid lines in Fig. 3(a). The determined figures show that the overall power consumption can be divided in a fixed processing power that does only depend on the DL to UL ratio Ψ and a Tx-Power dependent component that is independent of Ψ and added to the processing power. Polynomial fitting has been performed in a least square sense to find a polynomial that models the non-linear increase of the Tx-Power dependent component. From that we found that a second order polynomial is a very good approximation (see Fig. 3(a)).

For the determination of the file size dependent energy efficiency, files with different sizes have been transfered, while the Tx power is constant. In contrast to the previously described model for this investigation, the energy consumption per bit was calculated as

$$\frac{E}{bit} = \frac{\int_{t=0}^{T} P(t)dt}{PS} \tag{2}$$



(a) Tx-Power Dependent Energy Efficiency

(b) Energy Efficiency for Different File Sizes @ -6 dBm Tx Power

Figure 3: Original Measurement vs. Analytical Energy Model

where T determines the time that is needed to transmit a file of size PS. The results plotted in Fig. 3(b) show that the energy consumption per bit has to be modeled for three different areas. For a packet size of up to 20 kByte the shape is linear in a double logarithmic plot which again leads to a $1/x$ relationship. For packet sizes above 900 kByte the energy consumption per bit is constant and can be calculated based on the UL/DL ratio Ψ and the Tx-Power. For the transition area between 20 kByte and 900 kByte either the analytical expression for lower packet sizes or higher packet sizes can be applied, but increased errors have to be expected. The overall energy model for the packet size dependent energy consumption can be analytically expresses by the formula shown in Fig. 3(b).

References

[1] Lance Bloom, Rachel Eardley, Erik Geelhoed, Meera Manahan, and Parthasarathy Ranganathan. Investigating the Relationship Between Battery Life and User Acceptance of Dynamic, Energy-Aware Interfaces on Handhelds. *Mobile Human-Computer Interaction - MobileHCI*, 3160:43–79, 2004.

[2] Björn Dusza and Christian Wietfeld. Interference Robustness Measurements for IEEE 802.16e mobileWiMAX Systems. *Proceedings of The Seventh International Symposium on Wireless Communication Systems (ISWCS)*, September 2010.

[3] Björn Dusza and Christian Wietfeld. QDV - A QoS Enabled Packet Scheduling Scheme for Mobile WiMAX in UAV Swarms. *Proceedings of the International Conference on Computer Communication Networks (ICCCN) Workshop on Context-aware QoS Provisioning and Management for Emerging Networks, Applications and Services (ContextQoS)*, September 2011.

A Scalable Aspect-Oriented IP Stack for Resource-Constrained Devices

Christoph Borchert
Fakultät für Informatik, Lehrstuhl 12
Technische Universität Dortmund
{christoph.borchert}@tu-dortmund.de

Network protocol stacks are an important ingredient of today's system software. For example, all state-of-the-art operating systems for personal computers come with a TCP/IP stack. In the domain of resource-constrained devices, like wireless sensor networks, the Internet protocols are actually not extensively used due to their complexity. This work focuses on the design and implementation of a scalable TCP/IP stack for embedded devices with tight resource constraints. It discusses a methodology for analyzing variability of IP networking software and provides preliminary results in terms of memory footprint.

1 Introduction

Communication is essential for today's computer systems. Almost every recent personal computer comes with a built-in network interface, and devices like mobile phones have also access to data networks and interconnect to the global Internet. They employ successfully the TCP/IP protocol suite, which is the de facto standard for data communication. I attack the problem of bringing resource-constrained devices and TCP/IP together in order to achieve *interoperability* among embedded devices and common personal computers.

Dunkels [2] showed that even 8-Bit microcontroller architectures are suited for running a fully standard compliant TCP/IP stack despite of tight memory limits. This led to the development of *uIP* ("micro IP") and *lwIP* ("lightweight IP"), which are both open source implementations using the C programming language. uIP reduces the feature set of TCP/IP to an absolute minimum required for standard conformance, whereas lwIP implements most common features of TCP/IP and thus provides much greater functionality. Compared to the fixed minimal feature set of uIP, lwIP offers configurability of several features by means of the C preprocessor. A feature is for instance a particular protocol, like TCP, that is conditionally compiled by surrounding `#ifdef` directives. lwIP's source code is characterized by the excessive use of C preprocessor directives. Expressing software variability in that way is heavily criticized as error-prone, unreadable and unmaintainable [4]. There are mainly two drawbacks of textual preprocessing:

1. Dependencies of features are represented in the source code.
2. Features are not modularized, especially if they crosscut the implementation (many functions, files, and so on).

The first drawback applies for source code that is shared by more than one feature. The corresponding source code has to be encapsulated by `#ifdefs` for all particular features. The second drawback is valid for crosscutting concerns, which are features which affect many scattered lines of code. Crosscutting concerns cannot be implemented by the use of the C preprocessor in a modular way, because the resulting implementation is scattered, tangled, hard to understand and hard to maintain.

The following section outlines a methodology for the design and implementation of variable IP networking software that avoids the problems discussed above.

2 Design Approach

IP networking software for resource-constrained devices has to be statically configurable, like lwIP. Configurable software allows the selection and removal of features, and thus the adaptation to different devices and requirements. Protocol features that are not needed can be omitted in order to reduce memory consumption.

The development of configurable software has to focus on variability, which can be formalized by feature modeling [3]. The idea is to identify and document requirements in terms of *features*, which can be either mandatory or optional. A feature model encompasses all requirements that are imposed for the product being developed. The success of fine-grained configurable IP networking software relies mainly on high-quality feature models: It is crucial to have detailed information about the expected variability for the subsequent design and implementation process.

As a starting point for a feature model of the TCP/IP protocol suite, I use the official specification published by the Internet Engineering Task Force (IETF). In RFC1122 [1], the most important requirements are summed up and categorized into three different kinds: *MUST*, *SHOULD*, and *MAY*. These capitalized words determine the significance of each particular requirement.

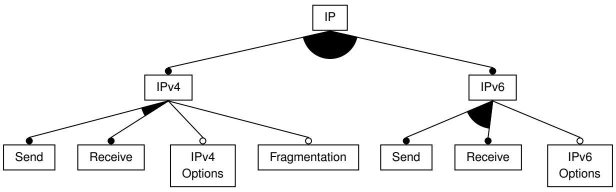
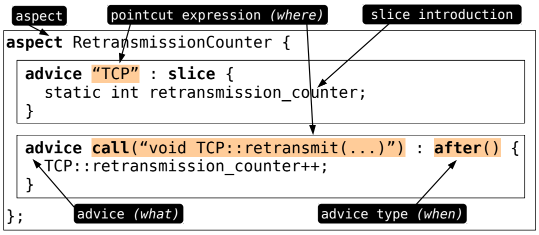


Figure 1: Simplified Feature Diagram of the Internet Protocol



```
aspect RetransmissionCounter {
  advice "TCP" : slice {
    static int retransmission_counter;
  }
  advice call("void TCP::retransmit(...)") : after() {
    TCP::retransmission_counter++;
  }
};
```

Figure 2: Syntax of AspectC++

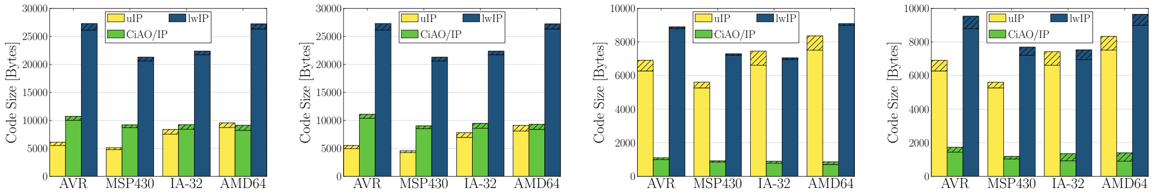
Figure 1 shows a simplified feature diagram of IP. A feature diagram graphically points out variability by using a tree layout. Each node represents a feature, which depends on its ancestor node. A filled circle at the lower end of an edge indicates a mandatory feature (c.f. RFC *MUST*), whereas a non-filled circle describes an optional one (c.f. RFC *SHOULD* or *MAY*). Cumulative features, of which a least one must be present, are grouped together by a filled arch at the upper ends of their edges.

Having analyzed the variability of each protocol of the TCP/IP suite, I developed a TCP/IP stack from scratch using *aspect-oriented programming (AOP)* [5]. The goal was to transfer the variability to the implementation level in a modular way in order to avoid the already discussed issues of the C preprocessor. The key concept behind AOP is *implicit invocation*: an *aspect* contains *advices*, which intercept the program's control flow and extend the underlying types. Advices target several *join points*, which are locations in the dynamic control flow or part of the static program structure. Join points are described via *pointcut expressions* in a textual form.

AspectC++ [5] extends the C++ programming language by AOP facilities. It consists of an *aspect weaver*, that compiles aspects into ordinary C++ code, which is *woven* into existing C++ source code files at relevant locations (i.e. join points). Therefore, advice code is *inlined* and produces no overhead compared to an implementation by hand (see [4]). Figure 2 outlines the syntax of AspectC++ by taking the example of a retransmission counter for TCP. The given aspect encapsulates two advices. First, the class `TCP` is extended by an integer member variable to store the counter's value. The static program structure is modified by a *slice introduction*. The second advice affects the function `TCP::retransmit(...)` and implicitly increases the counter *after* each function *call*. The ellipsis of the function's arguments in the pointcut expression ensures that overloaded functions are matched regardless their arguments.

3 Evaluation

To evaluate the aspect-oriented TCP/IP stack, I use a wide range of hardware platforms that cover 8- up to 64-Bit architectures. Figure 3 summarizes my preliminary results. *CiAO/IP* refers to the aspect-oriented implementation that is compared to uIP and lwIP



(a) TCP Client (b) TCP Server (c) UDP Tx w/o Checksum (d) UDP Rx w/o Checksum

Figure 3: Code size (ROM) for common use cases. Subfigures (a) - (d) show the minimal memory consumption of each IP stack. The lower sections of the stacked bars constitute the code size of the IP stacks themselves. The upper hatched sections indicate the cost for the application code that has to be added for full operation.

for common use cases. Each use case includes the Internet Protocol itself, and on IA-32 and AMD64, additionally Ethernet and the Address Resolution Protocol (ARP), since typical IA-32 and AMD64 systems feature an Ethernet adapter. The results show that aspect-oriented IP networking software is competitive to common C implementations in terms of program memory consumption. For the future, I plan to further investigate performance and scalability issues of the discussed systems.

References

[1] R. Braden. Requirements for Internet Hosts - Communication Layers. RFC 1122 (Standard), October 1989.

[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] K. Kang, S. Cohen, J. Hess, W. Novak, and S. Peterson. Feature-oriented domain analysis (FODA) feasibility study. Technical report, Carnegie Mellon University, Software Engineering Institute, Pittsburgh, PA, November 1990.

[4] Daniel Lohmann, Fabian Scheler, Reinhard Tartler, Olaf Spinczyk, and Wolfgang Schröder-Preikschat. A quantitative analysis of aspects in the eCos kernel. In *EuroSys 2006 Conference (EuroSys '06)*, pages 191–204, New York, NY, USA, April 2006. ACM.

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

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

Gabriele Kern-Isberner Joachim Biskup

Answerset Programming under Confidentiality Requirements in Multiagent Systems

Patrick Krümpelmann
Chair 1 - Information Engineering Group
Technische Universität Dortmund
patrick.kruempelmann@tu-dortmund.de

This report presents an overview of past and current work of the author in the Collaborative Research Center SFB 876 as well as future plans. It lays out the three lines of research with respect to the use of answerset programming under confidentiality requirements in multiagent systems. It focuses on the definition of confidentiality using answerset programming and gives a summary of it. In the end current and ongoing work is sketched.

This is the report about work and plans in the Collaborative Research Center SFB 876 "Collaborative Research Center SFB 876 - Providing Information by Resource-Constrained Data Analysis", project A5 "Exchange and Fusion of Information under Availability and Confidentiality Requirements in MultiAgent Systems". The project aims at developing theories of confidentiality for multiagent systems using two different approaches to non-monotonic reasoning, answer set programming (ASP) [2] and ordinal conditional functions (OCF) [3] as an instantiation of the System P. OCF in combination with c-representations [3] provide a rich semantics which satisfy many desirable properties while ASP allow for intuitive knowledge representation and comes along with several powerful and fast solvers which have proven to be practically usable which makes ASP especially interesting for resource-constrained data analysis. The reasoning component will be embedded in a multiagent system that is based on the preliminary work published in [1]. Complex inference operators are rarely used in current implementations of multi agent systems and models of confidentiality based on these are lacking. The project therefore makes the development of new models and the combination of those with existing techniques necessary. The authors current work on the project is on three lines. Firstly work on a definition of secrecy from the point of view of an autonomous epistemic agent with incomplete and uncertain information which is situated in a multiagent system. The resulting framework from this work will form the basis for the incorporation of different non-monotonic logics, operators of belief dynamics and confidentiality preservation into an model for an autonomous agent. The results of this work will be submitted to a conference in October. Secondly formalizing confidentiality in the context of answer set programming of which a a summary is given below. Thirdly work on the design and implementation of a flexible and modular multiagent framework that allows the use of plugins for different logics for the knowledge representation and plugins for important operators of the agent which include operators for belief change and operators that determine the behavior of the agent with respect to confidentiality preservation. The plugin architecture allows us to easily define and compare different types of agents and evaluate their performance. In the following extended logic programs and the answer set semantics are introduced and a brief summary of the formalization of confidentiality based on these is given. Afterwards an outlook on future work in the project is presented.

An extended logic program consists of rules over a set of atoms $\mathcal{A}$ using strong negation $\neg$ and default negation not. A literal L can be an atom $A \in \mathcal{A}$ or a negated atom $\neg A$. The complement of a literal L is denoted by $\neg L$ and is A if $L = \neg A$ and $\neg A$ if $L = A$. Let $\mathcal{A}$ be the set of all atoms and $Lit_{\mathcal{A}}$ the set of all literals $Lit_{\mathcal{A}} = \mathcal{A} \cup \{\neg A \,|\, A \in \mathcal{A}\}$. A rule r is written as $L \leftarrow L_0, \ldots, L_m, \text{not } L_{m+1}, \ldots, \text{not } L_n$. where the head of the rule $L = H(r)$ is either empty or consists of a single literal and the body $\mathcal{B}(r) = \{L_0, \ldots, L_m, \text{not } L_{m+1}, \ldots, \text{not } L_n\}$. The body consists of a set of literals $\mathcal{B}(r)^+ = \{L_0, \ldots, L_m\}$ and a set of default negated literals denoted by $\mathcal{B}(r)^- = \{L_{m+1}, \ldots, L_n\}$. If $\mathcal{B}(r) = \emptyset$ r is called a fact. A set of literals that is consistent, i. e., it does not contain complementary literals L and $\neg L$, is called a state S. A literal L is true in S, denoted by $S \models L$, iff $L \in S$ and false otherwise. The body $\mathcal{B}(r)$ of a given rule r is true in S iff each $L \in \mathcal{B}(r)^+$ is true in S and each $L \in \mathcal{B}(r)^-$ is false in S. A rule r is true in S iff $H(r)$ is true in S whenever $\mathcal{B}(r)$ is true in S. A state is a model of a program P if r is true in S for all $r \in P$. The reduct P^S of a program P relative to a set S of literals is defined as $P^S = \{H(r) \leftarrow \mathcal{B}^+(r) \mid r \in P, \mathcal{B}^-(r) \cap S = \emptyset\}$. An answer set of a program P is a state S that is a minimal model of P^S. The set of all answersets of P is denoted by $\mathcal{S}(P)$. An extended logic program P infers a literal L credulously, denoted by $P \models^c_{asp} L$, iff $L \in \cup\mathcal{S}(P)$. An extended logic program P infers a literal L skeptically, denoted by $P \models^s_{asp} L$, iff $L \in \cap\mathcal{S}(P)$. $P \models^\circ_{asp} S$ refers to any answer set inference relation. The corresponding sets of consequences are denoted by $Cn_{asp}(P) = \{L \in Lit_{\mathcal{A}} \mid P \models^\circ_{asp} L\}$. The answer set semantics defines the evaluation of queries $?L$ with $L \in Lit_{\mathcal{A}}$ as *yes* if $P \models^\circ_{asp} L$, *no* if $P \models^\circ_{asp} \neg L$ and *unknown* else.

The multiagent framework presented in [1] is adopted and a possibility to extend it towards the use of answer set programming for confidentiality is presented in the following as a basis for further discussion and development.

Definition 1. The *epistemic state* $\mathtt{bel}^\sigma_A$ of an agent A after a sequence of actions

σ is a tuple $\texttt{bel}_A^\sigma = (\texttt{is}_A^\sigma, \texttt{conf}_A^\sigma, \{\texttt{view}_{A,A_1}^\sigma, \dots, \texttt{view}_{A,A_m}^\sigma\})$ with individual belief base $\texttt{is}_A^{sigma} \subseteq \mathcal{L}^\mathcal{B}$, a personalized confidentiality policy $\texttt{conf}_A^\sigma$, and views, i. e., the beliefs of agent A about the beliefs of agents A_i, $\texttt{view}_{A,A_1}^\sigma, \dots, \texttt{view}_{A,A_m}^\sigma \subseteq \mathcal{L}^\mathcal{B}$. The belief set, the set of all inferences based on a belief base, is given by an belief operator $Bel(\texttt{is}_A^\sigma)$.

For the sake of clarity often two specific agents, namely Ann A and Bob B, are picked to illustrate definitions and terms since each speech act can be seen as an action involving two agents. Normally, the view of Ann is taken who wants to keep secrets from Bob. The general framework presented previously is going to be instantiated using extended logic programs as the underlying logical language. To this end the epistemic state of agents in this scenario needs to be elaborated. The first two components of the beliefs of an agent are instantiated by extended logic programs. The third component of the beliefs, the confidentiality policy is defined on the language of answer sets and therefore the one of answers to queries.

Definition 2. A belief base of an Agent A consists of a logic program, $\texttt{is}_A^\sigma \in \mathcal{P}$. A view of agent A on agent B of the belief base $\texttt{is}_A^\sigma$ of A is defined as an extended logic program, $\texttt{view}_{A,B}^\sigma \in \mathcal{P}$. A *confidentiality policy* $\texttt{conf}_A^\sigma$ is a set of literals, $\texttt{conf}_A^\sigma \subseteq Lit_\mathcal{A}$.

Example 1. For example $\texttt{is}_A^\sigma = \{a \leftarrow b, \text{not } \neg c., b., \neg c.\}$, $\texttt{view}_{A,B}^\sigma = \{a \leftarrow b, \text{not } \neg c., b.\}$ and $\texttt{conf}_A^\sigma = \{a, \neg b\}$.

Note, that alternative definitions of confidentiality policies are feasible and could be set to a program or even propositional formulas. This is left for future work. According to the instantiation of the agent's beliefs the belief operator $Bel(\cdot)$ is implemented by means of the inference operator based on the answer set semantics, i. e. $Cn_{asp}^\circ(\cdot)$.

Definition 3. A *revision operator* $*$ is a mapping from two logic programs to one single program, $P * P' = P''$.

For more detail on revision of logic programs refer to [5]. The result of a revision request is dependent on the input to be revised by in relation to the knowledge base of the agent being requested. Given a knowledge base $\texttt{is}_A^\sigma \subseteq \mathcal{L}$ and a set of sentences $\phi \subseteq \mathcal{L}$ an *acceptance function* f is defined as $f : \mathsf{P}(\mathcal{L}) \times \mathsf{P}(\mathcal{L}) \to \mathsf{P}(\mathcal{L})$ such that $f(\Psi, \phi) \subseteq \phi$. The acceptance function needs to be implemented by each agent. For more elaborate acceptance functions refer to [6]. A view $\texttt{view}_{A,B}^\sigma$ and a knowledge base $\texttt{is}_A^\sigma$ are called *compatible*, if $Cn_{asp}^\circ(\texttt{view}_{A,B}^\sigma) \subseteq Cn_{asp}^\circ(\texttt{is}_A^\sigma)$. Given an example belief base $\texttt{is}_A^\sigma = \{a., b \leftarrow a, \text{not } c.\}$ and a view $\texttt{view}_{A,B}^\sigma = \{b.\}$, then $Cn_{asp}^\circ(\texttt{view}_{A,B}^\sigma) = \{b\} \subseteq \{a, b\} = Cn_{asp}^\circ(\texttt{is}_A^\sigma)$. After every received or sent communication act an agent needs to change its beliefs. Therefore change operators for all components that are dependent on the type of act performed need to be defined. For the belief base a corresponding operator, though abstractly, is already defined. In the following the operators for updating an agent's views on other agents are described. Given a request for the evaluation of L the update function is defined depending on the sent answer: *yes*: $\texttt{view}^{\sigma'}_{A,B} = \texttt{view}^{\sigma}_{A,B} * \{L\}$; *no*: $\texttt{view}^{\sigma'}_{A,B} = \texttt{view}^{\sigma}_{A,B} * \{\neg L\}$; *unknown*: $\texttt{view}^{\sigma'}_{A,B} = \texttt{view}^{\sigma}_{A,B} - \{L, \neg L\}$. This definition demands that given the answer is *yes* the update of a view shall make sure that L should be true. $\texttt{view}^{\sigma'}_{A,B} \models^{\circ}_{asp} L$ for any $\circ \in \{c, s\}$. Likewise, after an answer *no*, one gets $\texttt{view}^{\sigma'}_{A,B} \models^{\circ}_{asp} \neg L$ for any $\circ \in \{c, s\}$. If the answer is *unknown* a contraction by the set $\{L, \neg L\}$ is performed which leads to a state where $\texttt{view}^{\sigma'}_{A,B} \not\models^{\circ}_{asp} L$ and $\texttt{view}^{\sigma'}_{A,B} \not\models^{\circ}_{asp} \neg L$ for any $\circ \in \{c, s\}$. Given a revision request for ϕ and an answer ϕ' the update function is defined as: $\texttt{view}^{\sigma'}_{A,B} = \texttt{view}^{\sigma}_{A,B} * \phi'$. Hence the view of the attacker is updated by incorporating the accepted input. This general setup raises requirements to the change operators for the component of the epistemic state of the agent.

Current work in line with the ASP formalization is dedicated to formalizing the requirements laid out in the last sections and defining operators satisfying these based on the framework presented in [4]. In the next steps all three current lines of work will be joint in one system, brought together and evaluated with the work on the OCF representation in the Project. Extensions towards handling of preference and plausibility as well as the evaluation with respect to resource-constrains are planned.

References

[1] Joachim Biskup, Gabriele Kern-Isberner, and Matthias Thimm. Towards enforcement of confidentiality in agent interactions. In Maurice Pagnucco and Michael Thielscher, editors, *Proceedings of the 12th International Workshop on Non-Monotonic Reasoning (NMR'08)*, pages 104–112, Sydney, Australia, September 2008. University of New South Wales, Technical Report No. UNSW-CSE-TR-0819.

[2] Michael Gelfond and Nicola Leone. Logic programming and knowledge representation — the A-Prolog perspective. *Artificial Intelligence*, 138(1–2):3–38, 2002.

[3] Gabriele Kern-Isberner. *Conditionals in nonmonotonic reasoning and belief revision: considering conditionals as agents*. Springer-Verlag, Berlin, Heidelberg, 2001.

[4] Patrick Krümpelmann. Dependency semantics for sequences of extended logic programs. *Logic Journal of the IGPL*, To appear., 2011.

[5] Patrick Krümpelmann and Gabriele Kern-Isberner. On belief dynamics of dependancy relations for extended logic programs. In *Proceedings of the 13th International Workshop on Non-Monotonic Reasoning (NMR)*, 2010.

[6] Patrick Krümpelmann, Matthias Thimm, Marcelo A. Falappa, Alejandro J. Garcia, Gabriele Kern-Isberner, and Guillermo R. Simari. Selective revision by deductive argumentation. In *Proceedings of the First International Workshop on the Theory and Applications of Formal Argumentation (TAFA'11). Barcelona, Spain*, 2011.

On the Inference-Proofness of Materialized Views

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

In this report my research during the last year of investigating the inference-proofness of a specific existing approach to vertical fragmentation of relational database instances is outlined. Moreover, it is discussed briefly how inference control can be established efficiently based on these results by employing so-called inference-proof materialized views of relational database instances. Finally, the long-term objective of designing a comprehensive solution to the problem of inference-proof materialized views is sketched.

My research during the last year was motivated by the observation that information has become one of the most important resources, which has to be protected. In order to protect information from undesired disclosures, confidentiality requirements are declared by setting up a confidentiality policy. Moreover, there is an increasing need for storing data cost-efficiently in our economy-driven society. One approach to achieve this goal is called "database as a service" paradigm and leads to third party service providers specialized on hosting database systems and offering the use of these database systems to their customers via Internet in return for payment of rent [8].

Obviously, there is a goal conflict between the discussed "database as a service" paradigm and confidentiality requirements because the service provider cannot be restrained from reading all cleartext information stored in its systems. To solve this conflict, some authors suggest to break sensitive associations by splitting relational instances vertically, which is referred to as vertical fragmentation. There are several different approaches to achieving confidentiality based on vertical fragmentation surveyed in [10] and for each of these approaches the corresponding authors describe how fragments of an original relational instance can be outsourced so that unauthorised (direct) accesses to confidential information are prohibited. But it is *not* shown that confidential information cannot be inferred by employing inferences, which may offer the possibility to infer confidential information based on the knowledge of non-confidential information [7]. Moreover, it is *not* considered that an attacker often has some a priori knowledge, which might enable him to infer confidential information [3].

In my diploma thesis a specific approach to vertical fragmentation – splitting a relational instance into one externally stored part and one locally-held part – has been analysed w.r.t. its inference-proofness [9]. More specifically, based on the seminal ideas proposed in [5, 6], a formalisation of this approach to vertical fragmentation is developed and analysed w.r.t. its inference-proofness within the framework of so-called Controlled Query Evaluation (CQE). This framework comprises several different approaches, which are all known to be inference-proof provably by limiting a user's information gain so that this user cannot infer protected information reliably based on his a priori knowledge and the (possibly distorted) answers to his queries [2].

The main result of this diploma thesis is that the approach to vertical fragmentation considered *can* be inference-proof as long as a user only has some a priori knowledge in terms of a rather restricted class of functional dependencies. But in general this restricted class of functional dependencies should not be sufficient to represent a user's a priori knowledge about the semantic constraints being valid in the database instance under consideration suitably. So, during the last year I did research on finding more expressive classes of a priori knowledge under which the inference-proofness of fragmentation still holds provably.

The main novel contribution of this research is that the inference-proofness of the approach to fragmentation considered still holds if an attacker's a priori knowledge consists of arbitrary unirelational and typed semantic constraints as long as they belong to the rather general classes of so-called Equality Generating Dependencies (EGDs) or Tuple Generating Dependencies (TGDs), which together comprise nearly all semantic constraints (cf. [1]). This result will be published in [4] and presented at ISC 2011 conference in China in October 2011.

Intuitively expressed, an EGD claims that the presence of some tuples in a relational instance r implies that certain components of these tuples are equal and a TGD claims that the presence of some tuples in r implies the existence of certain other tuples in r. Moreover, a constraint is unirelational if it refers to only one relational schema, and it is typed if there is an assignment of variables to column positions preventing the claim for equality of values being in different columns of r [1]. A well known example for a unirelational and typed EGD is an arbitrary functional dependency and an example for a unirelational and typed TGD is a join dependency.

As there are other approaches to achieving confidentiality by vertical fragmentation than the one treated in [4] (see e.g. [10]), an idea for future work might be to analyse the inference-proofness of these approaches. As these approaches free the client from storing data locally by resorting to encryption if necessary, the logic-oriented modelling of an attacker's knowledge has to be adapted suitably to reflect these circumstances. This research might be done by a student writing his master thesis, who is supervised by me, because both the development of the logic-oriented modelling of these approaches and the subsequent formal analysis of the inference-proofness are supposed to be similar to the ones presented in [4].

Seen from the point of view of inference control, the approach to vertical fragmentation considered in [4] 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 database instance vertically according to the inference-proof approach to vertical fragmentation considered and making only the "externally stored" part resulting from this fragmentation available to the pertinent user. So, this "externally stored" part can be treated as an alternative inference-proof database instance that does not contain any information which might enable the distinguished information receiver to infer any information considered to be secret.

Regarding the desired efficiency this generation of inference-proof materialized views has the advantage, that queries can be answered safely without employing costly mechanisms of (dynamic) inference control based on theorem proving: each query can be answered directly according to an existing inference-proof instance without the need of any monitoring. Of course, the generation of the desired materialized views might be of high computational complexity. But as the needed computations can be done beforehand, these computations can be seen as a kind of preprocessing as long as only queries are handled and updates do not occur.

Following up this idea of creating inference-proof materialized views, my research for the next year will deal with inference-proof materialized views created with existing algorithms for dynamic CQE, which are surveyed in [2]. Processing the identity query asking for a full relational instance, such an algorithm for dynamic CQE generates an inference-proof variation of the original instance as its output, which can be used as an inference-proof materialized view. As a long-term objective a comprehensive solution to the problem of inference-proof materialized views is intended to be designed. For that purpose it has to be evaluated which requirements in terms of confidentiality and availability can be fulfilled by each of the different possibilities to compute inference-proof materialized views. Moreover, it might be evaluated whether a combined use of some of the approaches under investigation increases availability without compromising confidentiality.

References

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

[2] Joachim Biskup. Usability confinement of server reactions: Maintaining inference-proof client views by controlled interaction execution. In Shinji Kikuchi, Shelly Sachdeva, and Subhash Bhalla, editors, *Databases in Networked Information Systems, DNIS 2010*, volume 5999 of *LNCS*, pages 80–106. Springer, 2010.

[3] Joachim Biskup, David W. Embley, and Jan-Hendrik Lochner. Reducing inference control to access control for normalized database schemas. *Information Processing Letters*, 106(1):8–12, 2008.

[4] 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. Springer, 2011. to appear.

[5] Valentina Ciriani, Sabrina De Capitani di Vimercati, Sara Foresti, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. Enforcing confidentiality constraints on sensitive databases with lightweight trusted clients. In Ehud Gudes and Jaideep Vaidya, editors, *Data and Applications Security XXIII, DBSec 2009*, volume 5645 of *LNCS*, pages 225–239. Springer, 2009.

[6] 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. Springer, 2009.

[7] Csilla Farkas and Sushil Jajodia. The inference problem: A survey. *ACM SIGKDD Explorations Newsletter*, 4(2):6–11, 2002.

[8] Hakan Hacigümüs, Sharad Mehrotra, and Balakrishna R. Iyer. Providing database as a service. In *Proceedings of the 18th International Conference on Data Engineering, ICDE 2002*, pages 29–40. IEEE Computer Society, 2002.

[9] Marcel Preuß. Untersuchungen zum Inferenzschutz von fragmentierten Speicherungen von Datenbankinstanzen. Diploma Thesis, Technische Universität Dortmund, 2010. http://www.marcelpreuss.de/.

[10] Pierangela Samarati and Sabrina De Capitani di Vimercati. Data protection in outsourcing scenarios: Issues and directions. In Dengguo Feng, David A. Basin, and Peng Liu, editors, *ACM Symposium on Information, Computer and Communications Security, ASIACCS 2010*, pages 1–14. ACM, 2010.

Belief Change Operations Under Confidentiality Requirements in Multiagent Systems

Cornelia Tadros
Chair 6 - Information Systems and Security
Technische Universität Dortmund
cornelia.tadros@tu-dortmund.de

In multiagent systems, several agents (i.a., 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. In our work in project A5 "Exchange and Fusion of Information under Availability and Confidentiality Requirements in MultiAgent Systems" of the "Collaborative Research Center SFB 876 - Providing Information by Resource-Constrained Data Analysis", we augmented an agent with additional components for declaring her confidentiality interests and, complementarily, components for controlling her interaction with other agents and effectively enforcing her declared interests.

We focus 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. The exchange of information relies on a common propositional language $\mathcal{L}_{pl}$. In this scenario, agent $\mathcal{D}$ holds the *role of a defender* against agent $\mathcal{A}$ in the *role of a (potential) attacker* who attempts to obtain confidential information against $\mathcal{D}$'s obligations or interests. Agent $\mathcal{D}$ is devised with the usual desired functionality of an interacting agent: $\mathcal{D}$ faces generally incomplete information (*current assertions* $\mathcal{R} \subset_{fin} \mathcal{L}_{pl}$) about her environment, e.g., trading stock and offers in an e-commerce scenario etc. Moreover, in order to plan how to achieve her personal and the joint goals and to guide her decisions in the process of planing, agent $\mathcal{D}$ must be capable to draw reasonable conclusions from the available information

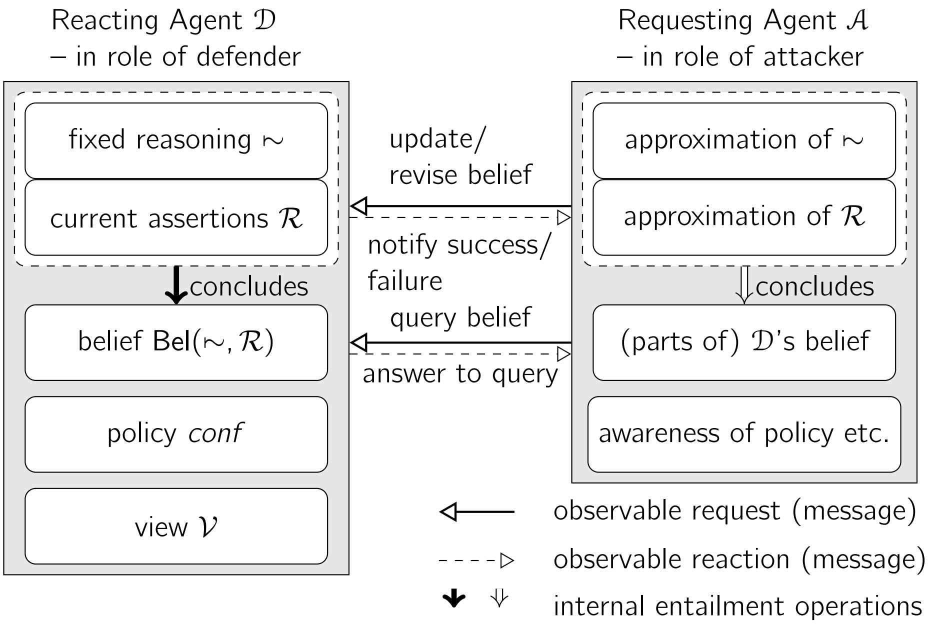

Reacting Agent $\mathcal{D}$
– in role of defender
fixed reasoning $\mathrel{|\!\sim}$
current assertions $\mathcal{R}$
concludes
belief $\mathsf{Bel}(\mathrel{|\!\sim}, \mathcal{R})$
policy *conf*
view $\mathcal{V}$
update/
revise belief
notify success/
failure
query belief
answer to query
Requesting Agent $\mathcal{A}$
– in role of attacker
approximation of $\mathrel{|\!\sim}$
approximation of $\mathcal{R}$
concludes
(parts of) $\mathcal{D}$'s belief
awareness of policy etc.
observable request (message)
observable reaction (message)
internal entailment operations

Figure 1: Interaction between two agents under confidentiality requirements

$\mathcal{R}$. This capability is modeled by a *consequence relation* (*fixed reasoning* $\mathrel{|\!\sim} \subseteq \mathcal{L}_{pl} \times \mathcal{L}_{pl}$) that maps assertions to conclusions. This reasoning can either be based on entailment or truth-value evaluation in propositional logic, used in classical database systems [2, 3, 5], or based on ordinal conditional functions (OCF) or other representations, used in non-monotonic reasoning [1, 4, 7]. Gathering all conclusions from her current assertions, $\mathcal{D}$ forms her *belief* $\mathsf{Bel}(\mathrel{|\!\sim}, \mathcal{R})$ about the environment:

$$\mathsf{Bel}(\mathrel{|\!\sim}, \mathcal{R}) := \{B \in \mathcal{L}_{pl} \mid \mathsf{con}(\mathcal{R}) \mathrel{|\!\sim} B\} \text{ where } \mathsf{con}(\mathcal{R}) := \begin{cases} \bigwedge_{F \in \mathcal{R}} F & \text{if } \mathcal{R} \neq \emptyset, \\ \mathit{true} & \text{otherwise.} \end{cases}$$

For confidentiality preservation, agent $\mathcal{D}$ is devised with two additional components, the policy *conf* (declaration of confidential information) and the view $\mathcal{V}$ (agent $\mathcal{A}$'s presumable view on $\mathcal{D}$'s fixed reasoning and current assertions), following the proposal in [4] for BDI agents.

Under the confidentiality aspect, agent $\mathcal{A}$ is a curiosity-driven reasoner about $\mathcal{D}$'s belief. Here, $\mathcal{A}$ is capable to conclude parts of $\mathcal{D}$'s belief after observing $\mathcal{D}$'s reactions to $\mathcal{A}$'s requests and accordingly approximating $\mathcal{D}$'s fixed reasoning and current assertions. In the following we will review our work in this project with two different instantiations of the outlined scenario.

In the area of classical database systems [5], agent $\mathcal{D}$ operates as a database server with a single client, agent $\mathcal{A}$. The database $\mathcal{R}$ corresponds to a propositional truth-value assignment which completely describes $\mathcal{D}$'s environment. Agent $\mathcal{D}$'s belief are all propositional formulas that evaluate to *true* with respect to the database $\mathcal{R}$. As a database client, agent $\mathcal{A}$ can request the truth-value of $A \in \mathcal{L}_{pl}$ (query) or request to insert/delete variables in the database (view update). These modifications may not violate integrity constraints $con \subset_{fin} \mathcal{L}_{pl}$ that define reasonable states of the database in the application context and must be evaluated to *true*. Agent $\mathcal{D}$ declares sensitive belief (here: *potential secrets*) as two disjoint sets $psec(TCP)$ and $psec(CCP)$ of propositional formulas collected in the confidentiality policy *conf* for agent $\mathcal{A}$. Informally, agent $\mathcal{A}$ is prohibited to conclude that $\mathcal{D}$ *currently* believes in a formula $B \in psec(TCP)$ (*temporary requirement*) and to conclude that $\mathcal{D}$ has ever believed in a formula $B \in psec(CCP)$, either previously or currently (*continuous requirement*). Against these prohibitions, agent $\mathcal{A}$ may exploit all released information (a finite set of valid formulas with respect to the database $\mathcal{R}$) and tracked effective updates to conclude the validity of sensitive belief in the current or preceding databases. Here, agent $\mathcal{A}$ reasons with propositional entailment and a syntax-based operator (*variable negation* [3]) that undoes effective updates. For enforcing confidentiality, agent $\mathcal{D}$ employs protocols that simulate $\mathcal{A}$ reasoning at runtime and refuse $\mathcal{A}$'s request if otherwise $\mathcal{A}$ was enabled to conclude sensitive current or previous belief. To judge the performance of the protocols under an availability policy of last-minute intervention ("refuse only if necessary"), in [5] we list other desirable properties of view update transaction protocols (beside confidentiality enforcement). Further, we show that the presented view update transaction protocol achieves all these properties while no other protocol can increase availability under the last-minute intervention policy without lacking one of these properties.

Beyond the classical complete propositional database model, in various application scenarios agent $\mathcal{D}$ cannot gain complete knowledge about her environment, but faces incomplete information $\mathcal{R} \subset_{fin} \mathcal{L}_{pl}$. From the available information $\mathcal{R}$, agent defeasibly concludes her belief about the environment by the consequence relation $\mathrel{|\!\sim}_{\kappa}$ defined by an OCF κ [1, 7]. In this scenario, in [6] we focus on the case that agent $\mathcal{A}$ requests queries about $\mathcal{D}$'s belief or requests revisions of $\mathcal{D}$'s belief. Via a belief revision request, agent $\mathcal{A}$ might add a formula $A \in \mathcal{L}_{pl}$ to $\mathcal{D}$'s current assertions $\mathcal{R}$. In the light of the additional information A, agent $\mathcal{D}$ might refute or withdraw previous belief. In the confidentiality policy $conf \subset_{fin} \mathcal{L}_{pl}$, agent $\mathcal{D}$ declares assertions believing in which agent $\mathcal{D}$ aims to hide from agent $\mathcal{A}$. Opposing $\mathcal{D}$'s confidentiality interest, agent $\mathcal{A}$ might desire to conclude what $\mathcal{D}$ certainly believes. For this purpose, agent $\mathcal{A}$ uses skeptical entailment (1) based on the following approximations: a set $\mathcal{B}^+ \subset_{fin} \mathcal{L}_{pl} \times \mathcal{L}_{pl}$ that describes $\mathcal{D}$'s observed behavior to draw conclusions, a set $\mathcal{B}^- \subset_{fin} \mathcal{L}_{pl} \times \mathcal{L}_{pl}$ that describes $\mathcal{D}$'s observed behavior not to draw conclusions, and a set $\mathcal{C} \subseteq \mathcal{L}_{pl}$ that describes which of the assertions propositionally entailed by $\mathcal{R}$ are visible to $\mathcal{A}$.

$$\begin{aligned}\mathsf{skeptical}(\mathcal{B}^+, \mathcal{B}^-, \mathcal{C}) = \{B \in \mathcal{L}_{pl} \mid & \text{ for each consequence relation } \mathrel{|\!\sim}' \\ & \text{possible under } \mathcal{B}^+,\ \mathcal{B}^- \text{ and } \mathcal{C} :\ \mathsf{con}(\mathcal{C}) \mathrel{|\!\sim}' B\}. \qquad (1)\end{aligned}$$

Informally, $\mathcal{A}$ considers a consequence relation possible iff this relation agrees with her approximations. In our work, we equipped agent $\mathcal{D}$ with procedures to control her reactions to queries and revisions by simulating $\mathcal{A}$'s skeptical reasoning at runtime. In particular, we formally proved that these procedures enforce confidentiality.

In ongoing research, we plan to implement the presented agents and evaluate the feasibility of our approach in comparison to confidentiality preserving agent with answer set programming developed also in this project. Further, we plan to augment $\mathcal{D}$'s belief $\mathsf{Bel}(\mid\!\sim, \mathcal{R})$ with degrees of beliefs where a higher degree expresses more confidence in the belief [1], based on an idea in [8] that "weakening of some degrees as a means to restrict access to information is a *more cooperative* way of communication than denying access altogether".

The work on this project is elaborated in the publications [5, 6].

References

[1] Christoph Beierle and Gabriele Kern-Isberner. A conceptual agent model based on a uniform approach to various belief operations. In Bärbel Mertsching, Marcus Hund, and Muhammad Zaheer Aziz, editors, *KI 2009*, volume 5803 of *LNCS*, pages 273–280. Springer, Heidelberg, 2009.

[2] Joachim Biskup. Usability confinement of server reactions: Maintaining inference-proof client views by controlled interaction execution. In Shinji Kikuchi, Shelly Sachdeva, and Subhash Bhalla, editors, *DNIS 2010*, volume 5999 of *LNCS*, pages 80–106. Springer, Heidelberg, 2010.

[3] Joachim Biskup, Christian Gogolin, Jens Seiler, and Torben Weibert. Inference-proof view update transactions with forwarded refreshments. *Journal of Computer Security*, 19(3):487–529, 2011.

[4] Joachim Biskup, Gabriele Kern-Isberner, and Matthias Thimm. Towards enforcement of confidentiality in agent interactions. In *NMR 2008*, pages 104–112, 2008.

[5] Joachim Biskup and Cornelia Tadros. Inference-proof view update transactions with minimal refusals. In *DPM 2011*, LNCS. Springer, Heidelberg, 2011. to appear.

[6] Joachim Biskup and Cornelia Tadros. Revising belief without revealing secrets. 2011. submitted soon.

[7] Gabriele Kern-Isberner. *Conditionals in Nonmonotonic Reasoning and Belief Revision - Considering Conditionals as Agents*, volume 2087 of *LNCS*. Springer, Heidelberg, 2001.

[8] Lena Wiese. Keeping secrets in possibilistic knowledge bases with necessity-valued privacy policies. In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann, editors, *IPMU 2010*, volume 6178 of *LNCS*, pages 655–664. Springer, Heidelberg, 2010.

Subproject B1
Analysis of Spectrometry Data with Restricted Resources

Sven Rahmann Jörg Ingo Baumbach

Resource-Constrained Analysis of Spectrometry Data

Dipl.-Inf. Dominik Kopczynski
Bioinformatics for High-Throughput Technologies
Chair 11 for Algorithm Engineering, TU Dortmund
dominik.kopczynski@tu-dortmund.de

Ion mobility spectrometry (IMS) devices are developed by B&S Analytics[1]. We introduce a new method to reduce the amount of data and extract appropriate features in measurements for an easier handling in further processes with respect to constrained resources on the used hardware.

Overview

Ion mobility spectrometry (IMS) is a technology to measure the presence and concentration of compounds in the air. Coupling an IMS with a multi-capillary column (MCC) for pre-separation yields more precise data. We obtain a two dimensional spectrum with retention time r, drift time d and the electric charge as the measured signal s. Measurements can be visualized as a two-dimensional heatmap, shown in Figure 1. A whole MCC/IMS measurement consists of 3 to 75 million data points depending on the resolution. Regions with a high signal intensity are called peaks. Peaks consist of several hundred data points. The challenge is to detect the peaks to draw conclusions from the position and shape of the peaks about the compounds.

Preprocessing

As we can see in the visualization the data is noisy. Therefore the data is pre-separated before further analysis. A feature that appears in all MCC/IMS measurements is the reaction-ion peak (RIP), visible as a continious run at drift time $d = 17.5$ms in Figure 1. For a RIP elimination a baseline correction is appropriate. Every signal in all MCC chromatograms is reduced by the median of the chromatogram it belongs to. To reduce

[1] `http://www.bs-analytik.de/`

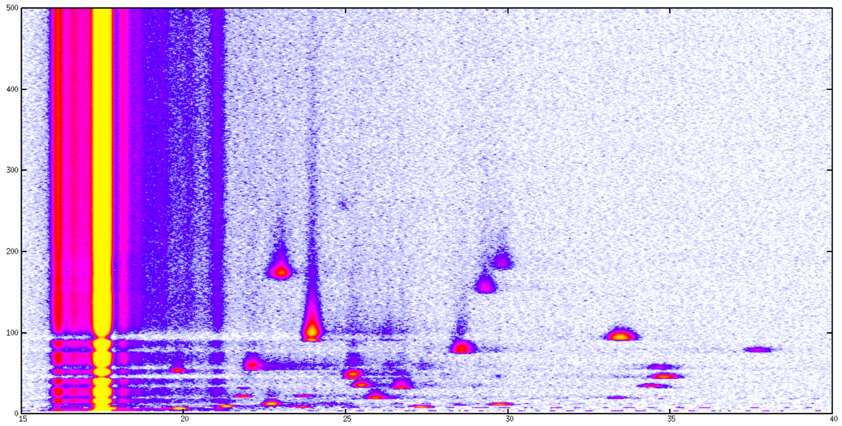


Figure 1: Heatmap of an MCC/IMS measurement. X-axis: drift-time d in ms; Y-axis: retention-time r in seconds; intensity increase: white - blue - purple - red - yellow

the noise we use a standard low-pass filter. Finally we are testing if a Gaussian blur or Savitzky-Golay filter yields better results for smoothing the data and if it is necessary to use further filters.

Describing Peaks with Models

The idea is to model the peaks with statistical functions for better comparisons [1] and further computations. A new approach is to find an appropriate model of statistical functions and to estimate the parameters with a statistic method. Because of the functionality of an MCC/IMS the peaks form a skewed curve in the retention time cross section as well as in the drift time cross section. In this case we chose the inverse gaussian (IG) distribution with parameters μ and λ. We modified the original IG with an additional parameter to set the origin o of the distribution.

$$IG(x;\mu,\lambda,o) = \left(\frac{\lambda}{2\pi(x-o)^3}\right)^{\frac{1}{2}} \cdot \exp\left(-\frac{\lambda((x-o)-\mu)^2}{2\mu^2(x-o)}\right)$$

The model is defined as a product of an IG in retention time r with an IG in drift time d. The volume under the curve equals 1. Thus we insert a last parameter for the volume or accordingly the weight ω in the measurement. The 2D peak model function is defined as:

$$M(r,d;\mu_r,\lambda_r,o_r,\mu_d,\lambda_d,o_d,\omega) = \begin{cases} \omega \cdot IG(r;\mu_r,\lambda_r,o_r) \cdot IG(d;\mu_d,\lambda_d,o_d) & \text{if } r > o_r \wedge d > o_d, \\ 0 & \text{otherwise} \end{cases}.$$

EM algorithm

Since peaks can overlap it is not appropriate to assign the data points to a single cluster. Thus it is more attractive to use the EM algorithm [2], which has the advantage of soft clustering. The algorithm estimates parameters in statistic models with given data points. Its an iterative process with two alternating phases: In the expectation "E" phase hidden values are estimated. This values discribe the weight of every data point to all models. The sum of all weights for one data point equals 1. The weight is estimated by the probability density function of the specific model. In the maximization "M" phase all parameters for all models are optimized with maximum-likelihood-estimators [3]. A typical approach is to work on a complete MCC/IMS measurement [4]. The adventages are that the preprocessing filters work better when the whole dataset is known and the decomposition of the peaks is more precise. Because our software concentrates on application areas in mobile devices the resources like storage or battery power are constrained. Hence it is necessary to build software that uses the hardware carefully. For this reason keep holding the whole measurement is not beneficial. The idea is to store just a few IMS spectra during the measurement and operate only in this window and subsequent discard the raw data after processing.

Therefore the EM algorithm is applied first on 1D IMS spectra belonging to one retention time point. After one EM execution one IMS spectrum for a single retention time is processed, hence we obtain parameterized 1D models in drift time. Now for all IMS spectra the EM process will be initiated. During this loop the detected 1D models at retention time $r + 1$ will be connected to their coherent models at retention time r. At the same time we use the estimated parameters of the models at retention time r to have good start parameters at $r + 1$. The results are n model chains [1]. After the first processing all model chains will be processed individually in a further EM step. In this step parameters in drift time as well as in retention time will be estimated for every peak. We obtain parameterized 2D models of the peaks.

Parameterized peaks have many advantages. Via parameterization the data will be extremely reduced. Depending on the size of the peaks (the amount of data points discribing a peak is in the range of a hundred to severeal thousand points) and the resolution of a measurement the factor of data reduction ranges between 10,000 and 250,000. Furthermore it is better to handle a data model for an alignment or resizing than discrete data points. Another issue where benefit from parameterization is the "ion-theft", visible in Figure 1 at retention time $r = 90$s. The peak at drift time $d = 34$ms "steals" ions from the peak at $d = 24$ms. This problem, which appears during a measurement, is generated by the function of an IMS. The ionization source which is used in the IMS does not ionize every molecule in the measured air. Thus we obtain incorrect data. With a statistical model it is easier to perform an ion reconstruction which is necessary for a correct 2D model parameter estimation. We developed a software, that already converts an MCC/IMS measurement file into a list of parameterized peaks.

Prediction of Models

A further issue is the prediction of compounds. Since the calibration and identification of one single metabolite is very costly in terms of time a prediction of the measurement (and the parameters) of an unregistered metabolite just by considering the chemical structure could be an alternative. So far three approaches came up. The first approach is to treat the peaks in the MCC chromatogram as poisson or inverse gaussian distributed, too. But the problem is that we have to assume that every molecule is injected into the MCC simultaneously. The second approach is to simulate the motion and collision of all molecules in the MCC. Of course such a simulation is very costly in terms of computation, but a prediction could be run at a compute server with no resource constrains since this computation need to be runed just one time with no critical time limits. The last approach is an abstraction of the full molecul simulation. We assume that the collisions in an MCC are irrelevant for the measurement. With appropriate mathematical functions we compute the time a molecule is carried in the mobile phase and sticks in the stationary phase. We obtain a retention time for every simulated molecule. For a simulation with a compound that is not already registered a method must be developed that computes correct parameters for the functions which assign the duration of the mobile and stationary phases. After a sufficient MCC simulation we will concentrate on an IMS simulation.

Open Issues

Next steps are to develop an alignment for peaks. Not only factors like temperature, velocity of carrier/drift gas or length of MCC/IMS tube change the measurement but also measurements of the same gas with different MCC/IMS devices vary in position and shape of the peaks. Another issue is the evaluation of feature selection. We want to compare all existing methods for pre-separation and peak detection. Thus we have to define criterias for the quality of the description of a spectrum with peak parameterization.

References

[1] D. Vogtland, Untersuchung von Ionenmobilitätsspektrometriedaten auf Peaks von Substanzen in verschiedenen Konzentrationen unter Einsatz von Glättungs- und Fittingverfahren, Diploma thesis, TU Dortmund, 2007.

[2] J. Bilmes, A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Technical report, 1998.

[3] R. Cheng and N. Amin, *American Statistical Association and American Society for Quality* (1981).

[4] D. Kopczynski, Datenreduktion und Merkmalsextraktion bei Ionen-Mobilitäts-Spektrometrie-Messungen, Diploma thesis, TU Dortmund, 2010.

Removing Adapter Sequences from High-Throughput Sequencing Data

Dipl.-Inform. Marcel Martin
Bioinformatics for High-Throughput Technologies
TU Dortmund
marcel.martin@tu-dortmund.de

High-throughput sequencing machines deliver increasing amounts of data. Current technology is able to determine, within one *run* of the sequencing instrument, the sequence of over 5 billion DNA fragments at a length of length 100 basepairs per fragment (*reads*).

For economic reasons, machines will operate at their highest possible capacity, minimizing the idle times. Since one instrument needs less than 14 days in order to finish a *run*, this constrains the maximum available for an analysis of that run.

There are many computationally expensive steps in processing the raw sequencing data until meaningful biological results are obtained. After on-machine image analysis, the reads need to be mapped to a known genome or assembled from scratch if no reference genome is known. Depending on the biological question, the next step would often be to find differences between the reference and the sequenced fragments. Those differences can then be classified as damaging mutations or harmless variations.

Many steps of that pipeline have already been highly optimized and so that they are currently not problematic regarding runtime. For analysis of small RNA fragments (microRNA), an intermediate filtering step is required, in which those parts of each sequenced DNA fragment are removed that do not belong to the original RNA molecule that was to be sequenced.

With our tool "cutadapt", which we describe in the following, that crucial part of the analysis pipeline for smallRNA/microRNA sequencing has now been sufficiently automated.

Overview High-throughput sequencing machines deliver huge amounts of data. Current sequencing technology is able to determine, within one *run* of the sequencing instrument, the sequence of over 5 billion DNA fragments at a length of length 100 basepairs per fragment.

For economic reasons, most machines will operate at their highest possible capacity, reducing the times at which the machine is idle as much as possible. Since one instrument needs less than 14 days in order to finish a *run*, this constrains the maximum time that can be used to analyze the data of that run.

There are many steps in processing the raw sequencing data until meaningful biological results are obtained. For example, there is the initial image analysis and quality value computation, which is done within the machines. After that, the DNA fragment sequences need to be mapped to a known genome or assembled from scratch if no reference genome is known. Depending on the biological question, the next step would often be to find differences between the reference and the sequenced fragments. Those differences can then be classified as damaging mutations or harmless variations.

Many steps of that pipeline have already been highly optimized and/or parallelized so that they are currently not problematic regarding runtime. For analysis of small RNA fragments (for example, microRNA), an intermediate filtering step is required, in which those parts of each sequenced DNA fragment are removed that do not belong to the original RNA molecule that was to be sequenced.

Through our work, that crucial part of the analysis pipeline for smallRNA/microRNA sequencing has now been sufficiently automated. The resulting tool – called "cutadapt" – is available as an Open Source tool.

As user feedback indicates, the program is also easy to use. We therefore argue that also the efficiency with which the highly constrained resource "human work time" is used has improved by providing this tool to the bioinformatics community.

Introduction The lengths of individual nucleotide sequences (reads) output by second-generation sequencing machines have reached 35, 50, 100 and more. When DNA or RNA molecules are sequenced that are shorter than this length, especially in small RNA sequencing experiments, the machine sequences into the adapter ligated to the 3' end of each molecule during library preparation. Consequently, the reads that are output contain the sequence of the molecule of interest and also the adapter sequence. An essential first task during analysis of such data therefore is to find the reads containing adapters and to remove the adapters where they occur. Only the relevant part of the read is passed on to further analysis. In some cases, finding adapters is a sign of contamination and the reads containing them must be discarded entirely.

For both tasks, we suggest to use cutadapt, which is a user-friendly program for the command-line, supporting a variety of file formats produced by second-generation sequencers. It especially supports color space data as produced by Applied Biosystems' SOLiD sequencer.

Cutadapt is the only stand-alone tool that can correctly trim color space reads. It supports FASTQ, FASTA and also SOLiD .csfasta/.qual input files. It outputs results in FASTA or FASTQ format. Gzip-compression of input or output files is automatically detected.

Implementation Cutadapt is mainly written in Python. The alignment algorithm is implemented in C as a Python extension module. The program was developed on Ubuntu Linux, but tested on Windows and Mac OS X.

Features The program was initially developed to trim 454 sequencing data collected by Zeschnigk et al. (2009). As insertions and deletions within homopolymer runs are common in 454 data, cutadapt supports gapped alignment. For small RNA data analysis by Schulte et al. (2010), the program was modified to support trimming of color space reads. It has also been tested and works well on Illumina data. Cutadapt can search for multiple adapters in a single run. It can optionally search and remove an adapter multiple times, which is useful when library preparation led to an adapter being appended multiple times. It can either trim or discard reads in which an adapter occurs. Reads that are outside a specified length range after trimming can also be discarded.

Performance In theory, adapter trimming with cutadapt is dominated by the computation of alignments, which is $O(nk)$, where n is the total number of the characters in all reads and k is the sum of the length of the adapters. In practice, other operations such as reading and parsing the input files take up more than half of the time.

With 35 bp color space reads and an adapter of length 18, cutadapt trims approximately 1 million reads per minute (0.06 ms per read) on a single core of a 2.66 GHz Intel Core 2 processor.

Color space reads Cutadapt correctly deals with reads given in SOLiD color space. When an adapter is found, the adapter and the color preceding it must be removed as that color encodes the transition from the small RNA into the adapter sequence and could otherwise lead to a spurious mismatch during read mapping.

Documentation Use cases are documented in the README file within the cutadapt distribution and also on the web site. Full documentation for all parameters is available by typing "cutadapt –help" on the command line.

Algorithm In the following, a character is a nucleotide or a color (encoded by 0-3). The first step in processing a single read is to compute optimal alignments between the read and all given adapters. Cutadapt computes either "regular" or slightly modified semiglobal alignments. Regular semiglobal alignments, also called end-space free alignments, do not penalize initial or trailing gaps. This allows the two sequences to shift freely relative to each other. When the "-a" parameter is used to provide the sequence of an adapter, the adapter is assumed to be ligated to the 3' end of the molecule and the behavior of cutadapt therefore is to remove the adapter and all characters after it. With regular semiglobal alignments, a short, usually random match that overlaps the beginning of a read would lead to the removal of the entire read. We therefore require that an adapter starts at the beginning or within the read. This is achieved by penalizing initial gaps in the read sequence, which is the only modification to regular overlap alignment.

Regular semiglobal alignment is used when the location of the adapter is unknown (assumed when the "-b" parameter is used). Then, if the adapter is found to overlap the beginning of the read, all characters before the first non-adapter character are removed.

After aligning all adapters to the read, the alignment with the greatest number of characters that match between read and adapter is considered to be the best one. Next, the error rate e/l is computed, where e is the number of errors and l is the length of the matching segment between read and adapter. Finally, if the error rate is below the allowed maximum, the read is trimmed.

Conclusion Cutadapt solves a small, but important task within sequencing pipelines, especially those for small RNA. It offers an easy-to-use command-line interface. If color space is to be processed, then cutadapt is the only standalone tool that supports this.

References

J. H. Schulte, T. Marschall, M. Martin, P. Rosenstiel, P. Mestdagh, S. Schlierf, T. Thor, J. Vandesompele, A. Eggert, S. Schreiber, S. Rahmann, and A. Schramm. Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. *Nucleic Acids Res*, 38(17):5919–5928, Sep 2010. doi: 10.1093/nar/gkq342. URL `http://dx.doi.org/10.1093/nar/gkq342`.

M. Zeschnigk, M. Martin, G. Betzl, A. Kalbe, C. Sirsch, K. Buiting, S. Gross, E. Fritzilas, B. Frey, S. Rahmann, and B. Horsthemke. Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. *Hum Mol Genet*, 18(8):1439–1448, Apr 2009. doi: 10.1093/hmg/ddp054. URL `http://dx.doi.org/10.1093/hmg/ddp054`.

Human breath analysis using MCC/IMS and GC/MSD

Kathrin Rupp

KIST Europe – Korea Institute of Science and Technology Europe Forschungsgesellschaft mbH

kathrin.rupp@kist-europe.de

Introduction

Ion mobility spectrometry (IMS) is generally used for direct breath analysis with respect to biomarker finding and gas trace analysis. Using IMS, ions are formed from the metabolites directly in air at ambient pressure, and the drift time within the spectrometer is measured. About 10 ml of breath is necessary to carry out a full analysis [1]. An IMS coupled to a MCC allows the identification of volatile metabolites occurring in human breath down to the ng/l- and pg/l- range of analytes in less than 500 seconds (see Figure 1).

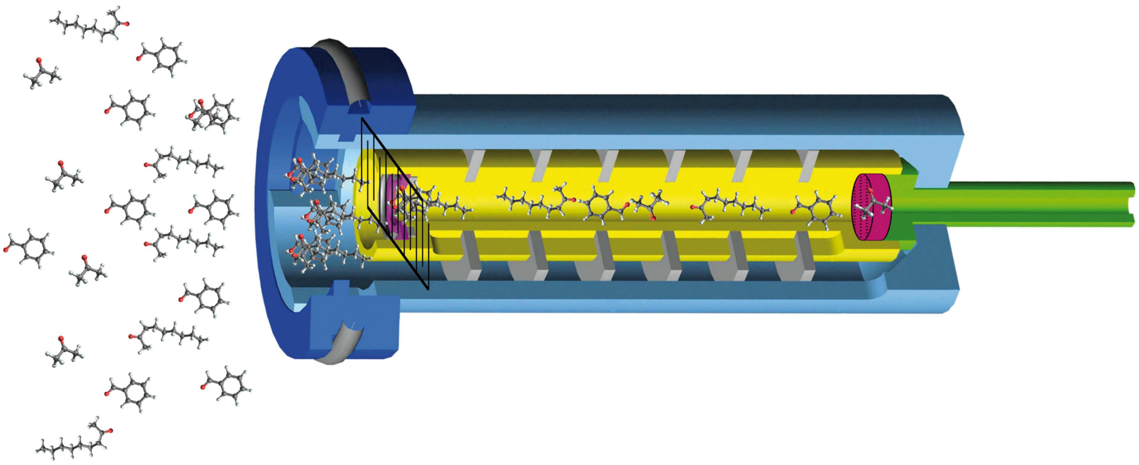


Figure 1: Working principle of an ion mobility spectrometer

The main aspect of this work is to create a database large enough to find specific metabolites in human breath, whereupon the human breath analysis can be used as medical diagnostics.

Parallel gaschromatographie (GC)-measurements were done to compare and confirm the MCC/IMS results. The samples are taken with tenax tubes with an adsorber inside, where the analytes adsorb. These tubes are heated for 10 min and the analytes enter the cryotrap, where they are focused. In the next step the cryotrap is heated and the analytes enter the column in the GC oven, where a pre separation takes place. After this the molecules are ionized by electron ionization and separated by the quadrupole and finally detected. That means the separation is carried out according to mass/charge ratio and because of the electron ionization you get characteristic fragments.

Processing

In cooperation with the Max Planck Institute (MPI) we are building an IMS database to use it as basis for further studies.

Medical samples are taken and measured parallel using MCC/IMS and GC/MSD (Gas chromatography coupled to mass spectrometry). Several Peaks should be identified and verified by comparing the spectra of the IMS and the GC/MSD. The samples for the GC-measurements were taken using TDS tubes.

An optimized method for measuring the medical samples using GC/MSD should be created.

The following chromatogram compares the results of measuring a Fishermen's Friend® (for simulation purpose) using MCC/IMS (Figure 2) and GC/MSD (Figure 3).

The breath measurement with Fishermen's Friend® (FF) (Figure 2) shows certain peaks, always found in the human breath and the menthol peaks. These menthol peaks are compared to the peaks in the GC measurement below (Figure 3).

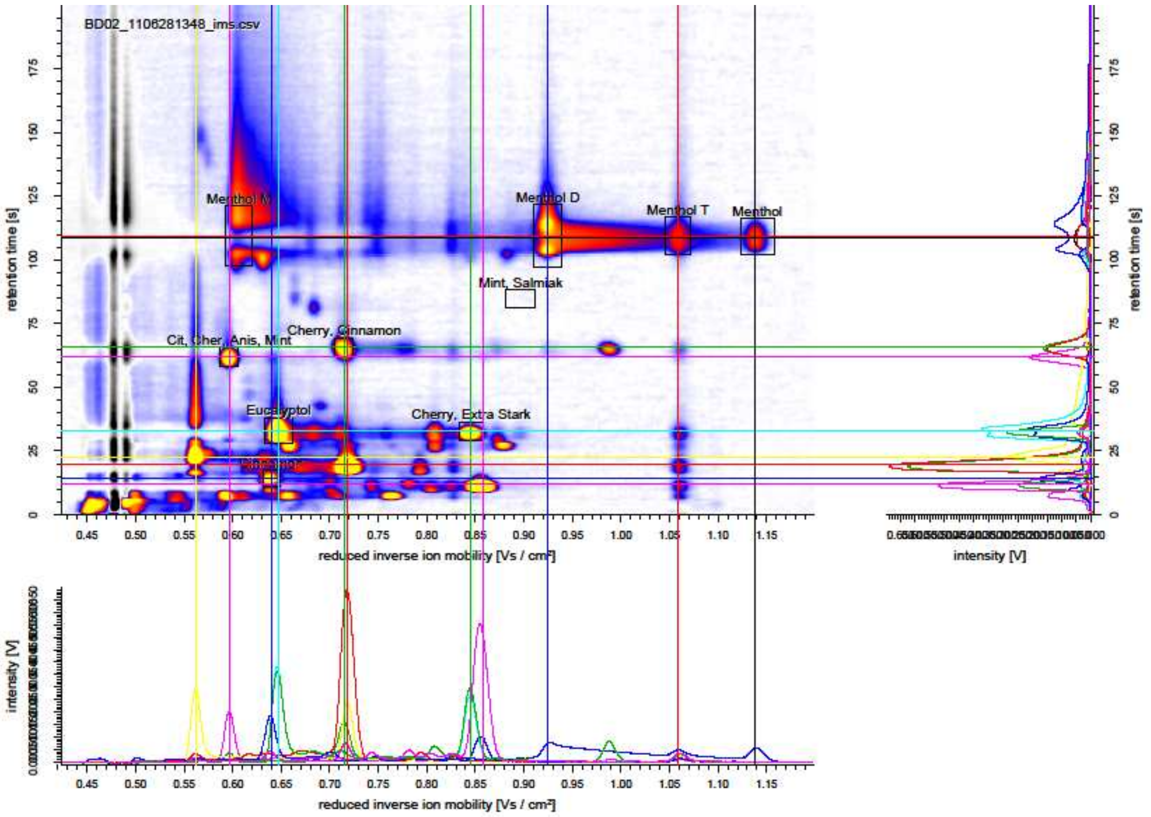


Figure 2: Heatmap of an MCC/IMS measurement.

```
File        :C:\GCMS Data\KIST_111010_F07061_AL_FF.D
Operator    : KR
Acquired    : 10 Oct 2011  12:04     using AcqMethod 111010_TDSA_substanceDB_HP-5ms_2.M
Instrument :   Instrument 1
Sample Name: KIST 111010 F07061 AL FF
Misc Info  :
Vial Number: 4
```

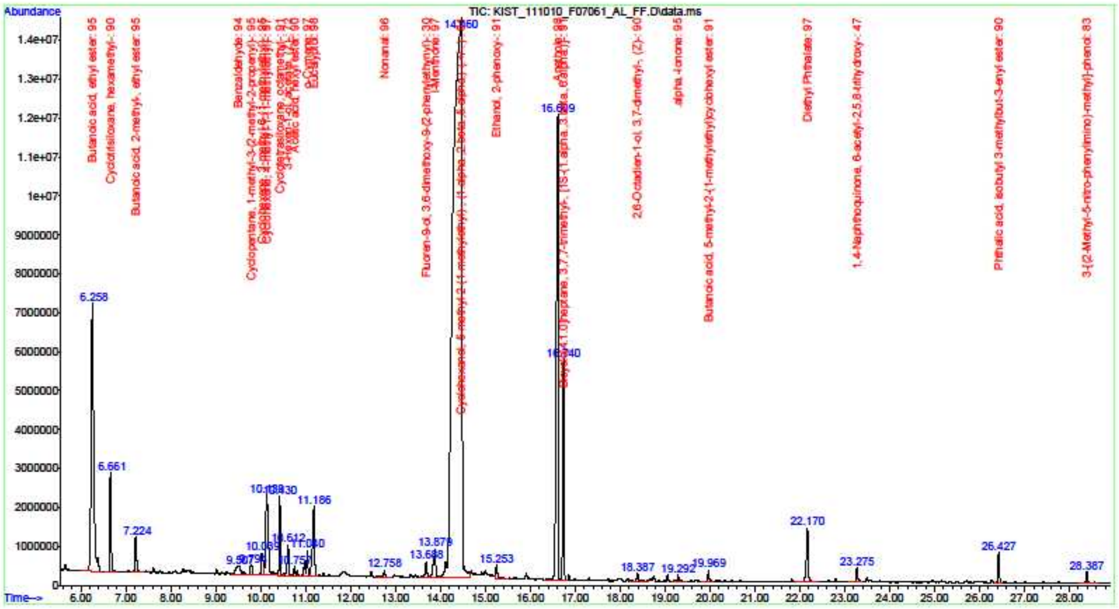


Figure 3: Chromatogram of a GC-measurement of Fishermen's Friend®

In the GC measurement we want to find several peaks, which we also presume to be present in the MCC/IMS to compare them. In this case you find a high number of peaks, which comes from various influences, not only from the breath itself. Other factors are the column, the tubes and so on; nevertheless the menthol peaks are easily identified, and confirm the menthol peaks in the MCC/IMS measurement.

Outlook

More reference measurements will be done to include them into the database. Further patient samples will be measured to find metabolites which are correlated to specific diseases.

The first step to find interesting metabolites already correlated to diseases is to look in the literature, acquire those substances, measure them with IMS an GC and finally add them to the database.

Another project is the toothpaste study. The aim of this study is to find out how long we can detect the toothpaste in human breath and how to do a correlation between the IMS and the GC measurements.

Various bacteria will be cultured and a model concerning the cell number and the metabolites will be established. The headspace from the cultured bacteria will be taken to identify the metabolites with IMS and GC and finally correlate the metabolites to infections or diseases.

References

[1] K.Rupp, S. Maddula, J.I. Baumbach, *Infections, drug delivery and metabolites detectable in human exhaled breath*, Poster HIPS-Symposium, 2011.

Characterizing Diseases based on MCC/IMS and GC/MS using Statistical Learning Techniques

Anne-Christin Hauschild

KIST Europe - Korea Institute of Science and Technology Europe
Forschungsgesellschaft mbH

Department Microfluidics and Clinical Diagnostics
Max Planck Institute for Informatics
Department Computational Systems Biology
ac.hauschild@kist-europe.de

It is common knowledge that the human breath contains information on the state of health of a person. This information is encoded by a combination of molecules produced by the human metabolism. The ion mobility spectrometer combined with a multi-capillary column (MCC/IMS) is a well known and sensitive technique to detect these volatile organic compounds (VOCs) within the human breath.
The application of statistical learning techniques is necessary to identify those metabolites / VOCs that are indicators for certain diseases. An outline is given, covering a general workflow starting from the data to the final biomarkers.

Introduction

In this report I will give an outline of the research I am going to do during my PhD studies, which started in July 2011. It will handle the tasks of acquisition and analysis of the data, produced by the previous mentioned MCC/IMS technique and data coming from a gas chromatograph coupled (GC) to a mass spectrometer (MS).

It is well known, that human exhaled air contains a combination of volatile organic compounds carrying potential information on the state of health of the human organism.

Several metabolites are already known as biomarkers for certain diseases, e.g. aceton is related to diabetes and nitric acid to asthma [2]. A drug, for example the anesthetic propofol, that is injected into the human blood system, can be detected in the breath of the patient [3].

In contrast to other methods the MCC/IMS is very sensitive; compounds down to the nanogram and picogram per liter range can be detected. Entering the MCC/IMS, the molecules are pre-separated by the multi-capillary column, further separated by the ion mobility spectrometer and detected by the Faraday plate, leading to a two dimensional set of data points.

The retention time (*rt*) within the multi capillary column, the inverse drift time (*1/K0*) in the ion mobility spectrometer and the measured signal (electric charge, *h*) can be visualized as a two dimensional heat map (see Figure 1).

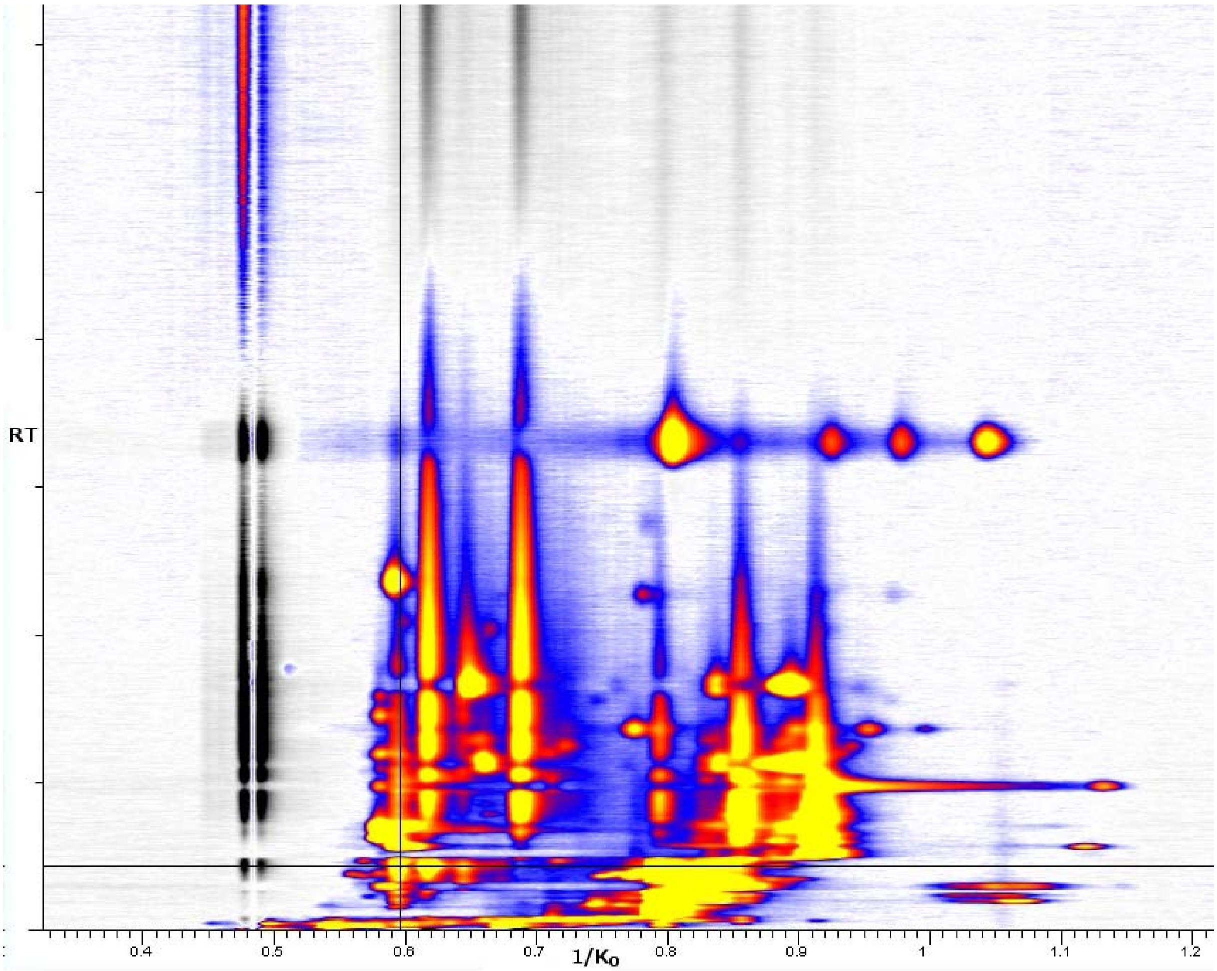


Figure 1: MCC/IMS spectrometer. The y axis represents the retention time, resulting of the multi-capillary column. The x-axis represents the drift time of the ions within the ion mobility spectrometer. The color of each point within the chromatogram represents the intensity of the signal.

In parallel to the investigations by the MCC/IMS, some of the samples will also be analyzed by the GC/MS. It is a widely-used technique in metabolomics, which quiet cost and time intensive compared to the MCC/IMS.

The GC/MS is also able to detect molecular compounds within the sample, but in contrast to the MCC/IMS a huge database of analytes is available to identify the molecule. Therefore this technique is included into this project.

Preprocessing

As previously mentioned, depending on the resolution, a MCC/IMS measurement contains up to 75 million entries. Even the analysis of just several hundred of these measurements by standard statistical learning techniques, would lead to huge problems in terms of performance and space.

Therefore several strategies have been designed during the last years, which are capable to decrease the noise and reduce the amount of features.

Bader et al. used lognormal detailing and wavelet transform to smooth and denoise the data [1]. The most intuitive way for feature reduction is to define regions within a two dimensional heat map that each represent a certain compound detected by the MCC/IMS as well as their specific position (retention and drift time) and height (quantity).

First of all there are software tools like VisualNow provided by B & S Analytik , which enables the user to manually select the promising areas.

Furthermore, there are automatic methods. At the beginning simple methods were used, for example: simple grid averaging or merging regions algorithms. A promising new approach is currently developed by Dominik Kopczynski, see reference [4] and his paragraph in this Techreport, for more details.

Workflow of the Project

The first step in the project is to establish a workflow to coordinate and combine the different information sources (see Figure 2).

The MCC/IMS and GC/MS data will be processed using the previously mentioned (VisualNow, B & S Analytics) and according software packages (AMDIS provided by the National Institute of Standards and Technology (NIST) [5] and ChemStation provided by Agilent Technologies [7]) respectively.

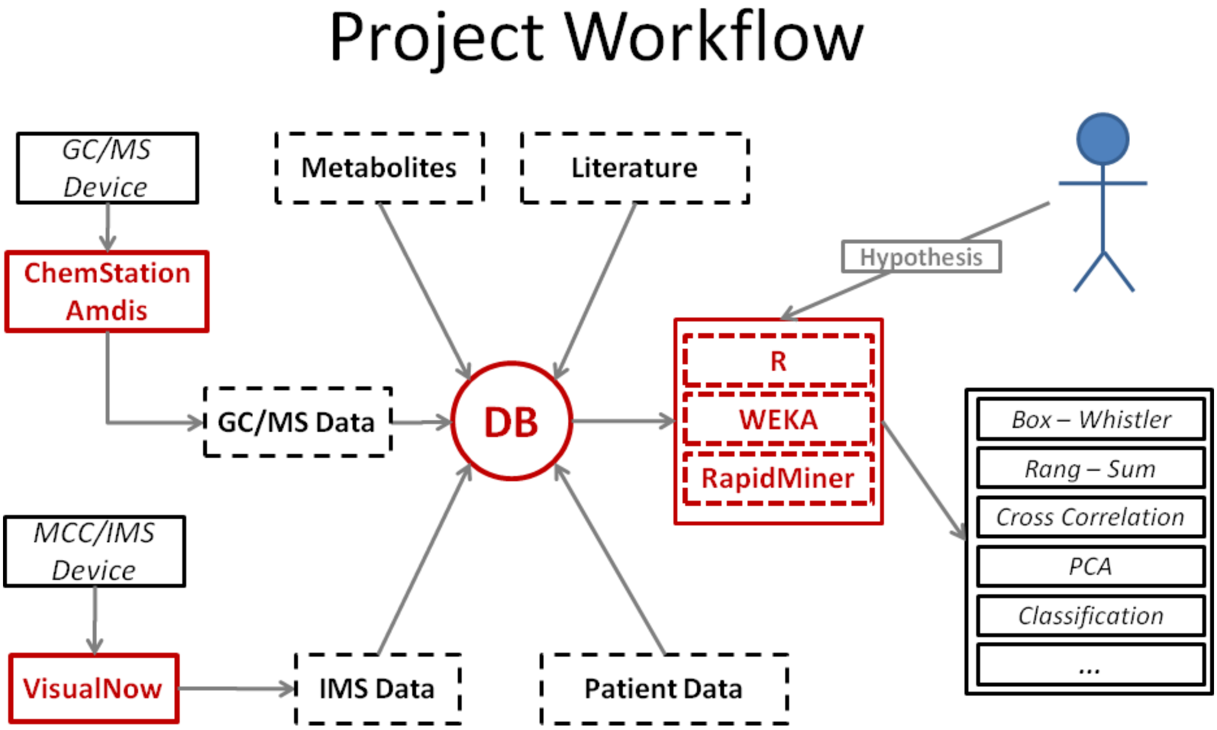


Figure 2: Sketch of the workflow of the project.

First of all the MCC/IMS data and the GC/MS data will be jointly stored in a database system. Assigning this data with the corresponding annotations like diseases or drug concentration, statistical learning techniques can be used to identify those metabolites / VOCs that are indicators for certain annotations.

[1] Bader, S. "Identification and Quantification of Peaks in Spectrometric Data", 2008.

[2] Baumbach, J. I., Vautz, W., Ruzsanyi, V. and Freitag, L. "Early detection of lung cancer: Metabolic profiling of human breath with ion mobility spectrometers"'Modern Biopharmaceuticals', Wiley-VCH Weinheim , 2005, pp. 1343-1358.

[3] Carstens, E., Hirn, A., Quintel, M., Nolte, J., Juenger, M., Perl, T. and Vautz, W. "On-line determination of serum propofol concentrations by expired air analysis," *Int. J. Ion Mobility Spectrom.* (13:1), 2010, pp. 37-40.

[4] Kopczynski, D. "Datenreduktion und Merkmalsextraktion bei Ionen-Mobilitдts-Spektrometrie-Messungen", 2010.

[5] " Automated Mass Spectral Deconvolution and Identification System (AMDIS)", http://chemdata.nist.gov/mass-spc/amdis/, 2011.

[6] "B & S Analytik", http://www.bs-analytik.de/, 2011.

[7] "Agilent ChemStation", http://www.chem.agilent.com/en-US/products/software/ chromatography/ chemstation/, 2011.

SFB 876 Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

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

Peter Marwedel Heinrich Müller Alexander Zybin

Methods for Analyzing PAMONO Biosensor Data (part of project B2)

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

A brief summary of the progress made in project B2 is given, covering the fields of image processing, classification and parameter optimization in the context of the PAMONO biosensor. The current status as well as planned work are described. The resulting methodology aims at enabling detection of viruses and other nano-objects in the data obtained from PAMONO.

1 Introduction

Imaging nano-scale objects, like e.g. viruses, by direct light microscopy is rendered impossible by the diffraction barrier, as discovered by Ernst Abbe in 1873: Objects with sizes below 200nm can not be imaged using this technique. The PAMONO[1] biosensor [22, 23], which is the practical focus of project B2, does not overcome the diffraction barrier, like e.g. STED microscopy [9], but enables indirect detection of nano-scale objects by observing effects they cause on the micrometer scale. These effects are due to surface plasmon resonance, resulting in a localized, micrometer-scale increase of surface reflectivity around a nano-scale object attaching to the mobilization layer of the sensor. Thus viruses in liquid samples can be detected indirectly. In order to do so, the sensor data must be analyzed. Developing methods for this analysis is the topic of the planned dissertation. The associated research can be divided into three main aspects:

1. Image processing of the time-series of images generated by the sensor
2. Classification of the observed image content
3. Optimization of the parameterized processing pipeline with respect to detection quality, speed and consumption of resources.

This report is structured as according to these aspects, presenting the current status in section 2 and the work plan in section 3.

[1] Plasmon Assisted Microscopy Of Nano-Objects

2 Current status

Image Processing: On the image processing side, the current setup uses the components described in [12]. They are assembled to a processing pipeline consisting of

(a) background removal (subtraction or division, fixed or sliding average),
(b) denoising based on Haar-wavelet coefficients [14],
(c) per pixel time-series matching to determine virus candidate pixels, and finally
(d) aggregation of adjacent virus candidate pixels to form virus candidate polygons, using the Marching Squares algorithm [8].

Each step is executed in real-time parallel processing on the GPU. The result is a segmentation of the spatio-temporal PAMONO sensor data, wherein the segments delineate virus candidate areas at certain points of time and certain positions on the sensor.

Classification: Separating true virus adhesions from erroneous detections is currently carried out on the polygon level. The classification pipeline detailed and evaluated in [18] uses polygon form-factors [11] as input features to estimate, whether or not a given polygon delineates an actual virus. A diverse set of classifiers is trained on manual classifications created by experts. Evolutionary parameter selection optimizes classification cross-validation accuracy [10] on the training data in an offline step. The optimized classifiers are then tested on previously unseen, manually classified examples to evaluate performance (70/30 stratified split validation [10]). The resulting average accuracies for different types of nano-particles are shown in Table 1. Averages were taken over different datasets obtained for 200nm and 280nm polystyrene spheres, as well as HIV virus-like particles (VLPs). The table shows accuracies attained by a one-class support vector machine (SVM) [4], and for two-class Naive Bayes (NB) [16], RIPPER rule induction (RI) [5], C4.5 Decision Trees (DT) [15], *k*-Nearest-Neighbor matching (kNN) [17] and support vector machine (SVM) [3]. The observed accuracies prove automated classification of PAMONO data through supervised learning a feasible approach.

Table 1: Automatic Classification Accuracies in %

Dataset	Avg.	One-Class	Two-Class				
		SVM	NB	RI	DT	kNN	SVM
280nm	89	85	**90**	89	89	**90**	89
200nm	92	89	89	93	**95**	94	90
HIV-VLP	87	82	87	88	88	88	**89**

Parameter Optimization: A survey of the literature pointed out existing multiobjective evolutionary algorithms (MOEAs) [20], such as SPEA2 [21] and NSGA-II [7], to be good candidates for the final parameter optimization.

3 Future Work

After the proof of concept presented in section 2, the pipeline is to be enhanced in a second iteration, developing new or extending upon existing methods for data analysis. With PAMONO sensor data as one field of application, general methods are to be developed. Applicability in more general contexts is to be verified by evaluating performance on e.g. astronomical data [19] and data from other microscopy techniques [9], which bear sufficient similarity to PAMONO data.

Image Processing: As a first step, a signal model will be developed, encompassing the data acquisition process and signal degradation. Its benefit is two-fold: Applying it forward generates synthetic data for validation, applying it backward serves signal reconstruction and may guide enhancement. A key issue encountered in this context is the low SNR because the amplitude of the desired signal is approximately 6% of the background signal. As the background is only approximately constant it introduces a further source of noise, besides the CCD sensor chip.

Efficient denoising methods are to be designed, which can account for the residual noise of the background signal, as well as exploit the temporal dimension of the data. To these ends, wavelet-based techniques are to be investigated for the temporal [6] as well as the spatial [13] dimensions. With regard to the latter, the small spatial extension of viruses needs to be accounted for.

A step from processing concrete PAMONO sensor data to general methods for analyzing noisy, background-polluted time-series data is to be taken.

Classification: In classifying PAMONO data, the set of employed features is to be extended: Highly discriminative features are to be identified and extracted. With regard to this, a fusion of virus candidate detection (currently based on time series analysis) and classification (currently based on polygon form factors) is desired, involving the design of combined spatio-temporal features as a natural step. In the temporal dimension, dynamic time warping [2] is to be explored, as well as amplitude-level features [1], which can be computed incrementally and are thus well-suited for GPGPU processing. Classification results will be validated on manual expert classifications and on the aforementioned synthetic data.

Parameter Optimization: The existing multiobjective evolutionary algorithms SPEA2 [21] and NSGA-II [7] will be applied to and evaluated on the parameter optimization problem. The goal is to find an approximation set of feasible solutions which is close to the theoretical Pareto-front of the problem.

References

[1] J. Assfalg, H.-P. Kriegel, P. Kroeger, P. Kunath, A. Pryakhin, and M. Renz. Similarity search in multimedia time series data using amplitude-level features. In *LNCS 4903*, 2008.

[2] D. J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. *Workshop on Knowledge Discovery in Databases*, 1994.

[3] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.

[4] Y. Chen, X. S. Zhou, and T. S. Huang. One-class SVM for learning in image retrieval. In *International Conference on Image Processing*, 2001.

[5] W. W. Cohen. Fast effective rule induction. In *Machine Learning: Proc. of the Twelfth International Conference*, 1995.

[6] R. Dahlhaus, J. Kurths, P. Maass, and J. Timmer, editors. *Mathematical Methods in Time Series Analysis and Digital Image Processing*. Springer, 2008.

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182 – 197, 2002.

[8] A. J. P. Gomes, I. Voiculescu, J. Jorge, B. Wyvill, and C. Galbraith. *Implicit curves and surfaces: mathematics, data structures and algorithms*. Springer, 2009.

[9] S. W. Hell and M. Kroug. Ground-state-depletion fluorscence microscopy: A concept for breaking the diffraction resolution limit. *Applied Physics B: Lasers and Optics*, 60:495–497, 1995. 10.1007/BF01081333.

[10] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

[11] G. Landini. Quantitative analysis of the epithelial lining architecture in radicular cysts and odontogenic keratocysts. *Head & Face Medicine*, 2, 2006.

[12] P. Libuschewski, C. Timm, D. Siedhoff, F. Weichert, H. Müller, and P. Marwedel. Improving nanoobject detection in optical biosensor data. In N. et al Callaos, editor, *Proceedings of the 15th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2011*, volume II, pages 236–240, Orlando, Florida, USA, 2011. IIIS.

[13] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 2008.

[14] C. R. Mittermayr, S. G. Nikolov, H. Hutter, and M. Grasserbauer. Wavelet denoising of gaussian peaks: a comparative study. *Chemometrics and Intelligent Laboratory Systems*, 34:187–202, 1996.

[15] J. R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artifcial Intelligence Research*, 1996.

[16] I. Rish. An empirical study of the naive Bayes classifier. Technical report, IBM Research Division Thomas J. Watson Research Center, 2001.

[17] Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk, editors. *Nearest-neighbor methods in learning and vision*. MIT Press, 2005.

[18] D. Siedhoff, F. Weichert, P. Libuschewski, and C. Timm. Detection and classification of nano-objects in biosensor data. *Microscopic Image Analysis with Applications in Biology*, 2011. Preprint, Accepted for Publication.

[19] J.-L. Starck and F. Murtagh. *Astronomical Image and Data Analysis*. Springer, second edition, 2006.

[20] E. Zitzler, M. Laumanns, and S. Bleuler. A tutorial on evolutionary multiobjective optimization. *Swiss Federal Institute of Technology (ETH) Zurich*, 2004.

[21] E. Zitzler, M. Laumanns, and L. Thiele. Spea2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. *Evolutionary Methods for Design, Optimisation and Control*, 2002.

[22] A. Zybin, E. Gurevich, F. Weichert, and K. Niemax. SPR detection of single nano particles and viruses. In *4th International Scientific Conference on Physics and Control (PHYSCON)*, 2009.

[23] A. Zybin, Y. Kuritsyn, E. Gurevich, V. Temchura, K. Überla, and K. Niemax. Real-time Detection of Single Immobilized Nanoparticles by Surface Plasmon Resonance Imaging. *Plasmonics*, 5:31–35, 2010.

Collaborative Research Center SFB 876 - Nanoobject Detection in Optical Biosensor Data

Pascal Libuschewski
Lehrstuhl für Graphische Systeme
Technische Universität Dortmund
pascal.libuschewski@tu-dortmund.de

This report presents the detection of nano-size objects in optical biosensor data. The novel PAMONO sensor is capable to visualize nano-size objects, like viruses, with a gray scale camera and only little magnification. To efficiently process, analyze and classify the biosensor image data a high performance approach is used, resulting in real-time diagnosis of virus occurrences in the sample. Also an estimate of the concentration can be obtained in real-time, if that concentration is not too high.

The contribution of this work is an optimization of the processing pipeline used for PAMONO sensor data analysis. The following objectives are optimized: detection-quality, speed and consumption of resources (e.g. energy, memory). Thus our approach respects the constraints imposed by medical applicability, as well as the constraints on resource consumption arising in embedded systems. The parameters to be optimized are descriptive (virus appearance parameters) and will be hardware-related (design space exploration) in the future.

The PAMONO sensor (Plasmon assisted Microscopy of Nano-Size Objects) [6–8] is a novel sensor which consists of a 50nm thick gold layer, which is mobilized on the upper side where the specimen are supplied in a liquid. On the bottom side of the gold layer a prism is attached which deflects the laser light from a laser to the gold layer and then through a magnifying lens into a camera. The PAMONO sensor unit produces a video stream, which is 1000 × 500 pixels in size and has a framerate of 30 frames per second. The image data is dominated by the reflected light of the gold layer and is superimposed by noise. The wanted signal of the small nano particles attaching to the mobilized gold layer is very small in amplitude, about 6% of the average signal. As a particle binds to the gold layer, for example by gene-antigene binding, the reflected light is influenced by the changed layer thickness resulting in a small increase of intensity at the spatial coordinates of the binding.

To detect the nano particles a GPGPU approach is used. The detection process consists of different steps. First the input data is preprocessed, for example by removing noise with wavelet denoising or with fuzzy logic noise reduction. In the second step, the pixels of interest are identified. As each particle adhesion produces an increased intensity in a small area around the adhesion (in a short period of time), each time series is matched to a variable pattern, resulting in a pre-classification of all pixels in space as virus candidates. Figure 1 shows an example time series of a virus adhesion and an optimized pattern. Each thread on the GPU matches one of the time series in a highly parallel manner. To match the time series with a characteristic pattern the time series is pre-processed to fit into the range of minus one to one. Then the distance between the current time series and the pattern is calculated. In the third step the single pixels are combined to polygonal segments, with a parallel marching squares algorithm. Polygons from different times are combined if they belong to one particle adhesion. In a last step the remaining polygons are classified as particle or non-particle, based on their form factors [2]. All steps, except the tracing of the polygons, are done in the temporal dimension, which gives advantages in parallelization as each pixel position is independent of others. This avoids the need for synchronization of the threads on the GPU. A detailed analysis of the scaling behavior of the algorithms can be found in [5].

As it is unclear which patterns, thresholds and form factor values should be used to seperate the particle class from the non-particle class, a genetic algorithm is used to optimize the manual parameters. For an introduction to genetic algorithms see [4]. In the following, the genetic optimization process is explained using the example of the pattern in the second pipeline step which can be seen in figure 1. More details of this step can be found in [3].

The pattern for the matching with the time series is converted into genes by representing each value of the pattern as a floating point number. Together these genes build the chromosome for the genetic algorithm, which is used to generate a population. As it is unlikely that a random start population holds an individual with genes good enough to detect a particle, the start population is defined manually with different simple patterns that can result in a good detection. Building on this start population the genetic algorithm evolves the population and after about 22.000 single runs of the pipeline, the improvement from one generation to the other is small enough to meet the stop criterion. Accuracy is used as the fitness function of the genetic algorithm, defined as $\frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{fn}+\mathrm{fp}+\mathrm{tn}}$ [1]. The true positives (tp) count the correct classified particles, the false negative (fn) count the not detected particles and the false positives (fp) count the detected artifacts. True

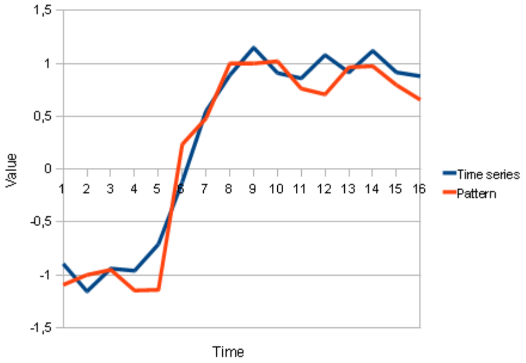


Figure 1: A scaled input time series (blue) is matched to the pattern (red).

Datasets	Precision	Recall	Accuracy
280 nm	95 %	90 %	86 %
200 nm	87 %	91 %	80 %
VLPs	80 %	86 %	71 %

Table 1: Detection results of different datasets, based on true positives, false positives and false negatives.

negatives (tn) are set to zero, as they are not properly defined in this scenario. To count these numbers, the genetic algorithm needs to verify the calculated results from one individual with some ground truth data which is given by manually classified datasets.

The results from the improved detection quality are shown in table 1, quantifying precision, recall and accuracy. As can be seen the detection quality depends on the particle size. This is due to the condition that the signal to noise ratio is worse for smaller particles. Small particles result in a lower intensity increase and therefore this lower intensity is closer to the noise level making it difficult to clearly identify the wanted signal.

A method to optimize the parameters of the detection pipeline was presented. An automated detection of viruses or nano particles in such data can be carried out in real-time, using GPGPU computing. Despite the early stage of development, the detection results are promising in terms of accuracy.

Future research in the analysis of PAMONO sensor data aims in an improvement of the overall detection quality. Therefore the noise level needs to be lowered resulting in a better differentiation between the noise and the signal. A 3-D fuzzy logic noise reduction algorithm is already implemented but not yet evaluated. Also the detection quality should be improved by more complex detection algorithms. Besides the detection quality speed and energy consumption need to be improved. In the next step a design space exploration will be done, exploring which hardware requirements fit the problem best. For this step the hardware will be fully simulated giving a control over many hardware parameters, enabling to use some kind of genetic- or learning-algorithm to find the best possible hardware.

References

[1] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Labs, 2004.

[2] G. Landini. Quantitative analysis of the epithelial lining architecture in radicular cysts and odontogenic keratocysts. *Head & Face Medicine*, 2, 2006.

[3] Pascal Libuschewski, Constantin Timm, Dominic Siedhoff, Frank Weichert, Heinrich Müller, and Peter Marwedel. Improving nanoobject detection in optical biosensor data. In N. et al Callaos, editor, *Proceedings of the 15th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2011*, volume II, pages 236–240, Orlando, Florida, USA, July 2011. IIIS.

[4] K. F. Man, K. S. Tang, and S. Kwong. *Genetic algorithms: concepts and designs*. Springer, 2001.

[5] Constantin Timm, Frank Weichert, Peter Marwedel, and Heinrich Müller. Design space exploration towards a realtime and energy-aware gpgpu-based analysis of biosensor data. *Computer Science - Research and Development, Special Issue "International Conference on Energy-Aware High Performance Computing (ENA-HPC)", Springer (accepted for publication)*, 2011. Publikation.

[6] F. Weichert, M. Gaspar, C. Timm, A. Zybin, E. Gurevich, M. Engel, H. Müller, and P. Marwedel. Signal Analysis and Classification for Surface Plasmon Assisted Microscopy of Nanoobjects. *Sensors and Actuators B: Chemical*, 151:281–290, 2010.

[7] A. Zybin. Verfahren zur hochaufgelösten erfassung von nanopartikeln auf zweidimensionalen messflächen, 2010. Patent DE102009003548A1.

[8] Alexander Zybin, Y Kuritsyn, E. Gurevich, V. Temchura, K. Überla, and K. Niemax. Real-time Detection of Single Immobilized Nanoparticles by Surface Plasmon Resonance Imaging. *Plasmonics*, 5:31–35, 2010.

Hardware/Software Co-Design and Multi-Objective Program Optimizations for GPGPU Applications under Resource Constraints

Constantin Timm
Lehrstuhl für Eingebettete Systeme
Technische Universität Dortmund
constantin.timm@cs.tu-dortmund.de

Latest chip design trends make GPGPU computing extremely interesting for embedded systems and especially for image processing systems. The utilization of such chips and the restricted availability of energy make an appropriate hardware/software co-design strategies and multi-objective program optimizations mandatory for the design of such GPGPU-based systems.

For scientific and industrial HPC (High Performance Computing) applications, utilizing GPGPU-capable GPUs at the desktop or server level is widely accepted [5]. Even in embedded systems, GPGPU (General Purpose Computing on Graphics Processing Units) can be used in order to accelerate parallel general purpose applications. The trend towards GPGPU computing can be noticed by the fact that GPU vendors, such as Nvidia and ATI, provide developers with dedicated GPGPU frameworks like CUDA [4] and Stream [1] to create applications for their GPUs. Moreover, initiatives such as OpenCL [2], make programming of GPUs easier. GPGPU-implemented applications comprise a wide area including video and image processing, simulation, cryptography or machine-learning [3].

The development of locally available specialized virus detection systems becomes increasingly important in the face of the growth of worldwide spreading virus infections. Advances in the medical sector enable the utilization of optical microscopy for the detection of viruses. Microscopy-based detection methods can be used for a rapid and distributed epidemic infection control. A novel technique which can achieve the latter is named PAMONO (Plasmon assisted Microscopy of Nano-Size Objects) [10]. It provides the possibility to selectively detect nano-objects. PAMONO not only enables an in-situ detection but also a detection in realtime which means that the result of a detection in progress can be visualized online while inserting the specimen. The processing system must cope with a certain frame rate. This requires novel and efficient algorithms for the identification of nano-objects and efficient systems with HPC acceleration techniques for processing the data [9].

The expected local availability of such biosensors at distributed sites, such as airports or seaports suggests that the number of concurrently used systems can be large. Due to the increase of the worldwide energy consumption for computing systems, green computing gets more and more important and therefore energy awareness should be one of the objectives when designing such a GPGPU-accelerated system. The major requirements on such a virus detection system can be summarized as follows:

1. Provide a precise virus detection rate.
2. The system should work in real-time.
3. Take energy-efficiency into account.

The usual GPGPU application design process does not take the decision for the most energy efficient platform – including resource restrictions – into account. Therefore, it is proposed to consider the parameters and capabilities of the platform and its resource restrictions in the GPGPU application process. The platform exploration is necessary to choose the most efficient platform. This means that a platform has to meet the real-time constraints of the input device and is most energy efficient. Integrating a GPGPU-equipped graphics card into an embedded system can be profitable in order to reduce the energy consumption of the complete system. The following theses have been proposed for reducing the energy consumption of an embedded system by integrating an additional graphics card:

1. The higher the idle time of the GPU is, the less the integration of a graphics card for GPGPU computations into the system is useful in terms of energy.
2. The communication overhead from the system memory to the graphics card memory is more important than discussed in literature.
3. Techniques for saving power on a GPU can enable the integration of this GPU for reducing the energy consumption of the total system.

Overall, the integration of an additional GPGPU-capable graphics cards/processor into a system is a low-cost way of integrating high-performance computing resources and a promising way of saving energy by accelerating applications, in particular data parallel applications. In the face of green computing and the utilization of GPGPU capable chips in small and mobile systems, it should be mandatory to take the energy consumption as an objective during the design process of a GPGPU based system into account. A design space exploration should exploit load balancing and the parallel processing scalability for a GPGPU application. These tow parameters have a direct impact on the energy efficiency of the application and must be considered at design time. In future work dynamic voltage and frequency scaling techniques should be utilized in more accurate ways, to decrease the overall power consumption of the system. This aims at reducing the idle time of the graphics card and reducing the energy consumption even more.

The issue of integrating GPGPUs in embedded systems for saving energy was addressed in [6]. A multi-objective design space exploration towards the optimization of the energy consumption under the restriction of meeting certain deadlines was addressed in [7, 8].

References

[1] AMD Corporation. ATI Stream SDK, 2010.

[2] Khronos Group. OpenCL Specification, 2010.

[3] Hubert Nguyen, editor. *GPU Gems 2*. Addison-Wesley, 2007.

[4] NVIDIA Corporation. Collection of Applications and SDK.

[5] J. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. Lefohn, and T. Purcell. A Survey of General-Purpose Computation on Graphics Hardware. *Computer Graphics Forum*, 26(1):80–113, 2007.

[6] Constantin Timm, Andrej Gelenberg, Peter Marwedel, and Frank Weichert. Energy Considerations within the Integration of General Purpose GPUs in Embedded Systems. In *Proc. of the Annual Int. Conf. on Advances in Distributed and Parallel Computing (ADPC)*, November 2010.

[7] Constantin Timm, Pascal Libuschewski, Dominic Siedhoff, Frank Weichert, Heinrich Müller, and Peter Marwedel. Improving nanoobject detection in optical biosensor data. *Proceedings of the 5th International Symposium on Bio- and Medical Information and Cybernetics (BMIC 2011)*, 2:236–240, 2011. Publikation.

[8] Constantin Timm, Frank Weichert, Peter Marwedel, and Heinrich Müller. Design space exploration towards a realtime and energy-aware gpgpu-based analysis of biosensor data. *Computer Science - Research and Development, Special Issue "International Conference on Energy-Aware High Performance Computing (ENA-HPC)", Springer*, 2011.

[9] F. Weichert, M. Gaspar, C. Timm, A. Zybin, E. Gurevich, M. Engel, H. Müller, and P. Marwedel. Signal Analysis and Classification for Surface Plasmon Assisted Microscopy of Nanoobjects. *Sensors and Actuators B: Chemical*, 151:281–290, 2010.

[10] Alexander Zybin, Y Kuritsyn, E. Gurevich, V. Temchura, K. überla, and K. Niemax. Real-time Detection of Single Immobilized Nanoparticles by Surface Plasmon Resonance Imaging. *Plasmonics*, 5:31–35, 2010.

Subproject B3
Data Mining on Sensor Data of Automated Processes

Jochen Deuse Katharina Morik

Leveling of Low Volume and High Mix Production

Fabian Bohnen
Lehrstuhl für Arbeits- und Produktionssysteme
Technische Universität Dortmund
fabian.bohnen@tu-dortmund.de

The application of conventional leveling approaches is limited to large scale production. This paper presents a PhD project which aims at developing a methodology for leveling of low volume and high mix production. It uses clustering techniques to group product types into product families considering manufacturing similarities. After that, a family-based leveling pattern is generated which describes a repetitive sequence of capacity slots for each family.

1 Introduction

Production leveling also referred to as production smoothing or heijunka is an essential element of the Toyota Production System and lean production respectively. The objective of production leveling is to balance production volume as well as production mix by decoupling production orders and customer demand [6]. Thus, unevenness in form of variation in the production schedule is reduced. This goes along with a decrease in overburden and waste. Leveling enables production to meet the customer demand without holding large volumes of inventory or spare capacities. Concurrently, it can be used to keep inventories to a controlled standard, diminishes or ideally avoids the bullwhip-effect in the value stream and shortens lead times. Leveling distributes production volume and mix to equable short periods. The sequence of these periods describes a kind of manufacturing frequency. According to this leveling pattern every product type is manufactured within a periodic interval. Thus, workload in production and logistic processes is balanced. Production leveling typically requires limited product diversity combined with a stable and predictable demand [4]. Due to that, the application of conventional leveling approaches (i.e. manufacturing every product type within a periodic interval) is limited to large scale production. For this scope of application, procedure models with different levels of detail can be found in literature (c.f. e.g. [10] or [9]). Wuthnow, for example, describes a systematic procedure to level the production of control units in the automotive industry and develops different production control methods [12].

Additionally, literature in the field of production leveling focuses on so called level scheduling approaches. These approaches aim at solving the so called production smoothing problem (PSP) which describes the problem of minimizing the variation of production rates and workload [3], [13]. In this context two alternative approaches are differentiated in literature, namely the single-level PSP and the multi-level PSP. The single-level PSP only considers production rates and workload at one production level (e. g. the assembly level) [7]. In contrast to that, the multi-level PSP includes production rates and workload at more than just one production stage [8]. Most of the approaches dealing with the PSP focus on flow shops (i.e. large scale production) in form of synchronized assembly lines in the automotive industry [13]. Considering these facts, the research objective of the PhD project presented in this paper is to develop a methodology for leveling of low volume and high mix production. In the following chapter the paper focuses on this methodology and the work conducted in this context. Based on that, future research work is discussed in chapter three and a conclusion is given in chapter four.

2 Methodology and Conducted Work

The methodology for leveling of low volume and high mix production is based on the systematic procedure for leveling developed in context of IGF research project 15865/N. Due to that, it is composed of two essential steps which are described in the following sections.

In the first step clustering techniques are used to group product types into families according to their manufacturing similarity. Utilization of these families for leveling requires that product types of one family can be manufactured one after another without or with minimal losses caused by changeover. Because of this, manufacturing oriented grouping criteria, especially criteria specifying similarities concerning production sequence and requirements are used [2]. For family formation different types of clustering algorithms are used which deliver different grouping results. These results are compared and validated using a so called desirability index (c.f. [11]). Up to now, this approach for product family formation for leveling has been validated with a real data set of about 300 product types.

Based on the formed families production leveling is realized in the second step by creating a family-oriented leveling pattern. Up to now, this pattern only aims at leveling of workload of a single maschine at one production level. Due to that, the methodology in this form refers to the single-level PSP. The leveling pattern describes a repetitive sequence of capacity slots for each family. The sequence of these slots is determined considering changeover times between the families. The problem of finding such a sequence can be transferred to the so called traveling salesman problem (TSP). The TSP describes the problem of creating a shortest salesman tour through a given number of cities [5]. After that the length of each capacity slot is determined by generating a start pattern which once contains each family and optimizing this pattern in the next step. In this context large capacity slots are divided and placed at different positions in the leveling pattern [1]. Both the TSP and the pattern optimization are problems in combinatorial optimization. Up to now, the methodology includes relatively simple greedy heuristics to solve these problems. Although these heuristics do not ensure that the best possible solutions are found, the methodology in its existing form has been validated successfully by using real data.

3 Further Research Work

Considering the methodology and the conducted work presented in the preceding chapter, future research work will be necessary regarding product family formation as well as leveling pattern creation.

In context of product family formation the utilized clustering method has to be analyzed especially concerning its ability for clustering of large data sets. Concerning that, the future research work will be related to the research work within Collaborative Research Center 876. Beyond that, the analysis of customer demands is an important aspect which has to be implemented in the methodology to generate reliable forecasts as input data for pattern creation. Due to the fact that these demands represent a form of time series, the future work on this topic will also be related to the work within Collaborative Research Center 876. This will especially affect the work in context of project B3 which aims at analyzing complex time series in form of sensor data.

In context of leveling pattern creation future research work will focus on transferring the multi-level PSP described in literature to the special application of leveling low volume and high mix production. After that exact and heuristic solution approaches will be analyzed and if necessary an adapted solution approach will be developed.

4 Conclusion

Up to now, the application of production leveling is limited to large scale production. The methodology which will be developed in context of the presented PhD projekt will point out how leveling can be implemented in low volume and high mix production to reduce waste in both production and logistic processes. Due to that, the PhD project presented in this paper focuses on a topic which is highly relevant for industrial application.

References

[1] F. Bohnen, M. Buhl, and J. Deuse. Systematic procedure for leveling of low volume and high mix production. In *Proceedings of the 44th CIRP Conference on Manufacturing Systems, 31.05.-03.06.2011, Madison, WI, USA*.

[2] F. Bohnen and J. Deuse. Leveling of low volume and high mix production based on a group technology approach. In *Proceedings of the 43rd CIRP International Conference on Manufacturing Systems 26.-28.5.2010, Vienna, Austria*, pages 949–956.

[3] N. Boysen, M. Fliedner, and A. Scholl. Sequencing mixed-model assembly lines: Survey, classification and model critique. *European Journal of Operational Research*, 192(2):349 – 373, 2009.

[4] A. Hüttmeir, S. de Treville, A. van Ackere, L. Monnier, and J. Prenninger. Trading off between heijunka and just-in-sequence. *International Journal of Production Economics*, 118(2):501–507, 2009.

[5] G. Laporte. A concise guide to the traveling salesman problem. *Journal of the Operational Research Society*, 61(1):35–40, 2010.

[6] J. K. Liker. *The Toyota Way*. McGraw-Hill, New York, 2004.

[7] J. Miltenburg. Level schedules for mixed-model assembly lines in just-in-time production systems. *Management Science*, 35(2):192–207, 1989.

[8] J. Miltenburg and G. Sinnamon. Scheduling mixed-model multi-level just-in-time production systems. *International Journal of Production Research*, 27(9):1487–1509, 1989.

[9] M. Rother, R. Harris, and J. Womack. *Kontinuierliche Fließfertigung organisieren*. Lean Management Institut, Aachen, 2004.

[10] A. Smalley. *Produktionssysteme glätten*. Lean Management Institute, Aachen, 2005.

[11] C. Weihs and G. Szepannek. Distances in classification. In P. Perner, editor, *ICDM*, volume 5633 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2009.

[12] A. Wuthnow. *Steuerung und Nivellierung von Wertströmen in der Automobilsteuergerätefertigung*, volume 3 of *Industrial Engineering*. Shaker, Aachen, 2010.

[13] M. Yavuz. An iterated beam search algorithm for the multi-level production smoothing problem with workload smoothing goal. *International Journal of Production Research*, 48(20):6189–6202, 2010.

SFB 876 - Project B3: Data Mining on Sensor Data of Automated Processes

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

Aiming at data mining on sensor data, it is the preliminary task of project B3 to enable the process analyzed to record and deliver data in an appropriate quality and quantity. This techreport outlines the necessary steps to qualify the process as well as to analyze and select the data-streams relevant.

1 Introduction

Within the Collaborative Research Center 876 project B3 targets the time-constrained analysis of sensor data recorded in automated interlinked processes. The real-time analysis of recorded data will yield information on the quality of the product currently processed at every monitored process-station. Firstly, identifying such defects as soon as possible offers the opportunity to eject the current product from the process-chain, avoiding production costs for the defective product at the remaining process-stations, reworking or rejection costs. In a second step online surveillance and evaluation of process-data allows for adaptations in the process parameters of subsequent process stations that will correct the effects of preceding faulty production steps and by this prevent the product from being rendered useless. Especially in production processes in which product quality cannot be monitored or tested directly at the product itself the described procedure is the only opportunity that assesses the quality produced [1] [3] [4].

This paper is organized as follows: In chapter two the work conducted during the first nine month is presented, focussing on installing the required IT-infrastructure and describing relevant parameters for the quality analysis. The third chapter discusses future research work and new research topics influenced by SFB876. Finally, a brief conclusion is given.

2 Conducted Work

The case-study analyzed in project B3 is a rolling mill for high quality long products of a major German steel producer. The conducted work, which is described in the following, is based on this case-study. Firstly, the installation of the IT-infrastructure necessary and secondly the work on determining relevant quality parameters and decision points in the process chain is presented.

2.1 IT-Infrastructure

In order to enable the entire process chain to collect, record and deliver the data of interest, the given IT-hardware infrastructure needed to be adapted. As of now, all relevant processes except for one station are equipped with process monitoring and data recording systems, which allow for collecting various digital as well as analog signals at a sampling rate of up to 100Hz. Currently, all process-data is transmitted in real-time to a centralized server, where it is stored. Based on the analysis to be conducted, the relevant signal-data is selected and preprocessed. In a last step, the data is converted from the recording system-specific file format into a csv-file and finally imported into a database for analysis. The data collected depends on the type of production process it derives from. In general the prototype process studied in this project consists of three distinct production technologies: hearth furnaces, rolling mills and separation machines.

Beyond that, relevant data is collected from ultrasonic quality tests. In comparison to the afore mentioned, reliable quality data becomes available with a delay of two to 14 days, depending on the product's heat-treatment. This type of data is complicated to integrate into the process-data database as it is collected manually which results in incomplete and sometimes not standardized information. In addition, it is stored in a different data type requiring a conversion. But as quality data is a prerequisite for training the prediction models, the conversion and standardization has to be performed regardless of the effort.

Backtracking the data that is recorded in the process-chain in order to link specific data packages to a physical product posed another challenge to be met. It was solved by assigning a unique ID to each physical product moving through the process-chain. This ID-number is included in the techno-string and by this linked to the recorded data.

The data types and IT-infrastructure described so far form the data basis on which data mining analysis can be conducted. However, as not all data types available at each station are relevant for quality predictions, a selection of parameters is a prerequisite.

2.2 Parameter and Decision Point Definition

Beyond enabling the process to record the data required, relevant data has to be selected and possible decision points within the process chain must be determined. The relevant data results from critical forming parameters (c.f. [5], [6]). For the first analysis the following data is recorded:

At the hearth furnace the different zone temperatures [°C] as well as the time spent at each zone [s] and the overall cycle time [s] is recorded. Additionally, the computed temperature at the steel block's core, top and bottom [°C] is written into the data base. At this point it is necessary to point out that due to technological restrictions resulting from the extreme temperatures in the hearth furnace, the actual material temperature cannot be measured. Therefore, computed values are recorded. The block-roll and finishing roll collect data of target and actual roll adjustment [mm], force applied [to], grooves used, forming temperature [°C] and rolling velocity [m/s].

Besides the definition of the relevant data for quality properties, identifying possible decision points in the process chain is a prerequisite for successful and efficient process control. The first decision point defined and analyzed is located directly after the hearth-furnace. At this point the recorded heating parameters can be analyzed in order to ensure a correct heating procedures for each block. Even though this process is believed to be critical for the entire process, first results show no obvious correlation between heating procedures and quality properties [7].

3 Further Research Work and Derived Research-Topics

In the upcoming research period, it will be necessary to introduce additional decision points into the analysis. The one analyzed next is located behind the block-roll. At this point in the process chain two major impact factors can be analyzed: Firstly, the surface temperature of the steel block can be verified by the means of a pyrometer installed in front of the block-roll. The core temperature cannot be measured, but it directly translates into the process parameters recorded at the block roll. An erroneous heating process may result in striking results when forming forces are analyzed. Further decision points for future analysis are at the finishing rolls and at the ultrasonic quality test facilities. Besides widening the focus step-wise on the entire process-chain, it will be necessary to work on the database structure to include additional quality-relevant data and to pre-process incoming data, especially the data gained from ultrasonic quality test.

From the research work conducted in the collaborative research center so far, several additional research interests were derived. One of which is to employ data mining techniques in the field of line balancing and sequencing in the context of mixed-model assembly lines [2]. This approach intends to reduce complexity in line balancing by analyzing the different products and their variants in order to determine master-variants representing a product family. Based on these product families, line balancing will be conducted to generate the idle-time minimal line-setup sequence. Within each line-setup the corresponding tasks may be scheduled according to a least over-work rule. The advantage of this optimization procedure is the reduced computational effort for balancing and sequencing which results from the pre-computation of the product families.

4 Conclusion

In the first nine month of project B3 it was the main task to enable the case-study process-chain to collect and record the data in a way it can be used in this project. Nonetheless, first analysis were conducted on the data gained, leading to first insights on parameter relevance. As lined out above the research work led to new approaches in balancing and sequencing assembly lines employing data mining techniques.

References

[1] B. Bugayev, Y. Konovalov, Y. Bychkov, and E. Tretyakov. *Iron and Steel Production*. Books for Business, 2001.

[2] J. Deuse, F. Bohnen, and B. Konrad. Renaissance der gruppentechnologie. *Zeitung für wirtschaftlichen Fabrikbetrieb*, 106(5):337–341, 2011.

[3] P. Figueiredo. *Technological learning and competitive performance*. Edward Elgar, 2001.

[4] A. Kugi, W. Haas, K. Schlacher, K. Aistleitner, H. Frank, and G. Rigler. Active compensation of roll eccentricity in rolling mills. *IEEE Transaction on Industry Applications*, 36(2):625–632, 2000.

[5] K. Lange, editor. *Umformtechnik: Grundlagen*, chapter Fließkurven, Fließortkurven und Formänderungsvermögen, pages 92–138. Springer, 2. edition, 2002.

[6] K. Lange, editor. *Umformtechnik: Grundlagen*, chapter Plastizitätstheoretische Grundlagen, pages 139–236. Springer, 2. edition, 2002.

[7] M. Stolpe, K. Morik, B. Konrad, D. Lieber, and J. Deuse. Challenges for data mining on sensor data of interlinked processes. Next Generation Data Mining Summit 2011: Ubiquitous Knowledge Discovery for Energy Management in Smart Grids and Intelligent Machine-to-Machine (M2M) Telematics, 2011.

Learning from Label Proportions on Distributed Sensor Data

Marco Stolpe
Department of Computer Science
Artificial Intelligence Group, LS 8
TU Dortmund University
marco.stolpe@tu-dortmund.de

In factories, distributed sensors can monitor the processing of individual products. The quality, however, can sometimes be assessed only at the end of the process. Prediction models could be devised that can detect potential errors earlier, in real-time, based on the current and previous sensor measurements. However, it is sometimes costly to track the errors of individual products. In such cases, only the proportions of errors for sets of products are known. Only few algorithms exist yet for this new kind of problem, the learning from label proportions. We have developed the LLP algorithm, which, on several standard data sets, shows better prediction performance than state-of-the-art methods and has a lower training and application time.

1 Introduction

In interlinked production processes, the physical quality of the final product often can only be assessed at the end of the process. However, based on data on how a product is processed, the quality of the final product could potentially be predicted before it reaches the final processing station and quality control, saving resources. Based on sensor measurements of the current and previous stations, supervised learning methods could train models that predict the quality of the final product in real-time and detect errors as early as possible. Moreover, the already distributed nature of sensors could be exploited by a decentralized analysis, avoiding the expensive transferal of all data to a central server. However, processes like in the rolling mill of a German steel producer pose several challenges to data mining [5]. Contrary to the familiar supervised learning scenario, labels aren't given for individual steel blocks, but for charges of blocks. Moreover, not many decentralized algorithms exist which can learn from observations whose attributes are vertically partitioned across several network nodes. For the new problem of learning from label proportions, we have developed the LLP (Learning from Label Proportions) algorithm [4], which is shortly described in the next section. It follows a discussion on how parts of the algorithm could be improved and distributed in the future.

2 Learning from Label Proportions

Given is a set X of n unlabeled observations with features $X_1 \times \ldots \times X_m$, drawn i.i.d. from an unknown probability distribution $P(X, Y)$. Y is a set of l categorical class labels, however, these labels are hidden. The set X is partitioned into h groups $G_i \subseteq X$ for which only the proportions $\pi_{ij} \in [0, 1]$ of labels y_j are known. From this information, a prediction function $g : X \to Y$ for individual observations should be learned, such that the expected loss $\mathbb{E}_P[L(Y, g(X))]$ is minimized. In the rolling mill case study, the observations correspond to the sensor measurements describing the processing of individual steel blocks. The groups correspond to customer orders. For each order, it is known how many blocks had a particular type of defect (if any).

The π_{ij} can be written as a matrix $\Pi = (\pi_{ij})$. The number of labels which result from applying $g(X)$ to all observations can be counted for each group, leading to a model-based label proportion matrix $\Gamma_g = (\gamma^g_{ij})$. The deviation of these two matrices can be measured by the average mean squared error $\mathrm{Err}_{MSE}(\Pi, \Gamma_g) = \frac{1}{hl} \sum_{i=1}^{h} \sum_{j=1}^{l} (\pi_{ij} - \gamma^g_{ij})^2$ over all matrix entries.

Let $\mathcal{C} = C_1, \ldots, C_n$ be a partitioning (clustering) of X that minimizes some quality criterion $q(\mathcal{C})$, as it could result from applying some partitional clustering algorithm. Under the assumption that the classes form clusters in the original space, the only problem left is to find out which cluster corresponds to which class. Let $\vec{\lambda}_{\mathcal{C}} = (\lambda_1, \ldots, \lambda_k), \lambda_i \in Y$ be a labeling of the clusters $\mathcal{C}_i$. The optimal labeling minimizes the difference between $\Gamma_{m_{\vec{\lambda}_{\mathcal{C}}}}$ and Π, where prediction model m maps each example to its corresponding cluster and assigns the cluster label λ to it:

$$\min_{\vec{\lambda}_{\mathcal{C}}} \mathrm{Err}_{MSE}(\Pi, \Gamma_{m_{\vec{\lambda}_{\mathcal{C}}}})$$

Since the number of clusters k is usually small, it is feasible to try out all different combinations of labelings exhaustively. In addition, also a greedy labeling strategy has been implemented.

To account for cases where the classes don't form clusters in the original space, the input space could be transformed. For example, with similarity based clustering methods, a weighting $\vec{w}$ of the features can be respected by using the weighted Euclidean distance $d_w(\vec{p}, \vec{q}) = \sqrt{\sum_i^m w_i(p_i - q_i)^2}, p, q \in \mathbb{R}^m$. The vector $\vec{w}$ whose associated clustering maximizes $q_{\vec{w}}$ and whose labeling minimizes Err_{MSE} can be considered optimal:

$$\min_{\vec{w}} \mathrm{Err}_{MSE}(\Pi, \Gamma_{m_{\vec{\lambda}^*_{\mathcal{C}}}}), \quad \vec{\lambda}^*_{\mathcal{C}} = \operatorname*{argmin}_{\vec{\lambda}_{\mathcal{C}^*}} \mathrm{Err}_{MSE}(\Pi, \Gamma_{m_{\vec{\lambda}_{\mathcal{C}^*}}}), \quad \mathcal{C}^* = \operatorname*{argmax}_{\mathcal{C}} q_(\mathcal{C})$$

The LLP algorithm solves this optimization problem by an evolutionary strategy. For each weight vector in a current population, X is partitioned by a clustering algorithm like k-Means with the weighted Euclidean distance. Then, one of the mentioned labeling strategies is applied to all clusterings and the fitness is evaluated according to the Err_{MSE}. The weight vectors then take part in a tournament selection, the values of parent weight vectors are passed on randomly to some newly generated children and their weight values are mutated randomly. LLP iterates until convergence or a user specified maximum number of generations. It can be shown that using k-Means with a bounded number of iterations, the total running time is only linear in the number of observations. Existing and new observations can be labeled by assigning the label of the closest cluster centroid.

For several group sizes and parameters, the prediction accuracy of LLP with the centroid model has been evaluated on eight different UCI data sets and compared to three state-of-the-art kernel methods for learning from label proportions. The evaluation was done by a 10-fold cross validation. For training, the observations were sampled uniformly into groups, the proportions calculated and the individual labels removed. For testing, the prediction accuracy was assessed on labeled observations. LLP with the centroid model had a prediction accuracy comparable to the existing methods, while the training took only linear time. When training additional classifiers like decision trees, random forests, kNN, Naive Bayes and the SVM with linear and RBF kernels on the observations as labeled by LLP, the best models achieved the highest average rank over all data sets for several group sizes, being significant in several cases. Also, none of the existing methods has *all* the properties of LLP: a good prediction performance, linear running-time, the ability to predict multiple classes and accounting for additional labeled examples, if available. Moreover, LLP is quite general, allowing for the easy incorporation of other input space transformations, partitional clustering algorithms, labeling strategies and performance measures.

3 Future Work

The exhaustive labeling strategy perfectly minimizes the Err_{MSE}, but it sometimes misses cluster labelings with a high accuracy. It should be tested if the stability of the labeling in terms of accuracy can be improved by doing a group-wise cross-validation, where the Err_{MSE} of a labeling is tested on a left-out subset of the groups. AOC-KK incorporates the Err_{MSE} directly into the optimization problem of kernel k-Means. The stable performance of LLP in comparison to AOC-KK could perhaps be explained by doing the clustering and labeling in subsequent steps, with the clustering acting as some kind of regularizer and the number of clusters k controlling the bias/variance trade-off. These questions should be further investigated, maybe by incorporating the Err_{MSE} criterion into the EM clustering algorithm and comparing its behavior and performance to LLP.

The one-class SVM is an unsupervised kernel method which can estimate the support of a high-dimensional distribution. Support vector clustering (SVC) [1] builds on a trained one-class SVM model and interprets the contours in the original space as clusters. As such, it could also be used with LLP. In comparison to k-Means, SVC has the advantage that the clusters can be arbitrarily shaped and it can also detect noise. Moreover, a global one-class SVM model across vertically distributed data can be calculated by sampling (see Das et al. [2]). The question is if and how the subsequent SVC step can also be distributed. The original SVC algorithm has a running time of $O(n^2)$. Lee and Lee [3] have proven that it suffices to only cluster equilibrium points, whose number is even lower than the number of support vectors. Thereby, they could improve the accuracy of the clustering and achieve an almost linear running time of the clustering step. It needs to be investigated if their method could also help with the distribution of SVC.

References

[1] Asa Ben-Hur, David Horn, H.T. Siegelmann, and Vladimir Vapnik. Support vector clustering. *The Journal of Machine Learning Research*, 2:125–137, 2002.

[2] K. Das, K. Bhaduri, and P. Votava. Distributed Anomaly Detection using 1-class SVM for Vertically Partitioned Data. *Stat. Analysis and Data Mining Journal*, 4:393–406, August 2011.

[3] J. Lee and D. Lee. Dynamic characterization of cluster structures for robust and inductive support vector clustering. *IEEE transact. on pattern analysis and machine intel.*, 28(11):1869–74, November 2006.

[4] M. Stolpe and K. Morik. Learning from label proportions by optimizing cluster model selection. In *Proc. of the 2011 Europ. conf. on Machine learning and knowledge discovery in databases - Vol. Part III*, pages 349–364, Berlin, Heidelberg, 2011. Springer.

[5] M. Stolpe, K. Morik, B. Konrad, D. Lieber, and J. Deuse. Challenges for data mining on sensor data of interlinked processes. In *Proc. of the NGDM 2011: Ubiquitous Knowlegde Discovery for Energy Management in Smart Grids and Intelligent Machine-to-Machine (M2M) Telematics*, 2011.

Mining Large Scale Data Sets using Online Learning

Christian Bockermann
Informatik Lehrstuhl 8
Technische Universität Dortmund
christian.bockermann@cs.uni-dortmund.de

The continuous growth of data available for training classifiers requires scalable methods for processing and learning from that data. Recently, stream mining or online algorithms have become a popular approach for learning on large data sets. This report documents the implementation of our stream-mining framework to handle data at a large scale. We propose the combination of online learning algorithms with the parallelization approaches of the *map-and-reduce* paradigm.

The objective of the Collaborative Research Center 876 is to investigate methods for analyzing large data sets under resource constaints. Considering the machine learning area, most of today's learning methods are *batch* algorithms that require multiple passes over the data. Often, this requires a considerable amount of data to be present in main memory which calls for new approaches handling that data with bounded resources. For datasets like the complete MNIST handwriting data[1], this renders traditional methods useless. Other data sources like click-stream data, electricity data or (system) log data generate likewise data volumes that need to be handled.

Streaming- or online-methods try to overcome this problem by providing approximate solutions using a single pass over the data set. Examples for such online algorithms have been proposed for various machine learning tasks such as *frequent itemset mining* [4, 2, 6, 7], classification [3, 8], regression and clustering.

In order to make these methods available in an easy to use way, we designed a flexible stream mining framework that implements various stream mining algorithms. Moreover,

[1] Here, we refer to the MNIST extended handwriting dataset, which contains 8 million datapoints of hand written characters. The data in SVM light format is about 17 GB in size. The dataset is available at `http://leon.bottou.org/papers/loosli-canu-bottou-2006`

we extended this framework to allow for a scaling to multi processor machines by following the *map-and-reduce* paradigm, which recently has become a major model for parallel computing on large sets of processors. By following a stream-oriented approach we seek to make the algorithms compatible to run on a single multi-processor machine as well as cluster setups provided by the *Hadoop* Map-and-Reduce framework[2].

1 Objectives and Considerations

Large data sets incur in various data generating processes, which we can divie into two different cases: (a) the setting of a *static data set* of large volume and (b) the *continuous generation of data* by some source. The former resembles the typical case of large sets that need to be processed with limitted resources, whereas the latter describes the setting of measurements a typically fixed set of features being generated by some sensor.

To easily incorporate the stream mining framework into the *RapidMiner* software, we focused on a close mapping of the data representation to *RapidMiner*'s `Example` class.

The objectives of the proposed framework are:

- Provide a clean abstraction layer for implementing online learning algorithms
- Implement a runtime environment for online-learning that allows for processing of large scale data sets or continuous data streams
- Provide mechanisms to empirically evaluate the algorithms with regard to their *error rate* and *memory usage*
- Allow for the integration of the framework into the *RapidMiner* software.

2 Online Algorithms implemented

The stream framework provides implementations of various online learning algorithms, which focuses on the case of *continuous data streams*. Below we outline the list of available algorithms.

Counting Elements in Continuous Streams

The task of counting the frequencies of elements or determining the *top-k* most frequent elements within an infinite continuous data stream is non-trivial, if the memory is limitted to a constant size.

The online counting algorithms implemented within the stream framework are *Lossy-Counting* and *Sticky-Sampling* [6], *Count-Min-Sketch* [2] and *Space-Saving-TopK* [7].

Algorithms for Computing Quantiles on Continuous Streams

A φ-quantile is a statistical measure that can be indicative for characterizing data sets.

[2] Hadoop is an open-source Java implementation of Map-and-Reduce, `http://hadoop.apache.org`.

Answering quantile queries on a continuous single-feature data stream precisely requires a considerable amount of memory, i.e. the frequencies of all encountered values.

Several online approximations for that task have been proposed. The following algorithms have been implemented within the stream framework: *Greenwald-Khanna* [5], *Window-Sketch Quantiles* [1] and *Ensemble-Quantiles*.

Classifiers for Data Streams

For classification tasks, the stream framework provides several online learners such as *Naive Bayes*, *Perceptron*, *Neural Networks*, *Very-Fast-Decision-Trees* [3], *Stochastic Gradient Descent* [8]. Based on different counting algorithms we also implemented several Naive Bayes versions for nominal features, such as Naive-Bayes with Lossy-Counting or Sticky-Sampling.

3 Large Scale Data Processing using Map&Reduce

To speed up the data processing on *large scale static data sets* the framework provides a simple abstraction of the *map-and-reduce* paradigm. This partitions a large data set into a disjoint set of finite streams. Algorithms providing models that allow to be merged can thus easily be parallelized with almost no changes. The *map-and-reduce* layer of the streams framework targets to be run on single multi-processor machines as well as being deployed on a multi-node Hadoop cluster by being compatible to the Hadoop *streaming API*. Figure 1 outlines the *map-and-reduce* paradigm.

Based on this abstraction layer we implemented a multi-threaded approach for training an SVM using multiple instances of the *Stochastic Gradient Descent* implementation.

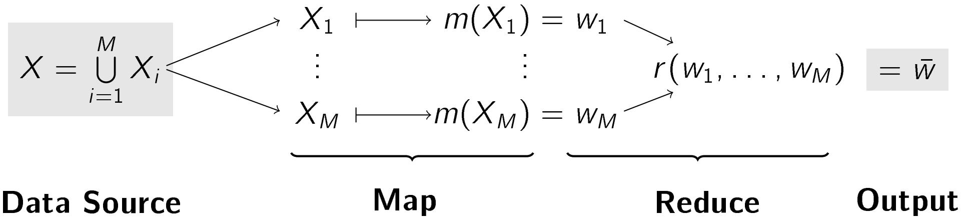

$X = \bigcup_{i=1}^{M} X_i$
$X_1 \longmapsto m(X_1) = w_1$
$\vdots$
$X_M \longmapsto m(X_M) = w_M$
$r(w_1, \ldots, w_M)$
$= \bar{w}$
Data Source
Map
Reduce
Output

Figure 1: Parallel stream-optimization with Map-and-Reduce to train an SGD classifier

Parallelized Training using Stochastic Gradient Descent

For training an SVM on large datasets, stochastic gradient descent (SGD) has been used as online optimization strategy. During optimization, the SGD algorithm passes over the data to adjust the prediction model in terms of the weight w and the intercept b. This single-threaded approach still requires a considerable training time. To make use of todays multi-processor architectures, we implemented a parallelized variant of SGD: We trained several SGD instances on disjoint blocks of the training data and then merged the resulting models. Each SGD is represented by a weight vector w_i and an intercept b_i. This training can be carried out in parallel as the *map* step. Finally we merge the w_i, b_i by computing the averages $\bar{w}$ and $\bar{b}_i$ in the *reduce* step. Using the map-and-reduce approach we have been able to improve the training time on the complete 8 million examples to below 10 minutes. We tested our implementations on a single multi-processor system, limitting the process memory to 8 GB.

4 Future Work

To incorporate the benefits of the stream framework into the *RapidMiner* tool, we started to implement a *StreamPlugin* that is continuously extended to include operators for all implemented online algorithms.

The first experiments on the parallelized SGD training showed promising results. Ongoing work is currently to run evaluations on more datasets and incorporate additional optimization strategies into the parallel SGD training, such as weighted averages based on the training error. We also performed data preprocessing of the traffic data of SFB876 project B4 using the *map-and-reduce* approach. As this also is a large scale data set, we will try to build a prediction model on that using parallelized online learning.

References

[1] A. Arasu and G. S. Manku. Approximate counts and quantiles over sliding windows. In *Proc. of the 23rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 286–296, New York, NY, USA, 2004. ACM.

[2] G. Cormode and S. Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. *J. Algorithms*, 55:29–38, 2004.

[3] P. Domingos and G. Hulten. Mining high-speed data streams. In *KDD '00: Proc. of the sixth ACM SIGKDD Int. Conf. conference on Knowledge discovery and data mining*, 2000.

[4] G. C. et al. Finding hierarchical heavy hitters in streaming data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(4):2, 2008.

[5] M. Greenwald and S. Khanna. Space-efficient online computation of quantile summaries. In *In SIGMOD*, pages 58–66, 2001.

[6] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In *VLDB*, pages 346–357, 2002.

[7] A. Metwally, D. Agrawal, and A. Abbadi. Efficient computation of frequent and top-k elements in data streams. In *Database Theory - ICDT 2005*, volume 3363 of *Lecture Notes in Computer Science*, pages 398–412. Springer Berlin / Heidelberg, 2005.

[8] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proc. of the 21st International Conf. on Machine Learning (ICML)*, 2004.

Subproject B4
Analysis and Communication for dynamic traffic prognosis

Michael Schreckenberg Christian Wietfeld

Performance of LTE for M2M Communication

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

The number of Machine to Machine (M2M) applications is increasing rapidly in cellular communication systems. Hence, the evaluation of the impact of M2M applications on common Human to Human (H2H) communication poses a huge challenge. By means of a Markovian model, which is parameterized by laboratory measurements and ray tracing simulations, an estimation of the behavior of Long Term Evolution (LTE) (regarding the blocking probability) for different traffic environments is proposed. The results show that particularly many devices with a very small amount of data influence the utilization of a LTE cell very intensely.

1 Motivation

Orthogonal Frequency-Division Multiple Access (OFDMA)-based cellular communications networks are typically designed for high data rates. However, new applications in the area of M2M communications can be found in these systems [2]. One example is the area of traffic estimation, where devices in cars collect sensor information about the traffic situation and transmit them to a server. Furthermore, a download of relevant information from a server, for example the current traffic situation, is needed. In a LTE cell with a radius of 5 km, up to 750 cars, which drive on a freeway, can be found. The behavior of M2M communication is different to H2H communication. Often M2M applications receive or transmit only a small amount of data or need a very low data rate. Hence, a framework for modeling M2M and H2H users in one cell is needed.

2 Markovian Model for Performance Evaluation of Cellular Networks

In order evaluate the behavior of a LTE cell, a tool chain consisting of a laboratory setup (LTE base station emulator, channel emulator for fast fading and User Equipment (UE), see [1]), ray tracing simulations and a Markovian model is used (see Fig. 1). Thereby, a typical urban outdoor scenario (*Vehicular A* channel model from the ITU) is assumed. For this scenario, we measured User Datagram Protocol (UDP) data rates for a single user dependent on the Signal to Noise plus Interference Ratio (SNIR) and the number of allocated Resource Blocks (RB) for one user. One RB is the smallest unit, which can be allocated to an LTE user. The distribution of the SNIR for the urban outdoor scenario is calculated by the ray tracing simulation.

In the next step, the behavior for many users in an LTE cell is modeled by a Markovian model. Assigning the user request to different classes of resources, each class is modeled by a dimension in the Markovian model. According to the reduction of dimension of Markovian models [3], a LTE cell with users with different Quality of Service (QoS) requirements and different SNIRs can be modeled. Thereby, the *j*th state denotes the allocation of *j* resource blocks from the LTE OFMDA signal.

We divide the users by the user requirements (H2H with a data rate of 1 Mbit/s; M2M with 10 kByte UDP data) and the SNIR. By means of the Markovian model, the blocking probability and the traffic intensity is evaluated for different user behaviors and transmission strategies in terms of the assignment of a different number of RBs.

3 Results of the Markovian Model

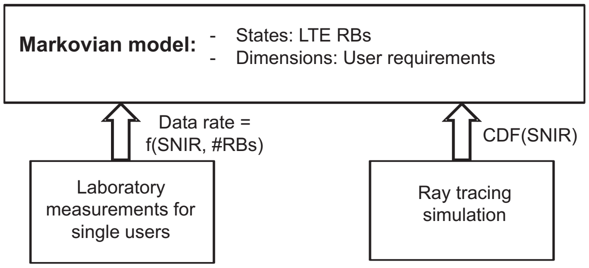
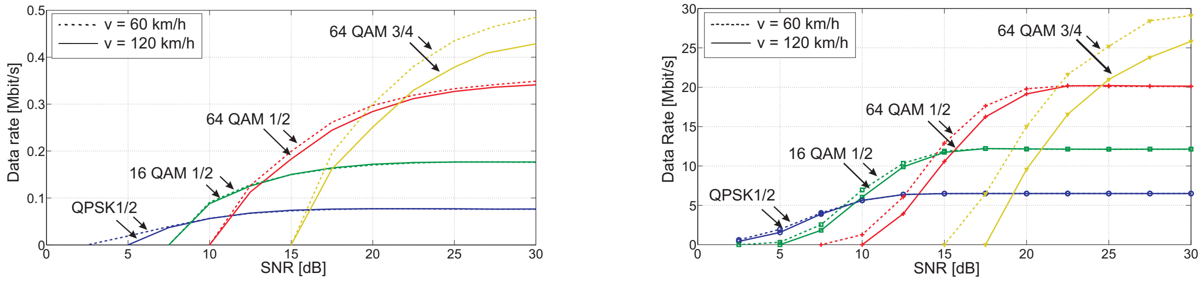

Markovian model:
- States: LTE RBs
- Dimensions: User requirements
Data rate = f(SNIR, #RBs)
CDF(SNIR)
Laboratory measurements for single users
Ray tracing simulation

Figure 1: Parameterized Markovian model

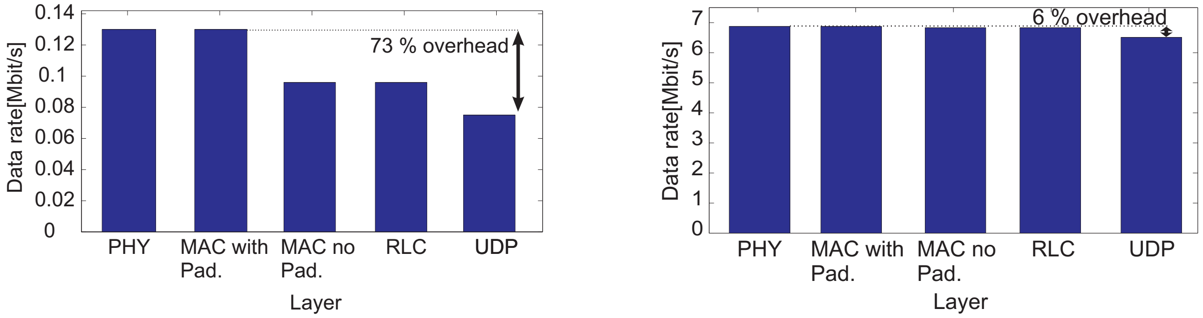
The UDP data rates vs. SNIR for 50 RBs per user and 1 RB per user for different Modulation and Coding Schemes (MCS) can be seen in Fig. 2. We assume that always the MCS with the highest data rate is chosen. For an SNIR of 20 dB a data rate of 19.5 Mbit/s for a 64 Quadrature Amplitude Modulation (QAM) with a code rate R = 1/2 and a velocity of 120 km/h can be achieved. This is a data rate of 0.39 Mbit/s per allocated RB. If one user allocates only 1 RB the data rate is only 0.28 MBit/s. The main reason for this effect is the relatively high overhead between Physical (PHY) layer and UDP layer for small data rates, because the Media Access Control (MAC) padding and the Protocol Data Unit (PDU) size in the Radio Link Control (RLC) layer is dependent on the incoming data rate. Therefore, the relatively overhead (see Fig. 3) for small data rates (1 RB; 73 % between PHY and UDP) is much higher than for high data rates (50 RBs; 6 % between PHY and UDP). Hence, we measured the data rate for all numbers of RBs per user and use these results as input for the Markovian model. Thereby, we identified that the impact of the SNIR is much stronger than the impact of the velocity (see. Fig. 2). Therefore, we differentiated the H2H and M2M users only by the SNIR. The signal receiver quality of the UEs is divided into three parts (1. SNIR $\leq$ 15 dB: represented by 10 dB SNIR. 2. 15 dB $<$ SNIR $\leq$ 25 dB: represented by 20 dB SNIR. 3. 25 dB $<$ SNIR: represented by 30 dB SNIR).



(a) 1 RBs per user

(b) 50 RBs per user

Figure 2: Measurements results for different MCS; Data rate vs. SNIR



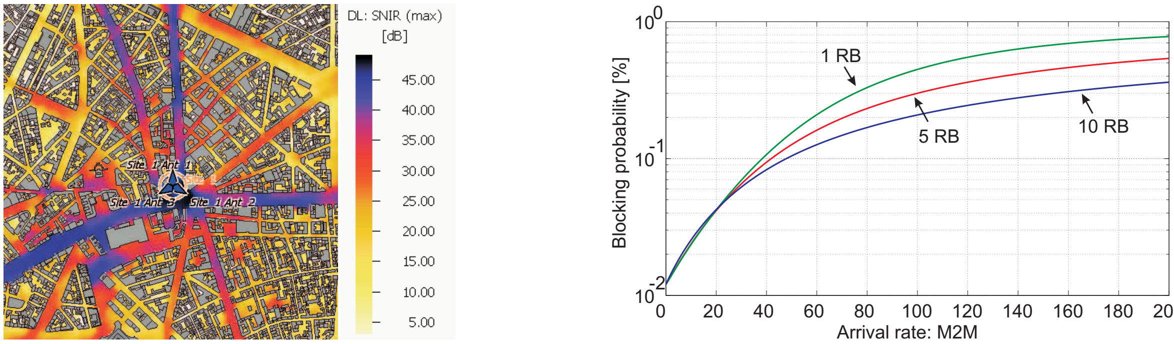
(a) 1 RB per user

(b) 50 RBs per user

Figure 3: Data data on different layers for values number of RBs; QPSK, MCS = 8

For a video streaming service (H2H) with 1 Mbit/s, 9, 3 or 2 RBs are needed dependent on the channel quality. The M2M users should download a 10 kByte file with an adjustable number of RBs. We assumed, that the users are homogeneously distributed in the urban scenario. The map for the SNIR from the ray tracing simulation is illustrated in Fig. 4a. The probability for an SNIR lower than 15 dB is 0.47, for an SNIR between 15 dB and 25 dB is 0.33 and for an SNIR higher than 25 dB is 0.20.

In Fig. 4b, the blocking probability of H2H users vs. arrival rate of M2M devices is



(a) Map for the SNIR in an urban environment

(b) Blocking probability vs. arrival rate of M2M devices for a different number of RBs per user

Figure 4: Ray tracing results and results from the Markovian model

shown for a different number of RBs per user. Hereby, a traffic for the video class of 3 is used. The blocking probability with 1 RB is worse than for 5 or 10 RBs per user due to the higher overhead. For an arrival rate of the M2M class from 50 Erlang and a transmission with 10 RBs, the blocking probability for the video class is 10 %. For the same assumtions, the blocking probability for the video class is 17 % if only 1 RB is used for the M2M class. Or rather, if the blocking probability should be smaller than 10 % (for QoS requirements) on average 38 M2M devices can be served per second when they transmit with 1 RB and on average 48 M2M devices when they transmit with 10 RBs. This is an enhancement of 26 %.

The Markovian model uses negative exponential distributions for the interarrival time of the arrival rate and the service rate. As next step, the influence of different distributions on the LTE performance should be observed. Furthermore, a direct connection between the ray tracing simulation and the channel emulator is planed. This makes an analysis of the LTE link in the laboratory with fast fading channels from die ray tracing simulation possible.

References

[1] C. Ide, B. Dusza and C. Wietfeld. *Mobile WiMAX Performance Measurements with Focus on Different QoS Targets*, accepted for presentation at 18th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN), Chapel Hill, North Carolina, USA, October 2011

[2] S.-Y. Lien, K.-C. Chen, *Massive Access Management for QoS Guarantees in 3GPP Machine-to-Machine Communications*, IEEE Communications Letters, vol. 15, no. 3, March 2011

[3] J. S. Kaufmann, *Blocking in a Shared Resource Environment*, IEEE Transactions on communications, vol. com-29, no. 10, October 1981

Lane-Specific Localization for Traffic Flow Prognosis

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

A lane-specific *Global Navigation Satellite System (GNSS)* positioning accuracy is required as input to improve sophisticated traffic flow modeling, but will be also useful for autonomous control of cars in future smart traffic environments. Within the first year of the SFB876 we investigate the accuracy of satellite-based positioning data in various system environments using a detailed system model, which includes all relevant impairments of the propagation channel as well as receiver behavior. Based on a multiscale simulation concept, which takes into account satellite mobility and constellations, satellite selection by the receiver using an evaluation of the respectively given satellite geometry and determination of *Time-of-Arrival (TOA)* impacted by shadowing and multipath propagation, the feasibility of a lane-specific positioning is investigated.

1 Simulative Position Accuracy Determination

State-of-the-art satellite simulations are widely used to estimate or validate minor to mayor modifications on running or future satellite systems. Several examples are known in literature in which satellite simulation facilities are of great value to analyze the performance of new features or the effect of additional hardware [3]. But all of those examples are analyzing a macroscopic problem. Hence simplifications like using empirical models are useful to minimize complexity and computing time for the simulations. On the other hand, those measures are missing the general validity for random local scenarios, and by that, also the applicability for lane-specific applications like real-time traffic predictions.

In order to determine the accuracy of actual satellite positioning systems and their combinations, a so-called *Simulative Positioning Accuracy Determination (SPAD)* was developed [1]. Based on the already published Multiscale Simulation Environment (MSE) [2] the SPAD, whose functionality is visualized in Figure 1, performs a realistic Time-of-Arrival (TOA) localization. In this case, the MSE is used to simulate the geospatial positions of the transmitter (satellite) and the receivers (mobile ground nodes) in OMNeT++. The satellite movement is also realistically included, using the appropriate NORAD Two-Line-Elements sets of the satellites, combined with the SGP4 algorithms [2].

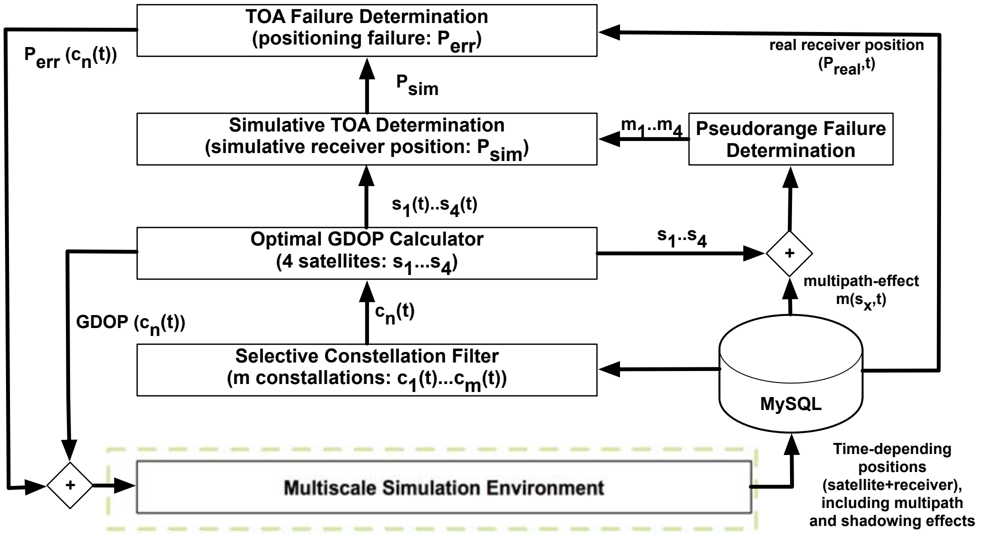


Figure 1: Visualization of the SPAD [1]

Afterwards, the so-called Selective Constellation Filter (SCF) creates different possible constellations ($c_1 ... c_m$), whereby every combination of satellites, just intra-system as well as inter-system ones is possible. Every single constellation c_n is analyzed to extract the optimal satellite geometry at a certain time and a given receiver position. By using the Geometric Dilution of Precision (GDOP) as a key performance indicator of the given constellation, the Optimal GDOP Calculator (OGC) analyzes iteratively every possible 4-satellite constellation of c_n to find the best-in-case GDOP value for the given time and receiver position. From this point on we just consider those four satellites ($s_1, ..., s_4$), because the MSE has already taken into account all shadowing and multipath effects, so that the extent satellites definitely contain the best possible four-satellite constellation. Then those $m_{i,err}$ are used, in addition to the known satellite positions $s_1 ... s_4$, to calculate the receiver's position P_{sim} using well-known trilateration approachs.The deviation of P_{sim} to the real position P_{real} is returned as P_{err} with the corresponding and already derived GDOP value, both depending on the used constellation $c_n(t)$ to the Simulation Control and Results Processing Unit of the MSE for visualization.

2 Satelllite Positioning Performance Evaluation

The main focus is a continuous analysis of satellite positioning accuracy, depending on the satellite position, the external conditions and the geospatial position of the receiver. In a first step, the comparison between the localization accuracy of different satellite systems like GPS, Galileo or a combination of both clarifies the variance in accuracy of those systems. Therefore, Figure 2 depicts a scatter plot of the simulated position estimation for a fixed point in a suburban environment. The measurement point is located in the point of origin. Furthermore, the plots contain the 95% quantile, which basically depict the expectable accuracy. It can be seen clearly by comparing the constellations of GPS, Galileo and Hybrid GPS/Galileo that the number of available satellites in orbit has a major impact on the achievable accuracy. In conclusion, the first result is that a combined usage of GNSS results in a higher positioning accuracy that reaches a lower bound of around 5m.

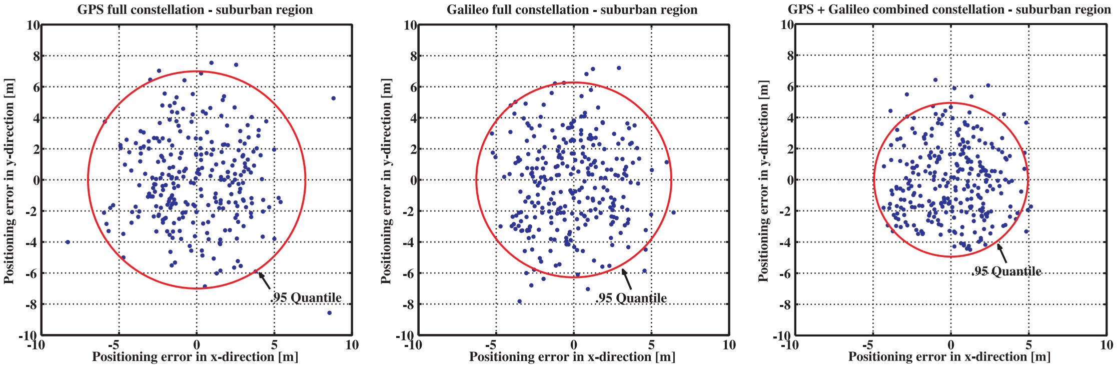


Figure 2: Simulated Position Error for GPS and Galileo Satellite Constellations in a Suburban Scenario for a Stationary Receiver [1]

Furthermore, it can be observed that the Galileo system will perform slightly better than GPS, which can be explained by the higher orbit of the satellites. It should be noted that the selected point of evaluation was static throughout the simulation. A more sophisticated evaluation of different positions belongs to our future works.

To substantiate the correctness of our simulation, Figure 3 left visualizes the coherence of satellites in range and the GDOP value corresponding to the developed framework. It is obvious, that both values depending on the respective actual system architecture are correlated to each other. In addition, the results of the simulation environment also have been validated with real-world measurements carried out with a GPS data logger, type V900 from Columbus. Referring to Figure 3 right, it can be clearly seen that the observed error distribution is very similar for both curves, which indicates that the simulation environment is working and is parameterized correctly. The measurement has been carried out for 20 minutes in the sub-urban region used for the corresponding evaluations explained above. The measurement period has been recalculated in the simulation environment with the equal GPS satellite constellation with the same *Two-Line-Element (TLE)* dataset and the date of the measurements to guarantee comparability.

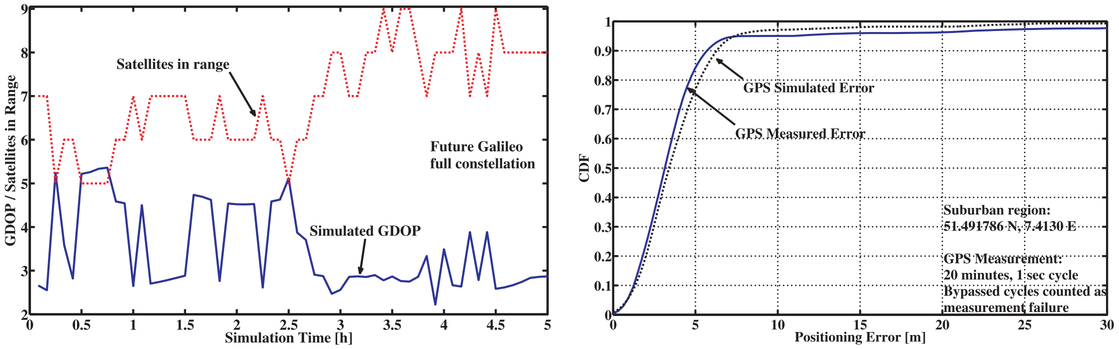


Figure 3: Comparison between geometric dilution of precision and satellite in range (left) and comparison between simulative and real-world localization accuracy (right) [1]

As a conclusion, it can be seen that the positioning accuracy is not sufficient for a lane specific positioning of cars for the traffic prognosis. Therefore, additional strategies are necessary to be able to detect the lane of the vehicle. A promising approach is data fusion mechanisms with real-time data from the vehicles CAN bus. This interface could be established via Bluetooth gateways and enables the evaluation of additional data like the flashing lights, steering angles and braking power. The first experiments have been successfully evaluated.

References

[1] Niehoefer, B., Lewandowski, A. and Wietfeld, C. , *Evaluation of the Localization Accuracy of Satellite Systems for Traffic Flow Predictions*, accepted for publication at Institute of Navigation - Global Navigation Satellite System (ION-GNSS), Technical Meeting, 2011, Portland, Oregon

[2] Lewandowski, A., Niehoefer, B. and Wietfeld, C., *Galileo Search-and-Rescue: Performance Aspects and new Service Capabilities*, International Journal on Satellite Communication and Networking, Special Issue: Emergency and Disaster Communication Systems Via Satellite, published online Oct. 2010

[3] Meng, X., Roberts, G. W., Dodson, A. H., Cosser, E., Barnes, J. and Rizos, C., *Impact of GPS satellite and pseudolite geometry on structural deformation monitoring: analytical and empirical studies*, Journal of Geodesy, 2004, Springer Berlin, Heidelberg

Traffic state reconstruction with "adaptive smoothing method" from incomplete traffic information

Timo Knaup
Physik von Transport und Verkehr
Universität Duisburg-Essen
knaup@ptt.uni-due.de

The recognition of traffic states and the prediction of traffic breakdowns on german highways is actually based on traffic data measured with stationary detectors, e.g. loop detectors. Today there are about 4500 loop detectors unregularly spreaded over nearly 2250 kilometer highway in North Rhine-Westphalia. Measurements contain average traffic speed (km/h), average traffic flow (veh/min) and proportions for cars and trucks, usually aggregated in one minute intervals. On basis of this data there are methods to reconstruct traffic conditions between the detectors. One is used in the actual work and is planned to be used for further improvements, especially the inclusion of floating-car-data.

1 Traffic state reconstruction

There are two possible methods to perform a reconstruction of spatio-temporal traffic patterns between locations of traffic measurements, e.g. loop-detectors. First the "adaptive smoothing method" presented in this report, second ASDA/FOTO introduced by Kerner in 1996-1998 [2]. The "adaptive smoothing method" is a twodimensional, spatio-temporal interpolation algorithm developed by Treiber and Helbing [4]. This method estimates a continuous mean velocity $V(x,t)$ as a spatio-temporal function from discrete velocity points $v_i = v(x_i, t_i)$, which be on hand at certain points x_i at certain times t_i.

It additionally filters out small-scale fluctuations and takes into account characteristics of information flow in different traffic situations. For detailed information, e.g. used formulae, look at [4] and [3]. The input values are given by stationary loop detectors, but also Floating-Car-Data can be a source. This is interesting for the B4 project.

Parameter	Value
Smoothing width in space coordinate (σ)	600 m
Smoothing width in time coordinate (τ)	60 s
Propagation velocity of pertubations in free traffic	70 km/h
Propagation velocity of pertubations in congested traffic	-15 km/h
Crossover from free to congested traffic (V_c)	60 km/h
Transition width between congested and free traffic (ΔV)	20 km/h

Table 1: Parameters used in calculations with the adaptive smoothing method.

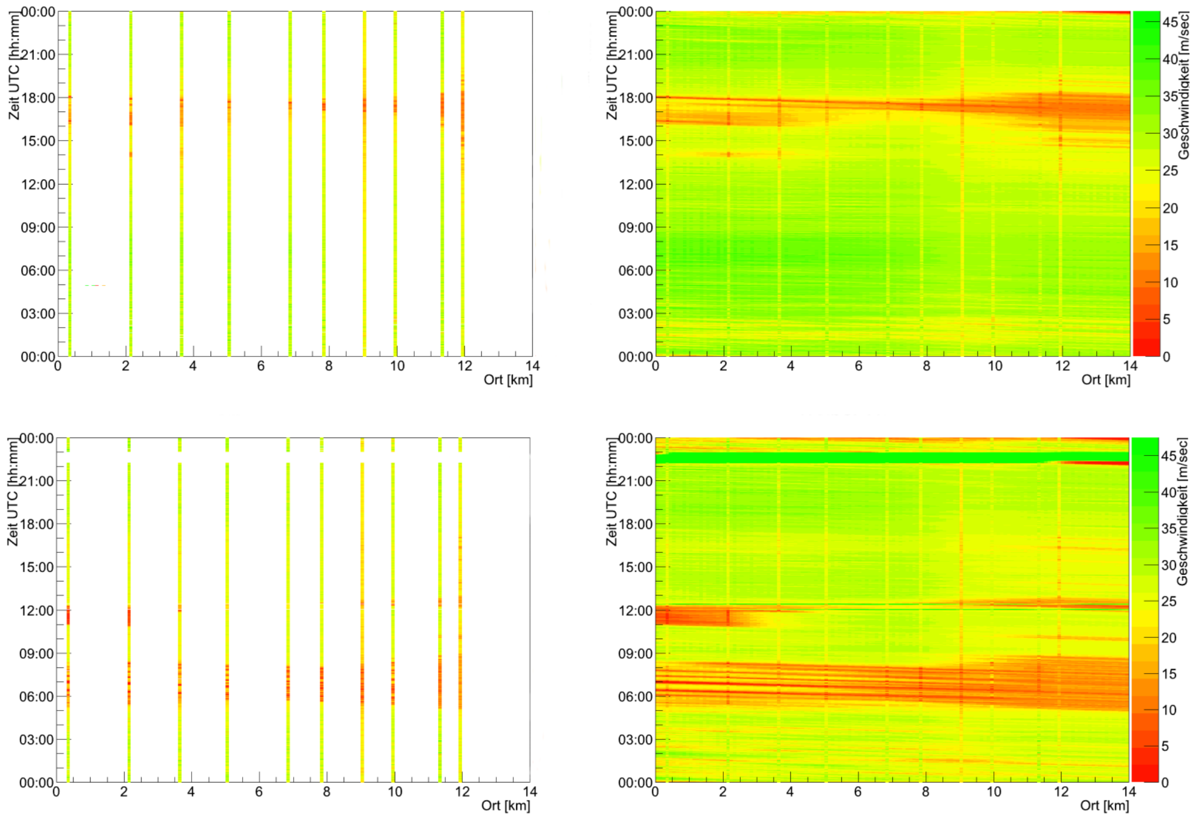


Figure 1: Reconstruction of spatio-temporal traffic patterns on the A44 at August 1 resp. August 8, 2010 by using the adaptive smoothing method. Green represents high whereas red represents low velocities.

In order to see, how the method works, we consider a 12 kilometer long section of the german highway A44 with 10 aperiodic spreaded loop detectors. This highway section is chosen as a test route for simulations and experimental drives in the B4 project. Table 1 shows parameters used in our calculation. They are taken from [3] and should yield good results.
Figure 1 shows the measured traffic data (left side) obtained from loop detectors aggregated over all lanes and the reconstructed traffic situation (right side) with the adaptive smoothing method for two different days in November 2010, more precisely November 8 and November 1, 2010. Some specific traffic patterns can be recognized, for example stop-and-go-waves in the region 0 km to 5 km between about 6:00 am and 9:00 am on November 8. As you can see the waves propagate against the traffic flow as assumed.
As mentioned above it is also possible to use Floating-Car-Data with this algorithm. The use of this kind of data is one target of the B4 project. Therefore we figured out, what kind of data is available and how it can be recorded.

2 Floating-Car-Data

In order to see what and how Floating-Car-Data can be obtained, we made a first experimental drive with two cars equipped with different data loggers which allow for recording of live data provided by vehicle internal sensors broadcasted via the vehicle internal CAN bus [1]. We recorded data from about 60 different kind of sensors like velocity, acceleration, status of brake light ten times each second. In both cars, GPS receiver were used as a second data source. The experimental drive was proceeded at August 10, 2011 in the morning hours on a highway sections around Dortmund. The detailed route is shown in figure 2. Both cars started with a 5 minute time delay at about 7:30h. Because of the

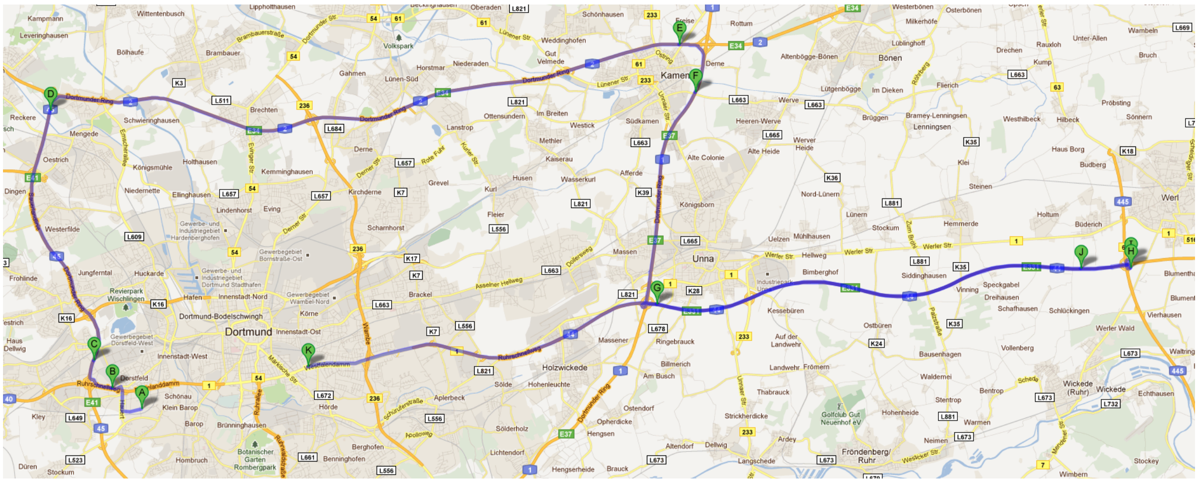


Figure 2: Experimental drive route.

summer holidays, no congested traffic occured on the freeways during the experimental drive, so both cars were able to travel with their desired velocity. In the "Vehicle speed" diagram in figure 3, the drops in velocity show freeway changing, they fit to the peaks in the "breaking processes" and "Vehicle ac-/deceleration"' diagrams also shown in figure 3. Further experimental drives are planned with focus on rush-hour traffic.

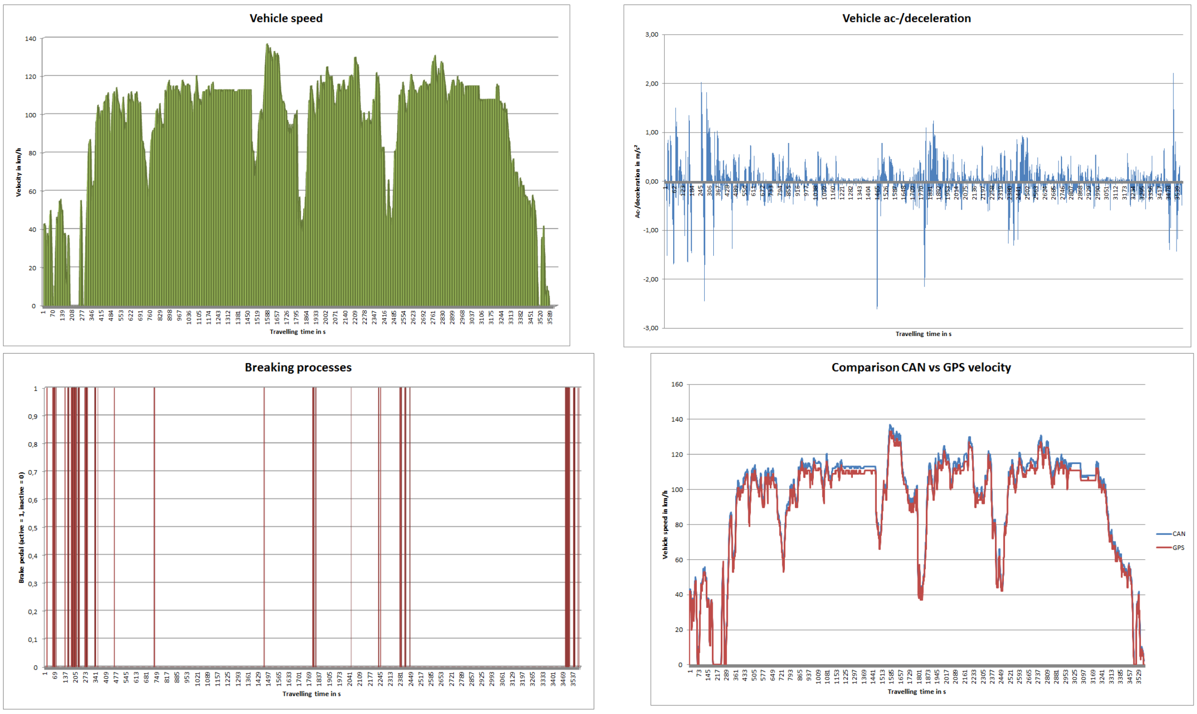


Figure 3: Visualization of the recorded CAN-tool data.

References

[1] http://de.wikipedia.org/wiki/Controller_Area_Network.

[2] B. Kerner. *The Physics of Traffic: Empirical Freeway Pattern Features, Engineering Applications, and Theory (Understanding Complex Systems)*. Springer, 2010.

[3] M. Treiber and D. Helbing. Reconstructing the spation-temporal traffic dynamics from stationary detector data. *Cooper@tive Tr@nsportation Dyn@mics 1, 3.1-3.24*, 2002.

[4] M. Treiber and A. Kesting. *Verkehrsdynamik und -siulation: Daten, Modelle und Anwendungen der Verkehrsflussdynamik (Springer-Lehrbuch)(German Edition)*. Springer, 9 2010.

Microscopic Traffic Simulation for generation of Virtual Floating Car Data

Daniel Weber
Physik von Transport und Verkehr
Universität Duisburg-Essen
weber@ptt.uni-due.de

We describe a microscopic modelling approach to freeway traffic flow. We use this model to simulate various penetration rates of vehicles providing Floating-Car-Data and present first results for travel time measurements based on these data.

1 Introduction

In recent years *Floating-Car-Data (FCD)*, i.e. cars providing their GPS measured localization and speed data, has become an important source for traffic information systems, providing travel time measurements and traffic state estimations. Our goal is to assess the quality and reliability of such FCD measurements. We will use microscopic traffic flow simulations to construct various traffic scenarios in which virtual FCD vehicles will serve as probes to gather the relevant data. Modern cars are not only equipped with GPS receivers but also possess a wide variety of sensors to detect the state of the car as well as its environment [3]. These *Extended Floating-Car-Data (XFCD)* pose a huge opportunity to gather additional traffic relevant information, e.g. about weather and road surface conditions to improve traffic information services.

2 Microscopic Traffic Flow Model

In microscopic traffic flow models vehicular traffic is treated as a system of interacting particles. The interactions of the individual cars lead then to macroscopic phenomena like traffic jams. One of the simplest such models is the Nagel-Schreckenberg (NaSch) model [2], a discrete stochastic cellular automaton model. Cars move on a lattice of L sites $i, i \in \{1, 2, ..., L\}$ which are either occupied by exactly 1 car or empty. The length of a cell corresponds to the space occupied by a car in a jam and is set to 7.5 m. The state of the lattice is described by the occupation variables of the cells τ_i, where $\tau_i = 0$ if cell i is empty and $\tau_i = 1$ if it is occupied. The cars are characterized by their position on the lattice $x_n(t)$ and their velocity $v_n(t)$, which is limited by a global maximum speed v_{max}. Another important variable is the spatial headway (the distance to the next car in front) d_n.

The time evolution of the system is governed by 4 update rules, which are applied to all cars at the same time (parrallel update):

1. Acceleration: $v_n(t + t_1) = min(v_n(t), v_{max})$.
2. Braking: $v_n(t + t_2) = min(v_n(t + t_1), d_n(t))$.
3. Randomization: $v_n(t+1) = \begin{cases} max(v_n(t + t_2) - 1, 0), & \text{w. prob.} p. \\ v_n(t_2) & \text{w. prob.} p - 1. \end{cases}$
4. Driving: $x_n(t+1) = x_n(t) + v_n(t+1)$

Without the randomization the model becomes totally deterministic, so that cars will travel with constant speed and distant headways. The randomization models fluctuations in driver behavior. Even freely moving cars may be travelling with speed less than v_{max} at some time and some drivers may overreact and reduce their speed more than needed to avoid an accident.

We consider a variaton of the NaSch-Model where the randomization parameter p is chosen according to the speed of the vehicle (*velocity dependent randomization* or *slow-to-start-rule*) [1]:

$$p_n(v_n) = \begin{cases} p_0 & \text{if } v = 0, \\ p & \text{if } v > 0 \end{cases}$$

with $p_0 > p$. The *slow-to-start-rule* reduces the jam outflow and thus stabilizes the jammed phase. This leads to a stronger separation between jammed and free-flow states. The model is implemented for a lattice with open boundary conditions. Cars are inserted on the left boundary with probability α, move through the lattice according to the update rules and leave at the right boundary with probablity β. This situation where cars cannot freely leave the lattice corresponds to a bottleneck situation, often encountered in real traffic (e.g. the merging of lanes at a construction site). The travel times measured in the simulation are calculated from the timestamps when the cars enter and leave the lattice (i.e. the travel time includes the passage of the bottleneck). For the Monte-Carlo simulations we used $\alpha = 0.5$, $\beta = 0.5$, $p_0 = 0.5$ and $p = 0.2$. The maximum velocity v_{max} was set to 5 cells per time step which for cells of 7.5 m length corresponds to a maximum speed of 135 km/h.

3 Virtual Floating-Car-Data

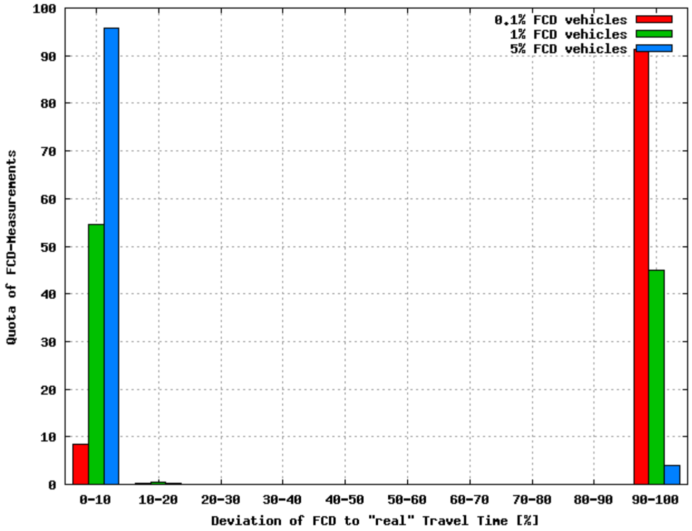


Figure 1: Deviation of FCD-measured travel time from “real” travel time for different FCD penetration rates

To extract virtual floating car data from the simulation, cars are randomly marked as FCD cars with probability *r* when they enter the lattice. For these cars position and velocity are logged in every time step and can be used for the reconstruction of the trajectory of the vehicles. We performed Monte-Carlo simulations for average penetration rates of 0.1%, 1% and 5%. The measured travel times were averaged over 5 minutes. As can be seen in Figures 1 and 2, even in such a simplified scenario an FCD coverage below 1% is clearly insufficient for a reasonable estimation of traffic characteristics. On the other hand the good agreement of travel times for a penetration rate of 5% has to be taken with some caution. In our simulation, there is only one type of car, while in reality the distribution of FCD vehicles is not necessarily the same as the distribution of all vehicles.

4 Conclusion and Outlook

At the moment we are testing the simulation for road networks and for lattices with defects, where on some section of the lattice a different driving behavior is enforced. We will then proceed to extend the simulation models to include more realistic driving behavior, traffic on multiple lanes and different vehicle classes, leading to realistic traffic scenarios in which the use of FCD can be evaluated. More detailed models will also allow to log additional data to assess the potential use of XFCD. A calibration of these

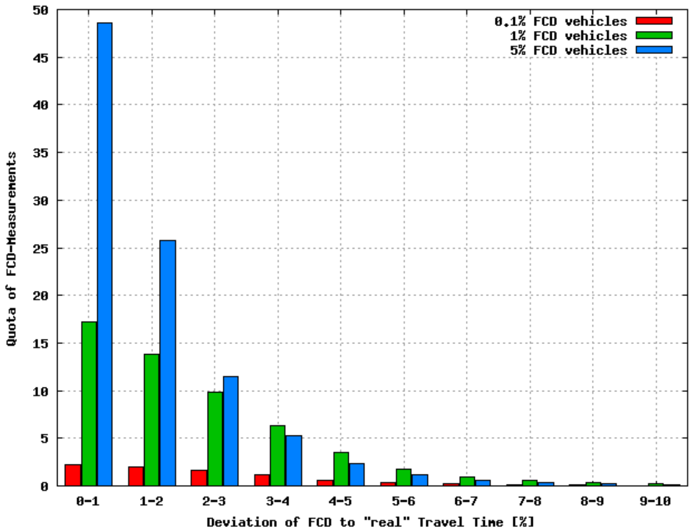


Figure 2: Zoom into the first bin of Figure 1

simulations will be done with data from static traffic sensors, probe drives with GPS equipped vehicles and data from weather stations along highways.

References

[1] R. Barlovic, T. Huisinga, A. Schadschneider, and M. Schreckenberg. Open boundaries in a cellular automaton model for traffic flow with metastable states. *Physical Review E*, 66:046113, 2002.

[2] K. Nagel and M. Schreckenberg. A cellular automaton model for freeway traffic. *J. Phys. I France*, 2:2221 – 2229, 1992.

[3] R. Quintero, A. Llamazares, D.F. Llorca, M.A. Sotelo, L.E. Bellot, O. Marcos, I.G. Daza, and C. Fernandez. Extended floating car data system - experimental study. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 631 –636, 2011.

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

Katharina Morik Alexander Schramm

Jarid 1c in neuroblastoma

Kathrin Fielitz
Oncology Lab - Children's Hospital
University Hospital Essen
Kathrin.Fielitz@uk-essen.de

Neuroblastoma is the most common solid extracranial malignancy in childhood. Since genetic predisposition is the main cause for this malignancy, we are seeking to identify genes relevant for neuroblastoma development. This is done, on exon level for chosen genes and also analyze genes which seem to cause a poor outcome for the patient, without taking remark to alternative exon usage. Momentarily the histone demethylase Jarid1c is in focus of attention. We detected differential location of the protein in different neuroblastoma cell lines and an effect of a Jarid 1c knock-down on protein expression as well as on the cell morphology.

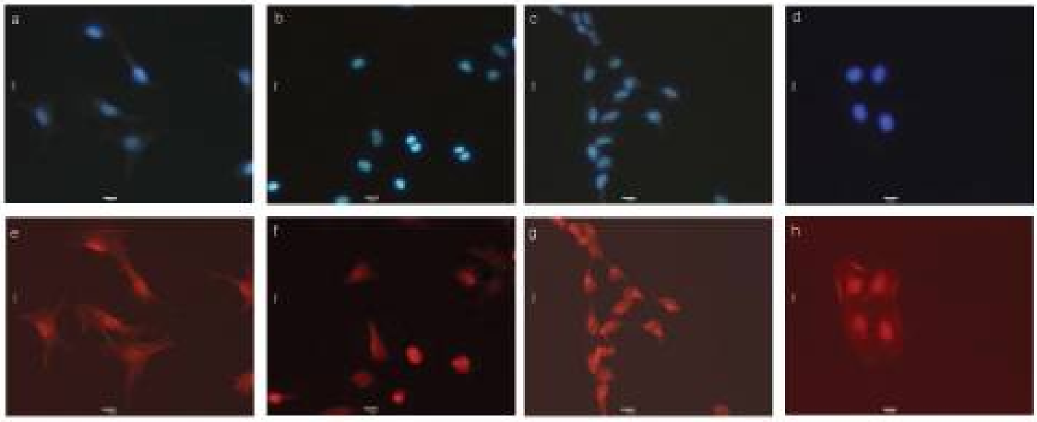
Neuroblastoma is the most common solid extracranial malignancy in childhood [1] and accounts for 15% of the deaths attributed to malignant conditions in children [2].
Neuroblastoma derives from the neural crest [3], multipotent migratory cell populations which give rise to tissue like the adrenal medulla, the parasympathetic ganglia or sensory neurons.
In the adult the major risks for developing a malignancy arise from age, exposition to carcinogenic agents and genetic predisposition. Obviously age cannot be a high risk factor for tumor development in children, as little as exposition to carcinogenic agents. Therefore, the genetic predisposition should be considered the main cause for malignancies in childhood.
In spite of the consistently ongoing advancement in developing and improving diagnostic and therapeutic possibilities, it remains difficult to make a reliable prediction about the course of the disease or outcome. So far the analysis of the mycn statushas been used for prognoses, since the amplification of this oncogene is associated with poor outcome [4]. Yet, mycn amplification is only found in 20% of neuroblastoma, and tumors with a single copy of the gene can also be aggressive. Therefore one aim of this project is the identification of further genes relevant to neuroblastoma development – and genes which

could serve as target structures for medical treatment.

We are trying to analyze the genes determined for differential exon usage, since exon analyses are much more precise than “normal” analyses. In previous experiments Affymetrix Microarrays, the data from which were used as a foundation for all the following experiments, showed that in patient groups with mycn amplification, an increased usage of different exons was detectable. So far we hypothesized that mycn conveys expression differences through alternative transcript usage, which we want to prove within this project. We also look at genes, which are expressed independently of the mycn status and which still lead to a poor outcome for the patient. Momentarily the focus is on the histone demethylase Jarid 1c. We analyze the expression and the effects of a knock-down through siRNA (small interfering RNA or silencing RNA), short RNA molecules, which can interfere with gene expression.

Histones are highly basic proteins around which DNA winds, forming nucleosomes an can be modified, for example through methylation. Lysine methylations exist in three different states – mono- di and trimethylations. (H3K4me3 means that lysine 4 on histone 3 is trimethylated.) The methylation of histones is associated with transcriptional activation and DNA damage response [5] [6]. As far as H3K4me3 is considered, it has been shown to regulate transcription [7]. It has been pointed out that Jarid 1c (SMCX) is able to demethylate H3K4me2/3 [8]. Jarid 1c was first described as the cause for X-linked mental retardation in 2005 [9].

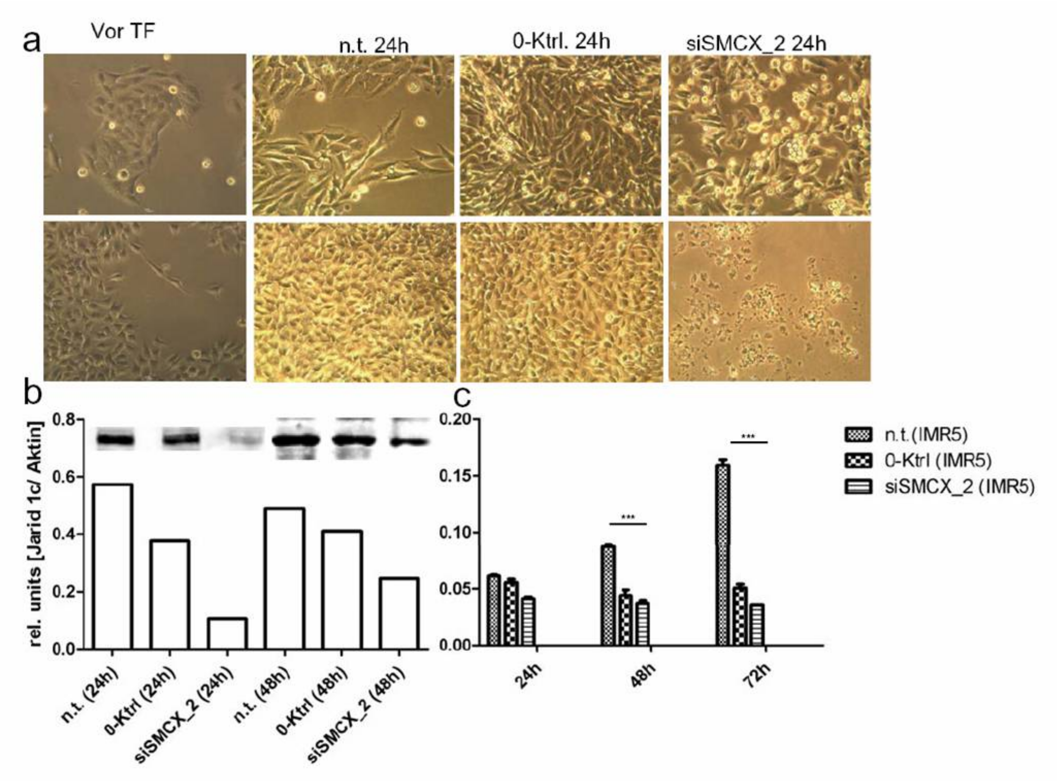
In the current experiments we were able to show that the protein expression in neuroblastoma cell lines (SHEP, NB 69, IMR 5 and IMR 32) (Fig.1) differs. In SHEP and IMR 32



1.jpg

Figure 1: Expression of Jarid1c (bottom row) and DAPI (top row) in neuroblastoma cell lines -SHEP (a, e); NB 69 (b, f); IMR 5 (c, g) and IMR 32(d, h). Scale bars equal 10um

the protein expression is localized predominantly to the nucleus, while in NB 69 and IMR 5 we find the expression more diffusely throughout the cell. The DAPI (4,6-Diamidin-2-phenylindol) staining serves for locating the nucleus within the cell, as it stains DNA.



2.jpg

Figure 2: The effect of Jarid1c knock-down in a) the cell morphology 24h after transfection (top row= SHEP, bottom row= IMR 5); b) protein expression in IMR 5 and c) cell proliferation in IMR 5. nt= non-treated. The graphs show average values +/- SEM

We were also able to show on protein level that Jarid1c is expressed in all these cell lines strongly (data not shown). The determination of the RNA expression of Jarid1c is in progress. Yet we analyzed the effect of siRNA transfection on the morphology of SH-EP and IMR 5, as well as on the protein expression in these cell lines and the effect of siRNA on the proliferation (Fig.2). We could very well see that a knock down of Jarid1c through siSMCX (40nM) lead to massive apoptosis in IMR 5 and to moderate apoptosis in SHEP, shortly after transfection (Fig. 2a).
It was also obvious, that the protein level in IMR 5 were decreased after knock down of Jarid 1c (Fig. 2b) and that the proliferation went down significantly, 48h and 72 h after transfection (Fig. 2c).
Based on our results we can assume that Jarid1c is a neuroblastoma relevant gene. One hint is given already through its strong expression in neuroblastoma patients with poor outcome. If we consider the effect of a knock-down of Jarid1c, which leads to a decrease in the amount of protein in the cell (Fig. 2b), increased apoptosis and decreased rate of proliferation, this will explain why Jarid1c could contribute so much to tumor development. We can conclude that the protein could be important for proliferation and cell

survival – when knocked down, the cells proliferate less and die.
Therefore we hope to soon be able to use these and coming results for better understanding of neuroblastoma biology. Precisely one of our goals is the ability to predict the outcome and that way a more appropriate treatment for the children. Hopefully we will be able to expand the aquired knowledge to different types of cancer as well.

References

[1] Oberthuer, A.; Berthold, F.; Warnat, P.; Hero, B.; Kahler, Y.; Spitz, R.; Ernestus, K.; König, R.; Haas, S.; Eils, R.; Schwab, M.; Brors, B.; Westermann, F.; Fischer, M.: Customized oligonucleotide microarray gene expression – based classification of neuroblastoma patients outperforms current clinical risk stratification. Journal of Clinical Oncology; 24, 5070 – 5078; 2006

[2] Janoueix-Lerosey, I.; Schleiermacher, G.; Delattre, O.: Molecular pathogenesis of peripheral neuroblastic tumors. Oncogene; 29, 1566 – 1579; 2010

[3] Brodeur, G.M.: Neuroblastoma: Biological insights into a clinical enigma. In Nature Reviews – Cancer; 3, 203 – 216, 2003

[4] Seeger, R.C.; Brodeur, G.M.; Sather, H.; Dalton, A.; Siegel, S.E.; Wong, K.Y.; Hammond, D.: Association of multiple copies of the N-myc oncogene with rapid progression of neuroblastomas. New England Journal of Medicine; 313(18):1111-1116; 1985

[5] Sanders, S.L.; Portoso, M.; Mata, J.; Bähler, J.; Allshire, R.C.; Kouzarides, T.: Methylation of histone H4 lysine 20 controls recruitment of Crb2 to sites of DNA damage. Cell; 119, 603 – 614; 2004

[6] Zhang, Y.; Reinberg, D.: Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tail. Genes and Development; 15, 2343 – 2360; 2001

[7] Liang, G.; Lin, V.W.; Yoo, C.; Nguyen, C.T.; Weisenberger, D.J.; Egger, G.; Takai, D.; Gonzales, F.A.; Jones, P.A.: Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proceedings of the National Academy of Sciences vol. 101; no. 19, pgs. 7357 – 7362, 2004

[8] Iwase, S.; Lan, F.; Bayliss, P.; de la Torre-Ubieta, L.; Huarte, M.; Qi, H.H.; Whetstine, J.R.; Bonni, A.; Roberts, T.M.; Shi, Y.: The X-linked mental retardation gene SMCX/Jarid1 defines a family of histone H3 lysine 4 demethylases. Cell, 128, 1077 – 1088; 2007

[9] Jensen, L.R.; Amende, M.; Gurok, U.; Moser, B.; Gimmel, V.; Tzschach, A.; Janecke, A.R.; Tariverdian, G.; Chelly, J.; Fryns, J.P.; Turner, G.; Reinhardt, R.; Kalscheuer, V.M.; Ropers, H.H.; Lenzner, S.: Mutations in the Jarid1c gene, which is involved in transcriptional regulation and chromatin remodelling causes X-linked mental retardation. The American Journal of Human Genetics, 76, 227 – 236; 2005

Functional validation of transcripts with alternative exon usage in neuroblastoma

Melanie Heilmann
Oncology Lab - Children's Hospital
University hospital 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 location in the abdomen or chest. The clinical presentation of this tumor can be very heterogeneous. The international Neuroblastoma staging system (INSS) classify the tumor in five stages according to the clinical presentation and age. This classification decides on therapy and outcome. In this project new neuroblastoma markers and new therapy targets will be identified to enable a better outcome prediction. Especially transcripts with alternative exon usage are in focus of this project, because the modification of splicing patterns has been shown to contribute to various diseases including cancer.

Neuroblastoma are the most common and deadly solid tumors in childhood. They account for 7-10 % of all cancers. The tumor of the autonomic nervous system derives from the neural crest tissue and usually arises in a paraspinal location in the abdomen or chest [1, 2]. This kind of cancer exhibits diverse and often dramatic clinical behavior. To enable an individual therapy it is essential that the molecular medicine will be consulted more often. In the last years many genetic features correlating with clinical outcome could be identified. It is known that the increase in gene copies of MYCN is connected with an unfavorable course [1]. These persons are classified as high-risk patients and treated accordingly. The amplifications can be observed in 20 % of the neuroblastoma but with the other 80 % without MYCN amplification, neuroblastoma can be found, showing an aggressive behavior as well [3]. To contribute to a better risk assessment this

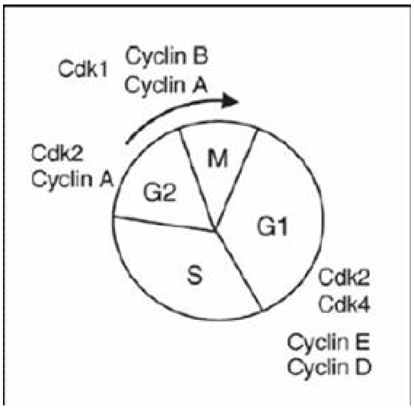


Figure 1: Cellcycle; the cycle consists of four phases, G1, S, G2 and M-phase. The transition between each phase are determined by the synthesis of cyclin and the activity of specific cdks [12]

project will deal with the identification of neuroblastoma relevant genes and transcripts. It will be focused on genes with an alternative exon usage especially. The usage of different exons could be established by alternative splicing. This post transcriptional mechanism allows the generation of protein diversity. Aberrant mRNA splicing has been described in many types of human cancer [4, 5, 6]. This change in mRNA can lead to the synthesis of new protein variants or the unbalanced expression of normal protein isoforms and this could initiate or sustain tumor growth [7]. Guo et al. 2011 could show that alternative splicing plays a significant role in high stage neuroblastoma and they suggested a MYCN-associated splicing regulation pathway [8]. The search of such genes is based on Affymetrix Exon array (HuEx 1.0 ST) data which give us an expression profile of the neuroblastoma. Transcripts with different isoforms. In the following steps the role of these transcripts for the neuroblastoma has to be clarified. One of these transcripts is cyclin B1 (CCNB1). CCNB1 is a cell cycle regulator. It activates the cyclin dependent kinase1 (cdk1) and promotes the passage through G2-phase to M-phase (fig.1). The deregulated expression of this gene is supposed to be involved in neoplastic transformation. So it could be a very interesting target for cancer therapies. The microarray shows that neuroblastoma express both a long and a short isoform. The long isoform is constitutively expressed whereas the short form is cell cycle dependent and predominantly expressed during G2/M-phase. Previous analyses indicated that the short isoform is higher expressed in patients with relapse (fig.2). So it can be hypothesized that the isoforms have different functions in the cell. More over RT-PCRs showed that various neuroblastoma cell lines (SH-EP, SY5Y, NB69, IMR32 and Wac II) express different levels of these isoforms and also the ratio between these isoforms is cell line dependent. Further investigations suggest that the knock down of both but not of the long isoform by siRNA causes a G2-arrest in SH-EP cells whereas Wac II cells are not affected by the siRNA. A G2-arrrest means that the cells continue to replicate their DNA but they do not divide. This experiment must be repeated with other cell lines in the near

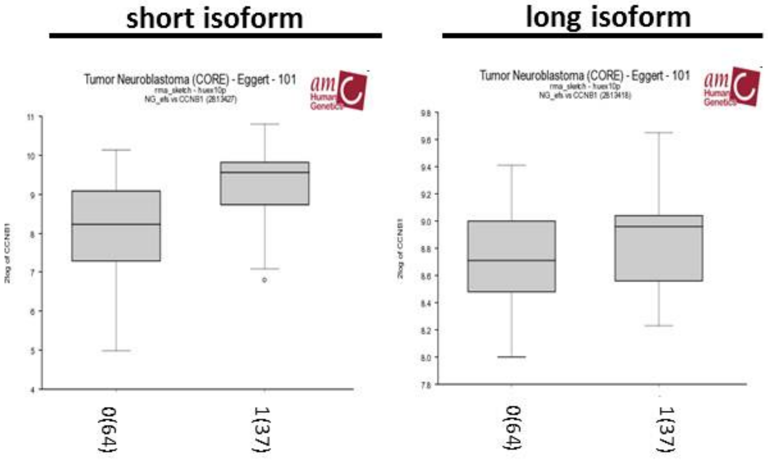


Figure 2: The short isoform of CCNB1 is associated with an unfavorable outcome; 0= event-free survival, 1=relapse]

future. In recent months I performed some experiments with a cdk1/cyclin B1 inhibitor (RO-3306) to check the effect on cell cycle and if it depends on the CCBN1 isoform expression pattern. To investigate the expression of further genes, which are predicted to express different isoforms, PCRs was and will be performed (these experiments are still in work). At the moment the expression of two transcripts, neurocan (NCAN) and plexinA4 (PLXNA4), are checked by real-time-PCR. PlexinA4 belongs to a semaphorin receptor family. They are single pass transmembrane receptors characterized by an intracellular GTPase activating domain. The results of Kigel et al. suggest that PLXNA4 may represent a target for the development of novel anti-angiogenic and anti-tumorgenic drugs [9]. The analyses of our data show that PLXNA4 is associated with favourable outcome and TrkA, a neutrophin receptor, and anti-correlated with MYCN amplification. Therefore the impact of TrkA and MYCN on the expression of NCAN and PLXNA4 should be investigated with the help of inducible cell lines (SY5Y-TR-TrkA and SH-EP-NMYC). Inducible cell lines contain a system which allows the cells to express a particular protein by the adding of a specific substance for example tetracycline. So a controlled expression of TrkA and NMYC is possible. The second transcript of interest is NCAN a brain-specific chondroitin sulfate proteoglycan of the extracellular matrix. In normal tissue it plays a role in cell-binding neurite outgrowth and adhesion [10]. But its role in cancer is still unknown. So this could be an interesting target for further experiments. Another project, which is planned, is the characterisation of Neuroblastoma 4s cells. Neuroblastoma classified as 4s show dissemination but a paradoxically favorable prognosis and a high rate of spontaneous tumor progression. Despite excellent prognosis for NB4s 10-25% of these patients nevertheless do not survive as the patient from whom our NB4s cells were removed [11]. In the next steps the expression pattern of TrkA isoforms of TrkA positive NB4s cells has to be analysed. Moreover in the next months I will investigate the expression of transcripts with different isoforms and if TrkA and/or MYCN impact on the expression of these. In further experiments I will analyse the role of these isoforms in the cell. Moreover I will continue the cdk1/CCNB1 inhibitor experiments and analyse if there is a correlation between the expression pattern of the CCNB1 isoforms and the effect of the inhibitor.

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] Brodeur GM, Seeger RC, Schwab M, Varmus HE, Bishop JM. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. Science 1984, 224:1121-4
[4] French P.J., Peeters J., Horsman S., Duijm E., Siccama I., Van Den Bent M.J., Luider T.M., Kros J.M., van der Spek P., Sillevis Smitt P.A. Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. Cancer Res 2007, 67:5635–42
[5] Gardina P.J., Clark T.A., Shimada B., Staples M.K., Yang Q., Veitch J., Schweitzer A., Awad T., Sugnet C., Dee S., Davies C., Williams A., Turpaz Y. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 2006, 7:325
[6] David C.J., Manley J.L. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev 2010, 24: 2343-64
[7] Pajares M.J., Ezponda T., Catena R., Calvo A., Pio R., Montuenga L.M. Alternative splicing: an emerging topic in molecular and clinical oncology. Lancet Oncol 2007, 8:349–57.
[8] Guo X., Chen Q., Song Y.K., Wei J.S., Khan J. Exon array analysis reveals neuroblastoma tumors have distinct alternative splicing patterns according to stage and MYCN amplification status. Medical Genomics 2011, 4:35
[9] Kigel B., Rabinowicz N., Varshavsky A., Kessler o. Neufeld G. Plexin-A4 promotes tumor progression and tumor angiogenesis by enhancement of VEGF and bFGF signaling 2011
[10] Rauch U., Feng K., Zhou X.-H. Neurocan: a brain chondroitin sulfate proteoglycan. Cell. Mol. Life Sci. 2001, 58:1842–1856
[11] Noesel M., Hählen K., Hakvoort-Cammel F., Egeler M. Neuroblastoma 4s A heterogenouse disease with variable risk factors and treatment strategies. American cancer society 1992, 80:834-843
[12] Pines J. Cyclins: Wheels within wheels. Cell Growth and Diiferentiation 1991, 2:305-310

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

Roland Fried Wolfgang Rhode

Threshold Optimization for Classification in the MAGIC Experiment

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

We introduce a method to choose an optimal discrimination threshold in the outcome of a random forest, which is used to classify gamma and hadron events in the MAGIC experiment. We choose the threshold to minimize the mean square error (MSE) of the estimation of the true number of gamma events. Estimating this number is an important step in the MAGIC analysis chain.

The MAGIC telescope on the canary island of La Palma is an imaging atmospheric Cherenkov telescope. Its purpose is to detect Cherenkov light [4] from particle showers in the atmosphere, induced by highly energetic gamma-rays, which have been sent out by astrophysical sources like active galactic nuclei (AGNs) [5]. The problem is that not only gamma rays induce such particle showers, but also many other particles summarized as hadrons, which are 100 to 1000 times more common than the gamma-rays of interest for the strongest sources [9]. So the gammas have to be separated from the hadrons via classification.

The initial situation is that we have a training sample of n events, where $n_{g.}$ is the number of gamma observations and $n_{h.}$ is the number of hadron observations in

this data set. In addition to the training sample, we have a sample of real data consisting of N events with $N_{g\cdot}$ and $N_{h\cdot}$ defined analogously, but unknown. We denote the numbers of observations in the training sample after the classification as $n_{ij}, i, j \in \{g, h\}$, where the first index indicates the true class and the second index the class as which the event was classified. Analogously we define $N_{ij}, i, j \in \{g, h\}$ as the corresponding numbers in the real data. Like $N_{g\cdot}$ and $N_{h\cdot}$ these values are unknown in actual application.
In addition to these two datasets in our specific astronomical setting we have access to so-called Off-data which is real data consisting only of negatives. We define $N_{h\cdot}^{off}$ as the number of hadron observations in this data set (which is equal to the total number of events in this data, N^{off}). Classifying this data set like the others we obtain N_{hg}^{off} and N_{hh}^{off}.

The True Positive Rate (TPR), which is also known as Recall or Efficiency, and the False Positive Rate (FPR) are defined as

$$TPR = \frac{n_{gg}}{n_{g\cdot}} \qquad \text{and} \qquad FPR = \frac{n_{hg}}{n_{h\cdot}}.$$

Obviously, both rates can only take values in the interval $[0, 1]$ and a high value of TPR and a low value of FPR is desired. These two aims are contradictory.

The aim of the classification in our astronomical application is however not primaryly a good separation of gamma and hadron events, that is a high TPR and a low FPR. Instead one is interested in estimating the real number of gamma events $N_{g\cdot}$ from the number of observations classified as such, $N_{\cdot g}$. The aim of our work is thus to minimize the mean sqare error (MSE) of the estimation of the number of gamma events $N_{g\cdot}$. To estimate this number we use

$$\widehat{N}_{g\cdot} = \frac{n_{g\cdot}}{n_{gg}} \left(N_{\cdot g} - N_{hg}^{off}\right) = \frac{1}{TPR} \left(N_{\cdot g} - N_{hg}^{off}\right). \tag{1}$$

In the MAGIC analysis chain, the MAGIC analysis and reconstruction software MARS [3; 6], the classification is usually done by random forests [1; 2]. The output of such classifiers is the fraction of trees which voted for an observation to be a hadron. This fraction is called hadronness in the MAGIC analysis. A threshold in the hadronness has to be applied to finally classify the data. As our aim is an optimal estimation of the number of gamma events $N_{g\cdot}$, we choose the threshold to minimize the MSE of the estimation.

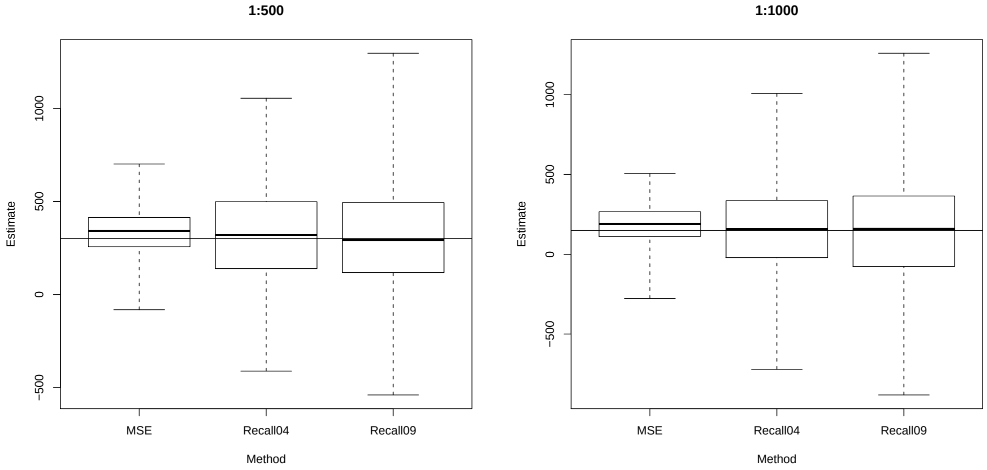


Figure 1: Boxplots of the estimates in the 500 samples with gamma to hadron ratio 1:500 (left) and with ratio 1:1000 (right). The thick line in the middle of each box represents the median of the estimates. Between the upper and lower boundaries of each box lie 50% of the estimates. The true number is marked by the long horizontal line [8].

It can be shown that under justifiable additional assumptions [8] the MSE of the estimation of the number of gamma events according to equation (1) becomes

$$\begin{aligned}\mathrm{MSE}\left(\widehat{N}_{g\cdot}|N_{g\cdot},N_{h\cdot},N_{h\cdot}^{off}\right) &= \frac{FPR^2}{TPR^2}\left(N_{h\cdot}-N_{h\cdot}^{off}\right)^2\\ &+ N_{g\cdot}\left(\frac{1}{TPR}-1\right)+\frac{FPR-FPR^2}{TPR^2}\left(N_{h\cdot}+N_{h\cdot}^{off}\right).\end{aligned} \tag{2}$$

We use this equation in an iterative algorithm to alternately estimate $N_{g\cdot}$ and minimize the MSE over the discrimination threshold until some convergence criterion is met. We thus receive the threshold with which the number of gamma events can be estimated with (approximately) the smallest error. This method can be regarded as an adaption of the thresholding method introduced in [7]. The difference is that we here use the MSE instead of the in our application unknown misclassification costs to be minimized. In fact, the MSE can be regarded as some kind of misclassification costs, as a single falsely classified observation leads to an increase in the MSE and the MSE reaches 0 for a perfect classification. In other words, the MSE implicitely weights every falsely classified observation. The method of minimizing the MSE thus implicitely handles the problem of unequal and unknown misclassification costs. It can be shown that the method also handles the problem of high class imbalance.

The result of tests with simulated data can be seen in Figure 1. It shows boxplots of 500 estimations of the number of gamma events for the method of minimizing the MSE presented here (MSE) and two methods which are currently used in the MAGIC analysis chain (Recall04/09). It can be seen that the MSE method typically estimates the true number with a smaller error than the two other [8].

References

[1] J. Albert et al. Implementation of the Random Forest Method for the Imaging Atmospheric Cherenkov Teleskope MAGIC. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 588(3):424–432, 2008.

[2] L. Breiman. Random Forests. *Machine Learning*, 45:5, 2001.

[3] T. Bretz, R. Wagner, and MAGIC collaboration. The MAGIC Analysis and Reconstruction Software. In *International Cosmic Ray Conference*, volume 5 of *International Cosmic Ray Conference*, pages 2947–+, 2003.

[4] P. A. Cherenkov. Visible emission of clean liquids by action of gamma radiation. *Doklady Akademii Nauk SSSR*, 2:451+, 1934.

[5] J. A. Hinton and W. Hofmann. Teraelectronvolt Astronomy. *Annual Review of Astronomy & Astrophysics*, 47:523–565, September 2009.

[6] A. Moralejo et al. MARS, the MAGIC Analysis and Reconstruction Software. In *International Cosmic Ray Conference*, International Cosmic Ray Conference, July 2009.

[7] V.S. Sheng and C.X. Ling. Thresholding for making classifiers cost-sensitive. In *Proceedings of the 21st national conference on Artificial intelligence*, volume 1, pages 476–481. AAAI Press, 2006.

[8] T. Voigt, R. Fried, M. Backes, and W. Rhode. Threshold optimization for classification in imbalanced data with unknown misclassification costs. *submitted to Advances in Data Analysis and Classification*, 2011.

[9] T.C. Weekes. *Very High Energy Gamma-Ray Astronomy*. Institute of Physics Publishing, Bristol/Philadelphia, 2003.

Periodicity detection in irregularly sampled light curves by robust regression and outlier detection

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

We investigate the application of regression techniques for periodicity detection in light curves. Light curves occur in astroparticle physics and are time series with special features like irregular periodic sampling and red noise. We consider robust instead of least squares (L_2) regression to calculate the periodogram, and fitting a distribution to the periodogram bars instead of assuming a predefined one to detect valid periods. Two studies are presented exploring the behavior of our proposed approaches in situations without and with outliers. This is a brief summary of the work submitted to Statistical Analysis and Data Mining, preprint available from the author.

Periodicity detection in nonuniformly sampled time series is a common task in various fields of study. We focus here on time series arising in astroparticle physics, called light curves, often with unequally spaced observation times due to bad weather conditions, lack of allocated telescope time and technical reasons. Moreover, astronomical cycles limit observability in a periodic manner, for example if a source is only visible in certain months. Such circumstances lead to periodic sampling. Light curve observations usually suffer from a varying measurement accuracy, which can be estimated for each observation time t_i, $i = 1, \dots, n$ leading to an estimate s_i which takes small values for high measurement accuracy. The resulting measurement errors $y_{w;i}$ are assumed to be independent normally distributed. So the present data is of the form (t_i, y_i, s_i), $i = 1, \dots, n$, where y_i is the observed signal. In addition to a period p_s influencing the sampling of the t_i, another period p_f may affect the observed signal in form of a periodic fluctuation. Besides, so-called red noise is expected as another signal component. A time series $y_r(t)$ is called

Model		Regression technique	
		Least squares	Robust
1	overlapping steps	PDM [16]	
2	steps	AoV [14]	
3	$a_1 \cos(2\pi t) + b_1 \sin(2j\pi t)$	LS [13]	
4	$a_0 + a_1 \cos(2\pi t) + b_1 \sin(2j\pi t)$	DCFT [5],SigSpec [12]	[1, 8]
5	$\sum_{j=0}^{2} a_j \cos(2j\pi t) + b_j \sin(2j\pi t)$	Fχ^2 [11]	
6	$\sum_{j=0}^{3} a_j \cos(2j\pi t) + b_j \sin(2j\pi t)$	Fχ^2 [11]	
7	periodic splines	GCV [10]	RCV [10]

Table 1: Characterization of some period finding methods as combinations of models and regression techniques. A more comprehensive list can be found in [17]. Underlined methods are performed using weighted regression.

red noise if its power spectrum $S(f)$ is proportional to $f^{-\alpha}$, $\alpha > 0$, i.e. if it follows a power law [18]. Altogether we assume the following model for a light curve (t_i, y_i, s_i), $i = 1, \dots, n$:

$$t_i = t^\star_{(i)}, \qquad t^\star_1, \dots, t^\star_n \sim \mathcal{D}(p_s), \tag{1}$$

$$y_i = y_{f;i} + y_{w;i} + y_{r;i}, \qquad y_{f;i} = g(t_i),\ y_{w;i} \sim \mathcal{N}(0, \sigma_i^2), \tag{2}$$

where $t^\star_{(i)}$ denotes the ith ordered observation time, $\mathcal{D}(p_s)$ is a periodic distribution depending on a sampling period p_s, $g(t) = g(t + p_f)$ is a periodic fluctuation in the observed signal of period p_f and $y_{r;i}$ is red noise. s_i is a given estimate of σ_i. Moreover one can allow the light curve to include substitutive outliers.

Fourier Analysis, as a standard method to find periodicities in an observed signal, cannot cope with the sampling situation described above, since equally spaced observation times are needed. A setting-adapted procedure, the so-called Deeming periodogram [4] has problems with periodically distributed observation times t_i and mainly finds the sampling period p_s (see [6]). Other methods for detecting periodic fluctuations in an irregularly and periodically observed signal have been developed. Many of them, such as the Lomb-Scargle Periodogram (LS) [13], Date-Compensated Fourier Transform (DCFT) [5], the Fast χ^2 Periodogram (Fχ^2) [11] but also methods like Phase Dispersion Minimization (PDM) [16] and the Analysis of Variance Periodogram (AoV) [14], which are usually called nonparametric, are based on fitting a periodic function to the light curve, typically by means of least squares (L_2) regression, sometimes using weights s_i^{-2} (see Table 1). A measure related to the coefficient of determination R^2, R^2 itself or the squared amplitude of the fitted function, is taken for the periodogram.

A few attempts have been made to use robust regression in this context until now (see Table 1 and further references in [17]). We investigate if robust regression techniques are reasonable alternatives to least squares regression. In doing so we compare different periodogram methods, all fitting a model of Table 1 by weighted L_2, Least Absolute Deviation, Huber- or Tukey M-regression [7] using weights s_i^{-2} and using R^2 (or the robust version [9], respectively) as periodogram. For $y_i = y_{w;i}$ and L_2 regression, R^2 is known to be $\mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)$-distributed, $\mathcal{B}$ denoting the beta distribution and m the degree of the fitted model [15]. As we use not only L_2 regression and not only white noise is assumed, we propose to robustly fit the parameters of the beta distribution using Cramér-von-Mises-Distance minimization [2] and determine peculiar periods using outlier detection [3].

In a performance study we apply our periodogram methods to artificial time series free of outliers, varying n, $\mathcal{D}(p_s)$, the amount of white and red noise, the presence or absence and shape of $y_{f;i}$, and p_f. We observe that robust methods work similarly well as L_2 methods in case of simple periodic fluctuations like sine-waves or triangular shaped periodic fluctuations. They detect less false periods in absence of a signal, but rarely find periods when the signal is a periodic peak with long interspaces. Fitting model 1 or 2 (see Table 1) with Huber M-regression works particularly well. An adaptive beta distribution leads to similar results as a $\mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)$ -distribution. In unclear cases the new approach leads to more conservative results, i.e. it detects less false positive periods in absence of a periodic fluctuation, but it occasionally misses an existing periodic fluctuation.

In a second smaller study introducing outliers we observe that using an adaptive beta distribution seems more suitable than the classical one and works well with least squares methods. When using robust regression, the true fluctuation period is not peculiar anymore.

Until now, we studied our methods in a quite stringent framework fixing a lot of constrains like linking the variance of red noise to that of white noise, letting the true period be one of the trial periods and assuming the periodogram bars to be independently distributed. It is an interesting question how well our methods work if the true period lies in between two trial periods. Will one of them be considered peculiar, or none of them? Moreover, our study indicates that it could be possible to detect even more differences between the different models using more disturbed settings, e.g. by increasing the level of noise.

References

[1] M. Ahdesmäki, H. Lähdesmäki, A. Gracey, I. Shmulevich, and O. Yli-Harja. Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. *BMC Bioinformatics*, 8(1):233, 2007.

[2] B. Clarke, P.L. McKinnon, and G. Riley. A fast robust method for fitting gamma distributions. *Statistical papers*, 2011. under revision.

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

[4] T.J. Deeming. Fourier analysis with unequally-spaced data. *Astrophysics and Space Science*, 36(1):137–158, 1975.

[5] S. Ferraz-Mello. Estimation of periods from unequally spaced observations. *The Astronomical Journal*, 86:619, 1981.

[6] P. Hall and M. Li. Using the periodogram to estimate period in nonparametric regression. *Biometrika*, 93(2):411–424, 2006.

[7] P.J. Huber and E. Ronchetti. *Robust statistics*, volume 1. Wiley, 1981.

[8] T.H. Li. A robust spectral analyzer for one-dimensional and multi-dimensional data analysis. 2009/0112954 A1, April 2009. US Patent Application.

[9] R.A. Maronna and V.J. Yohai. Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89(1–2):197–214, 2000.

[10] H.S. Oh, D. Nychka, T. Brown, and P. Charbonneau. Period analysis of variable stars by robust smoothing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):15–30, 2004.

[11] D.M. Palmer. A fast Chi-squared technique for period search of irregularly sampled data. *The Astrophysical Journal*, 695:496, 2009.

[12] P. Reegen. Sigspec i. frequency-and phase-resolved significance in fourier space. *A&A*, 467:1353–1371, 2007.

[13] J.D. Scargle. Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

[14] A. Schwarzenberg-Czerny. On the advantage of using analysis of variance for period search. *Monthly Notices of the Royal Astronomical Society*, 241:153–165, 1989.

[15] A. Schwarzenberg-Czerny. The distribution of empirical periodograms: Lomb-Scargle and PDM spectra. *Monthly Notices of the Royal Astronomical Society*, 301(3):831–840, 1998.

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

[17] 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*, 2011. submitted.

[18] J. Timmer and M. König. On generating power law noise. *Astronomy and Astrophysics*, 300:707–710, 1995.

Development of a Monte-Carlo simulation for propagating leptons

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 [4].
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 [5] 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. Currently the most important propagtion software is 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 MMC and its successor PROPOSAL

As mentioned above MMC is currently the main propagation program in the IceCube Monte-Carlo chain. MMC provides the possibility to propagate leptons and monopols through any type of media. 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. MMC takes the most important energy loss mechanisms into account:

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

From physical point of view MMC is a good choice for simulating leptons and monopoles. But several technical problems makes a revision of MMC necessary:

MMC is written in Java. The Problem of this is that apart from MMC the whole Monte-Carlo chain is writen in C++. Calling java-methods from a C++ program reduces the speed of simulation significantly.
Another issue is that MMC is nearly unmaintainable cause of missing comments and its unintuitiv structure. This makes it very hard to implement other physical effects someone could be interested in.

According to the issues above, we decided to develop a new propagation tool based on MMC but written in C++. The new propagator is called PROPOSAL
(**PR**opagation with **O**ptimal **P**recision and **O**ptimized **S**peed for **A**ll **L**eptons).

2 Status and Plans

The first and major step to develop PROPOSAL is translating MMC into C++. This task is completed.
A very important parameter is the final energy of the propagated particle. In figure 1 the distribution of the final energy calculated by MMC and PROPOSAL is shown. Within the statistical errors MMC and PROPOSAL produce the same results.

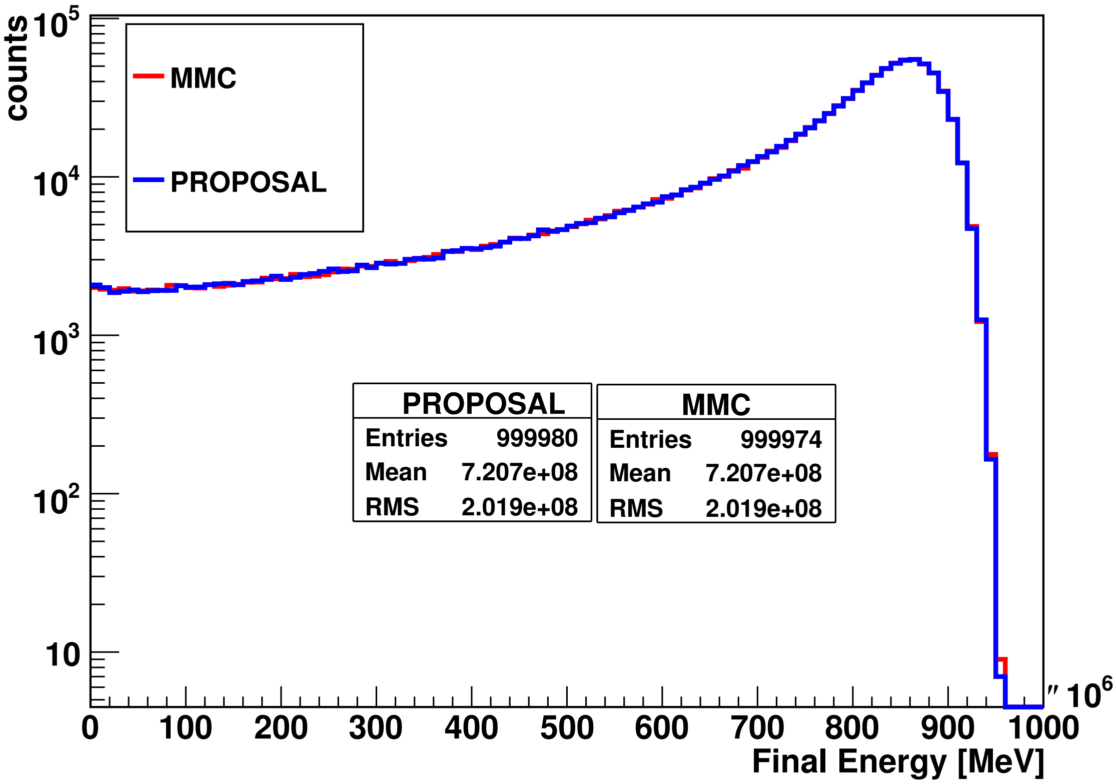


Figure 1: Distributions of the final energy of one million muons with an initial energy of 1 PeV calculated using PROPOSAL and MMC.

In Figure 2 the different energy loss mechanisms are compared to the calculations of Groom et al. [3] which are used as reference data.

The next step is to integrate PROPOSAL into the IceCube Monte-Carlo chain. It is planed to implement PROPOSAL for GPUs. In contrast to Java C++ is very suited to run on GPUs. This parallelization will speed up the simulation a lot and therefore save computing time and money.

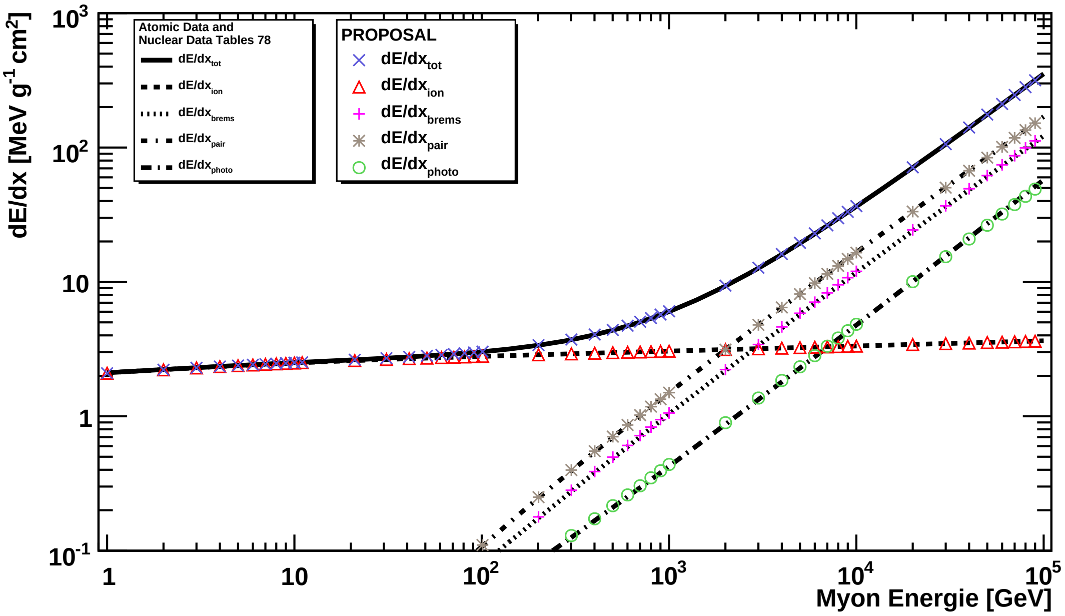


Figure 2: PROPOSALs results of energy loss mechanisms compared to reference data [3].

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] Donald E. Groom, Nikolai V. Mokhov, and Sergei I. Striganov. Muon stopping power and range tables 10-MeV to 100-TeV. *Atom. Data Nucl. Data Tabl.*, 78:183–356, 2001.

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

[5] 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.

A novel integrated data aquisition system for Cherenkov telescopes

Dominik Neise
Experimentelle Physik 5
Technische Universität Dortmund
dominik.neise@tu-dortmund.de

The current status of the Cherenkov telescope camera FACT [3], which is the object of the authors studies about data analysis under constraint conditions, is depicted in this short report. FPGA based highly intgraded data aquisition systems, such as the one used within FACT, might provide new ways to reduce the vast amount of data usually taken by this kind of detectors in an early stage.

1 Introduction

High energy cosmic rays consisting of a variety of particles such as protons, electrons and high energetic photons impinge the earth's atmosphere and cause so called *showers* of secondary particles. These showers of relativisic charged particles emit Cherenkov radiation, while beeing stopped in the atmosphere. This process lasts nanoseconds, thus an adequate photodetector is able to detect these blue flashes of light. [4]

2 The FACT camera

We are currently building a Cherenkov camera with a 2 GHz sampling rate of the photodetector signal in order to increase the ability to distinguish between gamma induced and proton induced Cherenkov showers, cf. [2] . Because only Cherenkov showers induced by high energistic primary photons carry information about their origin, while high energistic

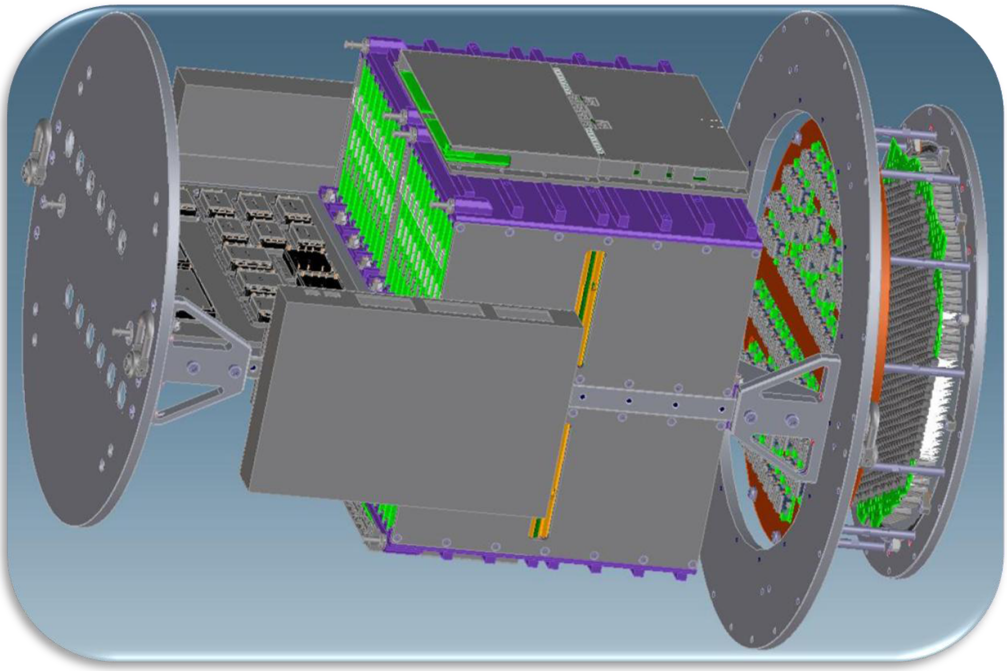


Figure 1: Schematic view of the FACT camera. The crates in the center house preamplifiers, trigger generators as well as the analog to digital converters.

protons, beeing charged particles, undergo deflections in intergalactic magnetic fields, thus loosing their original direction.

The FACT camera will be mounted on the refurbished mount of the former HEGRA CT3. The reflector area is about $10\,\mathrm{m}^2$, which leads to an estimated trigger rate of about 100Hz. The number of pixels is 1440, which is comparable to other Cherenkov cameras.

3 The Data Aquisition

The amount of data produced by a theoretical FACT camera beeing read out at a sampling rate of 2 GHz is about 6 TB/s. This amount of data can neither be transmitted from the camera to any recipient, nor would it be smart to store this vast amount of data. Apart from the fact, that 1440 ultra fast 2 GHz ADCs would consume a ridiculous amount of money and power. In order to reduce the amount of data, a highly sophisitcated trigger system was designed for the FACT camera. The typical signature of a Cherenkov shower is the coincident arrival of photons in a number of adjacent pixels in the camera. An extensive simulation has shown, that it is not necessary to implement elaborate *next*

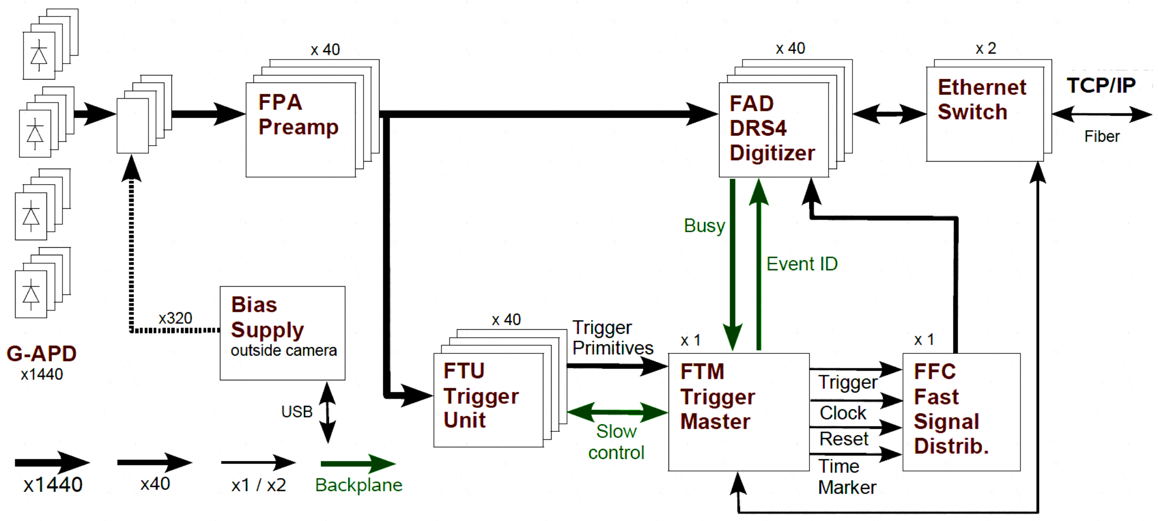


Figure 2: building blocks of the FACT camera

neighbor triggers, which would create the need for a lot of interconnections between the otherwise unconnected neighbors. Actually these simulations have shown, that the sumsignal of a number of neighboring pixels provides good means for finding coincident Cherenkov photons, and performs even better than next neighbor triggers towards lower energies.

4 Data Reduction

The DRS4 based digitizer is capable of storing a number of 1024 analog samples for each of the camera pixels, which results in a estimated data rate of only 0.3 TB/s. A remarkable feature of the DRS4 chip, is its capability to readout less than the entire 1024 samples. By reducing the number of read out samples, one can further reduce this amount to about 50 MB/s.

While this amount of data can be sent to a nearby data aquisition and storage system, neither the transmission of this kind of raw data from the usually remote telescope site to the interested scientist, nor the setup of a computing farm seems feasible. The on site data aquisition system is also capable of applying certain standard calibration and analysis steps, ending with the calaculation of the arrival time and number of photons in each of the camera pixels, further reducing the amount of data to about 3 MB/s. Only now it is possible to form a conclusive picture of the Cherenkov shower containing only few geometrical characterics of the shower such as, its size, orientation and overall light content. The number of characteristics typically is of the order of 20, so the amount of data is further reduced to about 16 kB/s.

The next step is to clean the data consisting of pictures of Cherenkov showers, which were induced by all kinds of cosmic rays, from the unwanted pictures of e.g. proton induced showers. This includes the use of classificators such as random forests [1], which are trained on simulated data. Since the qualitiy of classification highly depends on the quality of the simulated data, which depends not only on static telescope features but also changing ones such as the weather or the angle of the telescope during data taking, large computing facilities are needed here. So for the final step the data is sent to off site facilities for further analysis.

5 Outlook

In order tu reduce the amount calibration and cleaning steps, performed by the data aquisition computers outside the FACT camera, it is to be understood, whether FPGA based onbaord data calibration, zero suppression or even arrival time calculation are practicable.

References

[1] J. Albert et al. Implementation of the Random Forest Method for the Imaging Atmospheric Cherenkov Telescope MAGIC. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 588(3):424–432, 2008.

[2] E. Aliu et al. Improving the performance of the single-dish Cherenkov telescope MAGIC through the use of signal timing. *Astroparticle Physics*, 30:293–305, January 2009.

[3] H. Anderhub et al. A novel camera type for very high energy gamma-ray astronomy based on Geiger-mode avalanche photodiodes. *JINST*, 4:P10010, 2009.

[4] P. A. Cherenkov. Visible emission of clean liquids by action of gamma radiation. *Doklady Akademii Nauk SSSR*, 2:451+, 1934.

Test for spectral steadiness of gamma-ray data

Michael Backes
Experimentelle Physik 5 – Astroteilchenphysik
Technische Universität Dortmund
michael.backes@physik.tu-dortmund.de

Observations and detections of Active Galactic Nuclei (AGN) by Cherenkov telescopes are often triggered by information about high flux states in other wavelength bands. To overcome this bias, the VHE gamma-ray telescope MAGIC has conducted dedicated monitoring observations of nearby AGN since 2005. The goals of these observations are to obtain an unbiased distribution of observed flux states shedding light on the duty cycle of AGN and to investigate temporal variability and potential spectral changes during periods of different source activity. By testing predictions of theoretical models, like the correlation between the TeV flux level and its peak frequency predicted in leptonic emission models, monitoring observations deepen our knowledge about the emission processes in AGN. For this, a test of spectral steadiness is of special importance.

1 Blazars

By far, most of the known extragalactic emitters of very high energy (VHE, $E \gtrapprox 100$ GeV) gamma-rays are AGN. Among those, the largest subclass is comprised by blazars, being characterized by the relativistic jet pointing towards the observer [8]. They show non-thermal continuum emission ranging from radio to VHE gamma-rays. The emission is typically highly variable in all wavebands and on timescales ranging from minutes [1] to years [9]. Although theoretical models, generally, can explain the spectral shape of the observed blazar emission, the question whether electrons or hadrons are causing the high energetic electromagnetic emission in blazars is far from being settled. The exact spectral shape of the emission in the high energy and VHE region might shed light on this.

2 Cherenkov Telescopes

Imaging Atmospheric Cherenkov Telescopes (IACTs) are comprised by large mirrors and fast low light level detectors. They indirectly detect gamma-ray radiation in the VHE regime by sampling the visible Cherenkov light emission of the particle showers induced by gamma-rays hitting the atmosphere, c.f. [12]. Being much more sensitive than satellite experiments, like e.g. Fermi-LAT[1], IACTs suffer from their extremely limited fields of view, compared to the former. Thus, instead of all-sky surveys IACTs are used for deep single source exposures, often leading to a dependency on external triggers for the observation of already known sources. To overcome this dependence, monitoring observations independent of the source state are performed. Only by these, it is possible to obtain an unbiased distribution of observed flux states of the observed sources and to systematically investigate the possibility of changes in the emission spectra depending on the flux level of the sources and to compare such findings with the predictions by the theoretical models.

3 The MAGIC telescopes

The MAGIC Telescopes are the largest IACTs for VHE gamma-ray astronomy, featuring two times 234 m^2 mirror area. They are situated at 2200 m a.s.l. in the Observatorio del Roque de los Muchachos on the Canary Island of La Palma. The first MAGIC telescope has been in scientific operation since 2004. In the meanwhile, the separation of signal events from the hadronic background events and thus the sensitivity could significantly be improved by means of hardware and software developments [4].To further improve the analysis, alternatives to the used data mining method (random forest [2]) optimizing the sensitivity under resource constraints are as well under investigation [6, 7] as algorithms to minimize the systematic error of these methods [10]. In 2009 the second telescope started scientific operation, leading to a sensitivity improvement of a factor of two in the whole energy range from 50 GeV to several TeV, improving at the same time also the energy resolution [3].

4 Spectral Steadiness Test

Having monitored blazars with MAGIC over several years in an unbiased way [11] obviously the question arises whether or not these data may be used for common analysis. As blazars are generally variable in flux and spectral shape of their VHE emission the steadiness of the source's behavior must be tested. An example for this is depicted in

[1] http://fermi.gsfc.nasa.gov

Figure 1. The fit residuals (top panel), Cook's Distance [5] (second panel) as well as histograms of the fit residuals and the χ^2 test probabilities of Gaussian fits depending on the number of histogram bins are shown for a power-law (blue), a power-law+exponential cut-off (green) and a curved power-law (red) model fitted to data taken from an AGN over several years.

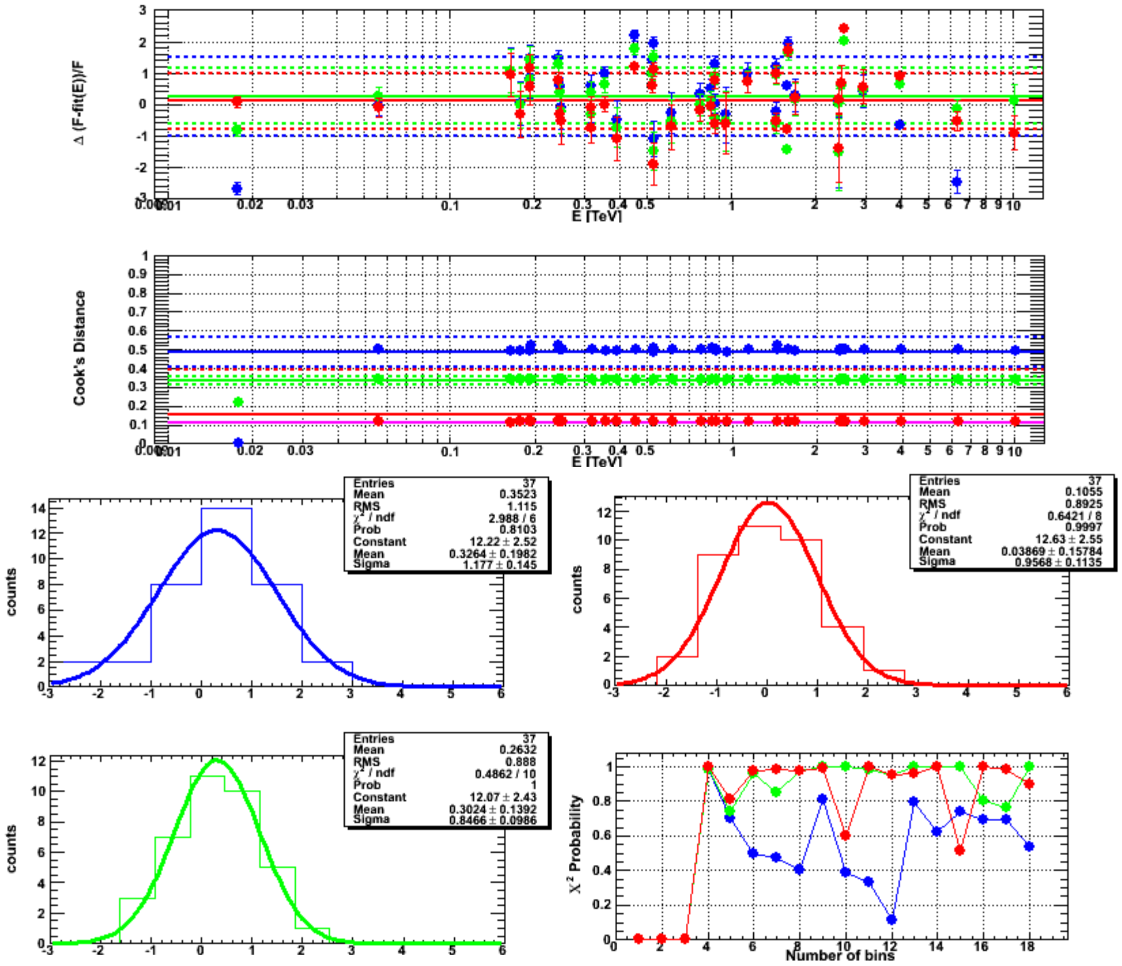


Figure 1: Test for spectral steadiness. Data taken during several years are comonly fit with different models (color-coded) and tested for spectral steadiness. Details can be found in the main text.

5 Conclusion and Outlook

In this report it has been shown that for the data sets under investigation all belong to a set of steady spectra and thus may be used for a common further analysis. Methodically, this result could further be strengthened with a method computing the Cook's Distance not only for single points but for sets of data points belonging to the same dataset or by the construction of confidence bands via bootstrapping methods.

References

[1] J. Albert et al. Variable Very High Energy γ-Ray Emission from Markarian 501. *Astrophysical Journal*, 669:862–883, November 2007.

[2] J. Albert et al. Implementation of the Random Forest Method for the Imaging Atmospheric Cherenkov Telescope MAGIC. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 588(3):424–432, 2008.

[3] J. Aleksić et al. Performance of the MAGIC stereo system obtained with Crab Nebula data. *Astroparticle Physics*, August 2011. submitted.

[4] E. Aliu et al. Improving the performance of the single-dish Cherenkov telescope MAGIC through the use of signal timing. *Astroparticle Physics*, 30:293–305, January 2009.

[5] R.D. Cook. Detection of Influential Observation in Linear Regression. *Technometrics*, 19 No. 1:15–18, February 1977.

[6] M. Helf. Gamma-Hadron-Separation im MAGIC-Experiment durch verteilungsgestütztes Sampling. Diplomarbeit, Technische Universität Dortmund, April 2011.

[7] M. Helf, K. Morik, M. Backes, and W. Rhode. Sampling for Gamma Separation – Data Mining on the Rocks. 2011. In preparation.

[8] P. Padovani and C. M. Urry. Luminosity functions, relativistic beaming, and unified theories of high-luminosity radio sources. *Astrophysical Journal*, 387:449, March 1992.

[9] A. Sillanpaa, S. Haarala, M. J. Valtonen, B. Sundelius, and G. G. Byrd. OJ 287 - Binary pair of supermassive black holes. *Astrophysical Journal*, 325:628–634, February 1988.

[10] T. Voigt, R. Fried, M. Backes, and W. Rhode. Threshold Optimization for Classification in Imbalanced Data with Unknown Misclassification Costs. *Advances in Data Analysis and Classification*, 2011. Submitted for publication.

[11] R. Wagner et al. Monitoring of bright, nearby Active Galactic Nuclei with the MAGIC telescopes. In *32nd International Cosmic Ray Conference, Beijing, China*, International Cosmic Ray Conference, page 1030, August 2011. Accepted for publication.

[12] T.C. Weekes. *Very High Energy Gamma-Ray Astronomy*. Institute of Physics Publishing, 2003.

Use of RapidMiner in the MAGIC data analysis

Nikola Strah
Lehrstuhl Experimentelle Physik 5 - Astroteilchenphysik
Technische Universität Dortmund
nikola.strah@tu-dortmund.de

To enable a deep and detailed study of physical properties of active galactic nuclei, observational campaigns with instruments in all wavelength bands have been performed. With MAGIC, a system of two Cherenkov telescopes, it is possible to achieve an excellent characterisation of the very high energy part of the spectrum (E>50 GeV). However, hadronic induced showers represent a significant background for the telescopes, therefore powerful methods of γ-hadron separations are needed to fully utilize the data taken with Cherenkov telescopes. Multivariate classification methods are commonly used for that. To be able to use other classification and advanced data mining methods, a RapidMiner environment is being implemented into standard analysis of MAGIC data. This will lead to an improvement of sensitivity and more efficient exploitation of data.

1 Introduction

MAGIC (Major Atmospheric Gamma-Ray Imaging Cherenkov Telescopes) is a system of two 17 m Cherenkov telescopes located at the Canary Island of La Palma built for observations in very high energy γ-ray astronomy. With the lowest energy threshold of all Cherenkov telescopes (50 GeV) and a high sensitivity [6], MAGIC is well aimed for detailed studies of γ-sources in energy region between 50 and 300 GeV, particularly for the study of blazars. Blazars are a subclass of active galactic nuclei which make up a largest group of extragalactic emitters at GeV/TeV energies. They are highly variable at nearly all wavelengths and their spectra are dominated by non-thermal emission [13].

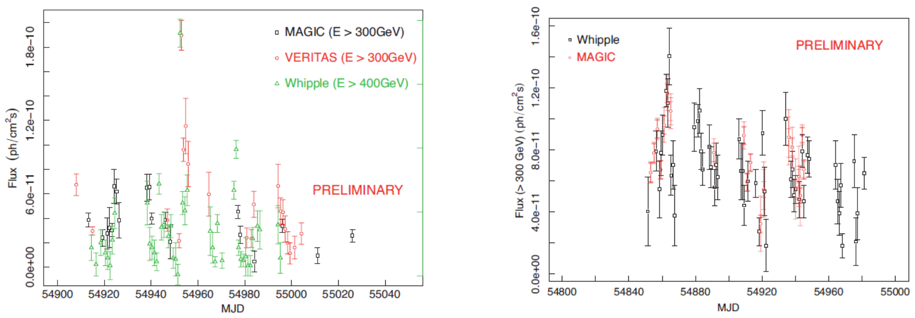


Figure 1: Very high energy lightcurve of Mrk 501 (left) and Mrk 421 (right) [2]

The spectra consist of two distinct broad components: low-energy bump originating from synchrotron emission of electrons and high-energy bump caused either by Inverse Compton scattering of synchrotron photons or from accelerated protons and ions. With the MAGIC telescopes it is possible to resolve the high-energy bump, to achieve an excellent coverage of the very high energy part of the spectrum and to study blazar variability in this energy region. Combined with data from other wavelengths, intrinsic properties of blazars can be studied, such as the structure of blazar emission zone [9] and their relativistic jets. This has been achieved in a monitoring campaign in 2009 on two near and bright active galaxies (blazars) Markarian 501 and Markarian 421 [2–4]. The unprecedented coverage of their spectra was achieved with quasi-daily observations of instruments from radio to TeV wavelengths. The resulting spectra provided a detailed insight into the physical properties of the both sources. Additionally, the lightcurves enabled a quantification and characterisation of the blazar variability (see figure 1 and [2]) .

2 The MAGIC data analysis

One of the most important problems of the data analysis of MAGIC is the discrimination between the signal and the background. The ratio of signal (recorded γ-photons) to hadronic background is 1:1000. Therefore a powerful background rejection method is needed to effectively use the observational results. The analysis of data largely relies on Monte Carlo (MC) simulations, which describe the development of the showers in the atmosphere, the Cherenkov light production, propagation through the atmosphere and subsequently the detector response. The shower image in the camera is usually parameterised with the set of Hillas parameters [11], which with some other additional parameters describe the charge distribution in the image. These parameters are used for γ-hadron-separation with a Random Forest (RF) method [5, 7], which is trained with MC simulations and so called OFF data (data, consisting mostly of hadron events, taken near the source but without it in the field of view) and is applied to data to analyse. This leads to a parameter called "hadronness", which in practice denotes a probability that an event is produced by a hadronic shower. The MAGIC data analysis is done inside the C++ framework called MARS [8], based on a C++ framework ROOT [1].

3 The implementation of RapidMiner into MAGIC data analysis

For the improvement of γ-hadron separation, which leads to a better sensitivity, other data mining methods can be used. This is particularly important in the energy region between 50 and 300 GeV, where the hadron events are much more difficult to distinguish from γ-events, enabling the deep spectral studies of blazar at those energies. A very good environment is offered by RapidMiner [12], a data mining tool which offers easy handling of different operators (such as sampling or attribute selection) and data mining methods (RF, neural networks, support vector machines). However, RapidMiner is Java based and MAGIC data are stored in a ROOT format. Therefore, an essential step in the analysis procedure is a corresponding ROOT-to-RapidMiner import operator, which is followed by a training and application of a selected learner on data and a RapidMiner-to-ROOT export operator. However, an adequately flexible import operator is still being developed and currently there are two approaches: converting ROOT data into ASCII format or using SQL database as back end for data storage (if the continuous read-out is needed e.g. for boosting) [10]. First preliminary results of use of RapidMiner RF show some improvements in the performance of the γ-hadron separation.

4 Conclusion and outlook

To enable deep studies of spectra of γ-emitting objects, such as blazars, it is necessary to improve γ-hadron separation. RapidMiner is a powerful data mining tool being implemented into analysis of Cherenkov telescopes data. Important parts of the analysis procedure are the import operator of MAGIC data in ROOT format into a RapidMiner-readable format and the export operator from RapidMiner to MARS. First results of attribute selection and application of RapidMiner Random Forest are promising. Further improvements can be achieved by the use of other features and data mining methods, such as support vector machines or neural networks.

References

[1] R Brun and F Rademakers. ROOT - An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research A*, 389:81 – 86, 1997.

[2] U Barres de Almeida, D Paneque, N Nowak, N Strah, and D Tescaro for MAGIC collaboration. Multifrequency variability and correlations from extensive observing campaigns of Mkn 421 and Mkn 501 in 2009. *Proceedings of 32nd International Cosmic Ray Conference, Beijing 2011*, 2011.

[3] A A Abdo et al. Fermi Large Area Telescope observations of Markarian 421: The missing piece of its spectral energy distribution. *Astrophysical Journal*, 736:131, 2011.

[4] A A Abdo et al. Insights into the high-energy γ-ray emission of Markarian 501 from extensive multifrequency observations in the Fermi era. *Astrophysical Journal*, 727:129, 2011.

[5] J Albert et al. Implementation of the Random Forest method for the Imaging Atmospheric Cherenkov Telescope MAGIC. *Nuclear Instruments and Methods in Physics Research Section A*, 588,3:424–432, 2008.

[6] J Aleksic et al. Performance of the MAGIC stereo system obtained with Crab Nebula data. *Astroparticle Physics*, 2011, submitted.

[7] L Breimann et al. Classification and regression trees. *Wadsworth*, 1983.

[8] A Moralejo et al. for the MAGIC collaboration. MARS, the MAGIC Analysis and Reconstruction Software. *Proceedings of 31th International Cosmic ray Conference, Łodz 2009*, 2009.

[9] G Ghisellini F Tavecchio, L Maraschi. Constraints on the physical parameters of TeV Blazars. *Astrophysical Journal*, 509, 2:608–619, 1998.

[10] M Helf. Gamma-Hadron-Separation im MAGIC-Experiment durch verteilungsgestütztes Sampling. *Diplomarbeit, Technische Universität Dortmund*, April 2011.

[11] A M Hillas. Cherenkov light images of EAS produced by primary gamma. *Proceedings of 19th International Cosmic ray Conference*, 1985.

[12] I Mierswa, M Wurst, R Klinkenberg, M Scholz, and T Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.

[13] C M Urry and P Padovani. Unified Schemes for Radio-Loud Active Galactic Nuclei. *Publications of the Astronomical Society of the Pacific*, 107:803, 1995.

Unfolding for Stacked Neutrino Sources with the IceCube Detector

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

IceCube is a cubic kilometre scale neutrino detector located at the geographic South Pole. Its construction of 86 strings was finished in the austral summer 2010/2011. IceCube is the most sensitive telescope for high energy neutrinos.

Because the neutrino flux of single sources might be to low to be measured, this analysis uses the stacking method with an unbinned likelihood method. Neutrino energy spectra for different source catalogues will be unfolded with the use of a new error estimation method using bootstrapping, developed in the collaborative research center 823. The used data was taken in 2009 when IceCube consisted of 59 strings.

1 Stacking

The stacking method treats multiple sources as one to increase the measured flux [8]. The sources selected for stacking are collected in different catalogues, each catalogue representing a different source type. Because the signal events add up faster than the background events, a stacking analysis is more sensitive to a discovery than a single source analysis, although one could only claim a discovery for a certain catalogue and not for an individual source. For this analysis an unbinned likelihood method will be used [5].

1.1 Catalogues

In a source catalogue multiple objects are collected which share a common pattern. One famous example is the Messier catalogue listing astronomical objects which resembled comets but were not.

Interesting catalogues for this work are catalogues used in previous stacking analyses e.g.

- Starburst Galaxies [3]
- TeV Milagro sources [2]
- Multiple Fermi catalogues [1]
- Supernova remnants with nearby molecular clouds [6]
- CSS/GPS catalogue [12]

including some updates. As well as new catalogues like massive binary systems based on the Einstein catalogue [13].

2 Unfolding

The energy reconstruction is obtained with an unfolding algorithm introduced in the software RUN [4]. The program used for the unfolding is an enhanced version of RUN written in C++ named TRUEE [11] [10].

Up to three different variables can be used for the unfolding. These variables should have a good correlation to the target variable, in this case the energy.
A MRMR algorithm [7] implemented in the feature selection extension [14] for RapidMiner [9] is used to narrow down the available variables to ten. These ten variables were than further investigated with the available functionalities in TRUEE.

An unfolding result retrieved by unfolding only 500 monte carlo events is shown in Figure 2. Usual unfolding analyses in IceCube use approx. 40 000 Events and are therefore able to produce smaller bins and cover a larger energy range. The offset between the theoretically predicted Bartol flux and the measured data points results from the low number of events used for the unfolding. Therefore, only the slope is comparable. The slope from the data fit $\gamma = -3.0$ doesn't rejct the null hypothesis $H0 : \gamma = -3.4$ within a 5σ confidence interval.

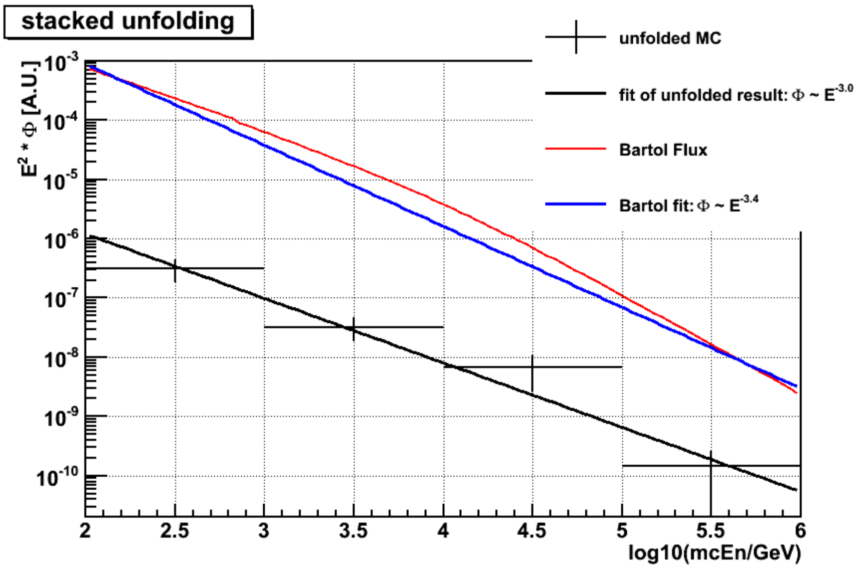


Figure 1: The unfolding result of 500 monte carlo events is presented in black, as well as the corresponding fit. The distribution is proportional to a power law with a slope $\gamma = -3.0$. The theoretical prediction is shown in red the corresponding power law fit in blue is proportional to a slope $\gamma = -3.4$. The offset results from the low number of unfolded events.

References

[1] A. A. Abdo et al. Bright Active Galactic Nuclei Source List from the First Three Months of the Fermi Large Area Telescope All-Sky Survey. *Astrophysical Journal*, 700:597–622, July 2009.

[2] A. A. Abdo et al. Milagro Observations of Multi-TeV Emission from Galactic Sources in the Fermi Bright Source List. *Astrophysical Journal Letters*, 700:L127–L131, August 2009.

[3] J. K. Becker, P. L. Biermann, J. Dreyer, and T. M. Kneiske. Cosmic Rays VI - Starburst galaxies at multiwavelengths. *ArXiv e-prints*, January 2009.

[4] V. Blobel. An Unfolding Method for High Energy Physics Experiments. *ArXiv High Energy Physics - Experiment e-prints*, August 2002.

[5] J. Braun, J. Dumm, F. de Palma, C. Finley, A. Karle, and T. Montaruli. Methods for point source analysis in high energy neutrino telescopes. *Astroparticle Physics*, 29:299–305, May 2008.

[6] V. Cavasinni, D. Grasso, and L. Maccione. TeV neutrinos from supernova remnants embedded in giant molecular clouds. *Astroparticle Physics*, 26:41–49, August 2006.

[7] Chris H. Q. Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. In *2nd IEEE Computer Society Bioinformatics Conference (CSB 2003), 11-14 August 2003, Stanford, CA, USA*, pages 523–529. IEEE Computer Society, 2003.

[8] Andreas Gross. *Search for High Energy Neutrinos from AGN classes with AMANDA-II*. PhD thesis, Universität Dortmund, February 2006.

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

[10] N Milke, M Doert, and W Rhode. Solving time-dependend inverse problems with truee: Examples in astroparticle physics. *Paper in progress*.

[11] N Milke, W Rhode, and T Ruhe. Studies on the unfolding of the atmospheric neutrino spectrum with icecube 59 using the truee algorithm. *IceCube Collaboration Contributions to the 2011 International Cosmic Ray Conference*.

[12] C. P. O'Dea. The Compact Steep-Spectrum and Gigahertz Peaked-Spectrum Radio Sources. *The Publications of the Astronomical Society of the Pacific*, 110:493–532, May 1998.

[13] A. M. T. Pollock. The Einstein view of the Wolf-Rayet stars. *Astrophysical Journal*, 320:283–295, September 1987.

[14] Benjamin Schowe and Katharina Morik. Fast-ensembles of minimum redundancy feature selection. In Oleg Okun, Giorgio Valentini, and Matteo Re, editors, *Workshop on Supervised and Unsupervised Ensemble Methods and their Applications - SUEMA 2010*, 2010.

MAGIC Analysis Frameworks

Malwina Uellenbeck
Lehrstuhl für Astroteilchenphysik E5b
Technische Universität Dortmund
malwina.uellenbeck@tu-dortmund.de

This report is a short summary of the study of very high energy (VHE) γ-radiation from active galactic nuclei (AGN) using the **M**ajor **A**tmospheric **G**amma-ray **I**maging **C**herenkov (MAGIC) telescopes on the Canary Island, La Palma. A significant part of this report is devoted to the development and improvement of analysis tool and analysis of MAGIC data taken on AGNs.

1 High-energy γ-rays and Blazars

High-energy γ-ray above an energy of some GeV cannot reasonable be observed by satellite detectors with their too small detection areas. In this very high energy (VHE) γ-ray domain, since the late 1980s, ground-based Cherenkov telescopes have proved to be very successful in energy range between 50GeV and 50TeV, adding significant information to the understanding of and for modeling galactic and extragalactic γ-ray sources and emission mechanism [7]. One of these most interesting extragalactic γ-ray sources are Blazars. Blazars belong to the class of AGN and are characterized by relativistic jets oriented toward the Earth. They are also characterized through a continuous **S**pectral **E**nergy **D**istribution (SED) with no or weak emission lines and two broad humps (the first one in the UV to soft X-ray and a the seconde one in the GeV range). Beside this their flux was found to be variable at all observed frequencies and in time scales ranging from minutes to years [3].

All these interesting results in the γ-ray physics require a high developed hardware and also software environment. The analysis software of the MAGIC experiment, which provide the "high level" results, in particular the flux and the light curve determination, is still under development and improvement process. One of the most important tasks in the improvement of the analysis software are the efficient exploitation of data (e. g. running

Monte-Carlo Simulations on GPUs) and the improvement of the sensitivity with advanced data mining methods and tool like RapidMiner [5].

2 The MAGIC analysis chain

In general there are three primary goals in every analysis for an **I**maging **A**tmospheric **C**herenkov **T**echnique (IACT) experiment:

1. Verification between γ-like and hadron-like events, using tools like Random Forest for the so-called γ-hadron separation
2. Determination of the primary γ-ray energy of the γ-like events, which allow to drive an energy spectrum of a detected γ-ray source like the AGNs
3. Proper determination of the incoming direction of the γ-like events from a γ-emitter. If the position of the γ-emitter is known and its dimension is a point source for the telescope, then the direction information can be used for better γ-hadron separation and a more defined energy determination [4].

Currently, the MAGIC analyzers are using the software package called MARS (**M**AGIC **A**nalysis and **R**econstruction **S**oftware), which is written in C++ language and is based on the ROOT framework [1]. This software package was developed to cover all signal processing steps and providing robust tools starting from the reading of the uncompressed raw data and extending to the calculation of light curves and energy spectra of a dedicated γ-ray source [4].

The main steps of the MARS analysis software are as follows and illustrated in Fig. 1:

- The calibration of FADC information for each pixel into number of photoelectrons (ph.e.) which includes also the timing information of the signal (**Callsito**)
- Pixels of each telescope MAGIC I and MAGIC II, which contain noise were removed by the image cleaning method and the calculation of image parameters using survived pixels is conducted (**Star**)
- Merges MAGIC I and MAGIC II Star files into a single stereo files. At the same time, it also performs some trigonometric calculations to calculate the stereoscopic parameters (**Supertstar**)
- The Training of γ-hadron matrices and training of energy estimation matrices is accomplished. For the first one, a subsample of Monte-Carlo simulated γ-events is used vs. a subsample of background events from real data. For the energy estimation only a subsample of Monte-Carlo simulated γ-events is used (**Osteria**)

[1] An object oriented data analysis framework, http://root.cern.ch

- The output of Osteria, calculated γ-hadron and energy estimated matrices, are applied to a test sample of Monte-Carlo γ-events, to background data and also to signal data (**Melibea**)
- The γ-hadron separation cuts are applied to the data and the number of excess events is calculated. Beside this the effective on-time can be determined from a dedicated data sample and the effective area from the corresponding Monte-Carlo simulated γ-sample. Furthermore the spectrum and light curves are determined in corresponding bins of estimated energy (**Fluxlc**)
- The energy spectrum in bins of true energy is obtained. Here the energy spectrum is unfolded taking into account the energy resolution of MAGIC and other analysis effects (**Unfolding**)

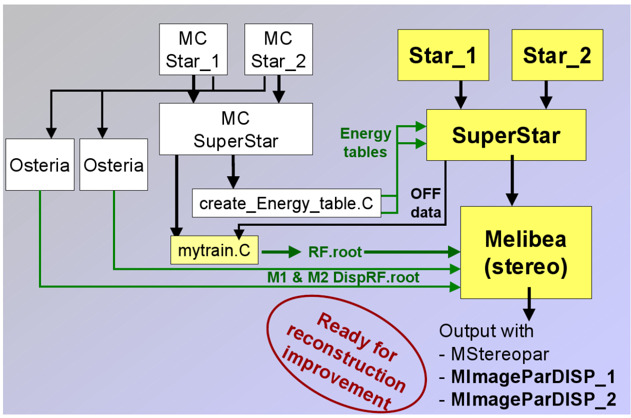
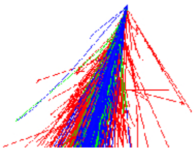


Figure 1: Analysis chain for data of the MAGIC telescopes starting from Star level [1]

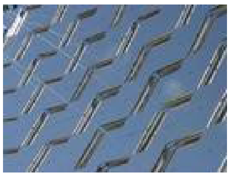
As shown above Monte-Carlo (MC) simulations play an important role at several steps in the analysis chain. The most disadvantage of these simulations is that the production of such data is very CPU time consuming and therefore more efficient algorithm and GPU environment are needed.For the Monte-Carlo simulation of MAGIC telescopes data also here a chain of programs is used.

3 The MAGIC Monte-Carlo simulation chain

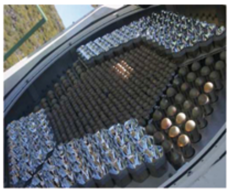
The Monte-Carlo simulation programs for the MAGIC telescopes are divided into three different steps.



`In step 1` the development of γ-ray initiated air shower is simulated with the software called **CORSIKA [2]**. Cherenkov photons arriving on the ground around the telescopes location are stored in binary files containing all the relevant parameters.



`In step 2` the information about the Cherenkov photons coming from the simulated air showers (CORSIKA) is passed to the program called **REFLECTOR**. Furthermore this program calculates then the reflection of the surviving photons on the mirror dish to obtain their location and arrival time on the camera plane.



`In step 3`, the last step, the program called **CAMERA** simulates the behavior of the MAGIC photomultiplier camera, trigger system and also the data acquisition system. In this program also realistic pulse shapes, noise levels and gain fluctuations obtain from the real MAGIC data have been implemented into the simulation chain.

4 Conclusion and Outlook

As mentioned before, further improvements of the MAGIC analysis and Monte-Carlo chains are possible and can be achieved by the use of other features and data mining tools, such as GPUs, RapidMiner [5] and TRUEE [6].

References

[1] P. Collin. Current status of osteria/melibea for stereo-data analysis. Talk on the MAGIC Stereo-analysis Workshop in Dortmund, January 2010.

[2] D. Heck. Extensive Air Shower Simulations with CORSIKA and the Influence of High-Energy Hadronic Interaction Models. *ArXiv Astrophysics e-prints*, March 2001.

[3] C. C. Hsu, K. Satalecka, M. Thom, M. Backes, E. Bernardini, G. Bonnoli, N. Galante, F. Goebel, E. Lindfors, P. Majumdar, A. Stamerra, and R. M. Wagner. Monitoring of bright blazars with MAGIC telescope. *ArXiv e-prints*, July 2009.

[4] Daniel Mazin. *A study of very high energy gamma-ray emission from AGNs and constraints on the extragalactic background light*. PhD thesis, December 2007.

[5] Ingo Mierswa, Martin Scholz, Ralf Klinkenberg, Michael Wurst, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In *In Proceedings of the 12th ACM SIGKDD International PONZETTO and STRUBE Conference on Knowledge Discovery and Data Mining*, pages 935–940. ACM Press, 2006.

[6] N Milke, M Doert, and W Rhode. Solving time-dependend inverse problems with truee: Examples in astroparticle physics. *Paper in progress*.

[7] Robert Marcus Wagner. *Measurement of very high energy gamma–ray emission from four blazars using the MAGIC telescope and a comparative blazar study*. PhD thesis, November 2006.

Combined application of classification and spectral reconstruction in the IceCube analysis

Natalie Milke
Lehrstuhl für Experimentelle Physik 5b
Technische Universität Dortmund
natalie.milke@tu-dortmund.de

The measurement of the atmospheric neutrino energy spectrum provides information about the diffuse neutrino flux from extragalactic sources. A relative increase of the spectrum toward higher energies could be evidence for neutrino producing hadronic processes in the cosmic high energy accelerators, such as Active Galactic Nuclei or Gamma Ray Bursts. IceCube is a cubic kilometer large neutrino detector located in the glacial ice at the geographic South Pole. IceCube permits the detection of neutrinos with energies beyond 10^6 GeV. To obtain the final neutrino energy spectrum a classification algorithm is applied on the IceCube data to select interesting events, followed by an unfolding algorithm to estimate neutrino spectrum using the measured attributes. In the current work the multivariate method Random Forest is used to separate neutrino events from the background events. For the spectrum estimation the new unfolding algorithm TRUEE is used, developed within the Collaborative Research Center SFB 823.

Neutrinos from interactions of cosmic rays with the Earth's atmosphere represent a background for the extragalactic neutrinos. Thus, a precise measurement of the atmospheric neutrino flux is important for understanding this background. Since the spectral index of the flux distribution depending on neutrino energy is lower for extragalactic neutrinos (following the spectral behavior of Fermi accelerated cosmic rays $\gamma \sim 2$ [5]) than for atmospheric neutrinos ($\gamma \sim 3.7$ [7]), a contribution of extragalactic neutrinos would cause an enhancement of the flux in the high energy region of the spectrum.

The detection of extragalactic neutrinos for understanding of cosmic ray production in cosmic accelerators is one of the main goals of IceCube [6]. IceCube is a cubic kilometer large neutrino detector located at the geographic South Pole. It consists of 5160 digital optical modules (DOM) arranged along 86 strings in the depth between 1450 and 2450 m in the glacial ice. While traveling through the ice the high energy neutrino-induced muons produce Cherenkov light which can be detected by the DOMs providing directional and energy information of the muon track and thus also of the primary neutrino. Even during its construction, the partially built IceCube was able to take data. In this paper the atmospheric neutrino sample from the measurement with the IceCube 59 (IC 59) string configuration is used.

During the measurement not only the relevant neutrino-induced events are detected. Despite the big amount of glacial ice above the detector the background events from atmospheric muons exceed the signal events by several magnitudes. Although the cut on the zenith distribution reduces significantly the background, selecting only events coming from below the horizon and thus traveling through the Earth, there are still coincident atmospheric muon events misreconstructed as upgoing neutrino-induced events. Therefore the application of a multivariate method is necessary to classify events in the measured data and select a high purity sample of the interesting events. We use the Weka-Random Forest [3] included in the framework RapidMiner [8]. For a detailed description of application of Weka-Random Forest in the IceCube analysis see also the report of Tim Ruhe.

For the training of the algorithm and the subsequent determination of the detector response matrix in the unfolding analysis the Monte Carlo simulation of the signal neutrino-induced events and the background atmospheric muon events is used. The complete event development from the particle generation, interaction probabilities, propagation and signals in detector are considered in the simulation to describe the events realistically.

The classification of measured data allowed to obtain a sample with the purity of minimum 95 %. Thus, the amount of remaining misreconstructed background events is neglegible for the following energy spectrum determination.

The energy of the primary particles is convoluted with the interaction probability and the detector finite resolution and limited acceptance. Therefore the neutrino energy has to be estimated from energy-correlated, measured observables. For this purpose a regularized unfolding algorithm TRUEE [9] is developed and applied. TRUEE - Time-dependent Regularized Unfolding for Economics and Engineering problems is written in C++ using ROOT [4] and is based on the Regularized UNfolding ($\mathcal{RUN}$) algorithm [2] [1] written in FORTRAN 77. The new software contains additional user-friendly functions and is easy to install and use in combination with modern software. In the following examples for the application of the main new functions and the unfolding analysis of the IC 59 data are introduced.

The measured observables in the detector and other calculated characteristic attributes have different dependency on the neutrino energy. The user has to choose which observables can be taken for the unfolding. In TRUEE three different attributes can be used for the unfolding at the same time. TRUEE provides automatically correlation plots, which demonstrate the dependency of an attribute on the primary energy. By observation of these scatter plots and the corresponding profile plots the user decides which variables can be used for the unfolding (see Fig. 1).

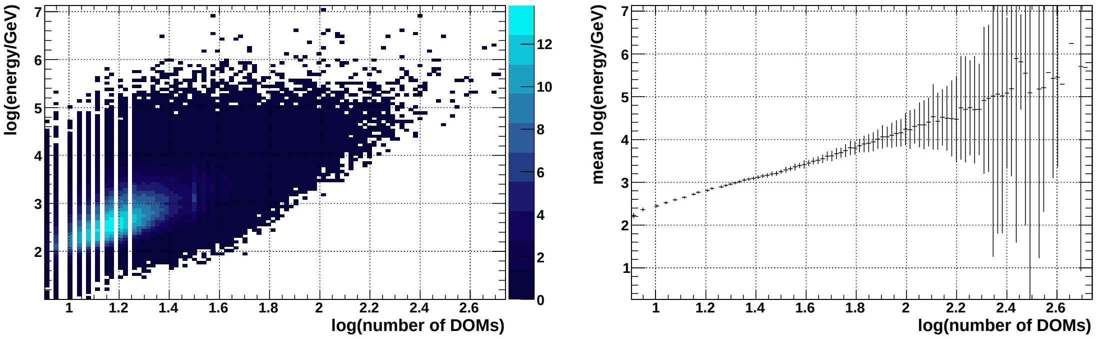


Figure 1: Scatter plot and related profile histogram to check the correlation between the event energy and the measured attribute. An optimal correlation is presented by a monotonically changing profile function with small uncertainties. Here as an example the IceCube observable *number of DOMs* is shown.

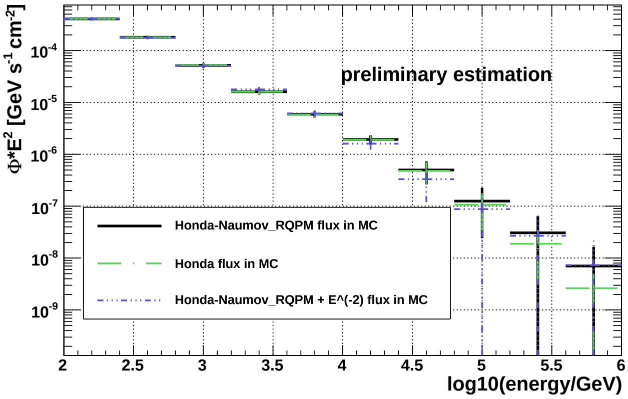


Figure 2: An example of the atmospheric neutrino energy spectrum from 10 % of IC 59 data unfolded with TRUEE. Three unfolding results are shown, using different simulated distributions to determine the detector response. The uncertainties are calculated by the software from the covariance matrix considering the error propagation. The spectrum is weighted by squared energy for a better illustration.

Further the performance of the unfolding is checked using the new test mode. A fraction of simulated events is used as data sample. In this case the unfolding result can be compared to the true distribution and settings of the algorithm can be chosen.

An example of the preliminary estimation of the atmospheric neutrino energy spectrum is shown in Fig. 2 detemined from 10 % of the IC 59 data. The presented analysis chain is an exemplary procedure, that is being optimized for the final determination of the neutrino energy spectrum of the full year data. The introduced analysis was presented at the International Cosmic Ray Conference 2011 in Beijing [10].

References

[1] V Blobel. The run manual: regularized unfolding for high-energy physics experiments. *OPAL Technical Note TN361*, May 1996.

[2] V Blobel. An unfolding method for high energy physics experiments. *arXiv:hep-ex/0208022v1*, Jan 2002.

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

[4] R Brun and F Rademakers. Root - an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research A*, 389(1 - 2):81 – 86, 1997.

[5] E Fermi. On the origin of the cosmic radiation. *Phys. Rev.*, 75:1169 – 1174, 1949.

[6] F Halzen and S R Klein. Icecube: An instrument for neutrino astronomy. *arXiv:1007.1247v2*, 2010.

[7] M Honda and et al. Calculation of atmospheric neutrino flux using the interaction model calibrated with atmospheric muon data. *Phys. Rev. D*, 75(4):26, 2007.

[8] I Mierswa, M Wurst, R Klinkenberg, M Scholz, and T Euler. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.

[9] N Milke, M Doert, and W Rhode. Solving time-dependend inverse problems with truee: Examples in astroparticle physics. *Paper in prep.*

[10] N Milke, W Rhode, and T Ruhe. Studies on the unfolding of the atmospheric neutrino spectrum with icecube 59 using the truee algorithm. *IceCube Collaboration Contributions to the 2011 International Cosmic Ray Conference.*

Data Mining for IceCube with RapidMiner

Tim Ruhe
Experimentelle Physik 5
Technische Universität Dortmund
tim.ruhe@tu-dortmund.de

IceCube is a 1 km^3 neutrino telescope located at the geographic South Pole. The large number of reconstructed attributes as well as the small signal to background ratio in the search for atmospheric neutrinos makes IceCube well suited for a detailed study within the scope of machine learning. In our study a systematic feature selection using the MRMR-algorithm was carried out. Furthermore, a Random Forest was trained and tested in order to improve the event selection of the IceCube atmospheric neutrino analysis. Finally the forest was applied on data. A good agreement between Monte Carlo expectations and data was observed.

The IceCube neutrino telescope [1] was completed in December 2010 at the geographic South Pole. There are 5160 Digital Optical Modules (DOMs) mounted on 86 vertical cables (strings) forming a three dimensional array of photosensors. The spatial distance between individual strings is 125 m. IceCube strings are buried at depths between 1450 m and 2450 m corresponding to an instrumented volume of 1 km^3. The spacing of individual DOMs on a string is 17 m [1, 2, 8].

Atmospheric neutrinos are produced in extended air showers where cosmic rays interact with nuclei of the Earth's atmosphere. Within these interactions mainly pions and kaons are produced which then subsequently decay into muons and neutrinos [5].

The measurement of the atmospheric neutrino spectrum, however, is hindered by a dominant background of atmospheric muons. A rejection of atmospheric muons can be achieved by selecting upward going tracks only since the Earth is opaque to muons. However, a small fraction of atmospheric muons is still misreconstructed as upward going. For the starting point of this analysis (the so called Level 3) where many advanced reconstruction algorithms have already been run and the dominant part of the atmospheric

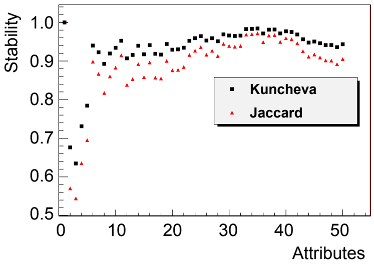


Figure 1: Stability estimation for the MRMR Feature Selection depicting the Jaccard and Kuncheva's index. The stability of the feature selection goes into saturation as the number of attributes increases. For a number of attributes ≥ 20 both stability measures lie well above 0.9. One should note that both indices reach the maximum of 1.0 if only one attribute is selected indicating that there is one single best attribute for the separation of signal and background.

muons has already been removed, we expect $N_{back} \approx 9.699 \times 10^6$ background events and $N_{sig} \approx 1.5788 \times 10^4$ signal events in 33.28 days of IceCube in the 59-string configuration. This corresponds to a signal to background ratio of $R = 1.63 \times 10^{-3}$. Approximately 2600 reconstructed attributes where available at Level 3.

The low signal to background ratio in combination with the large number of attributes available at Level 3 makes this task well suited for a detailed study within the scope of machine learning.

Prior to our studies precuts where applied on $v_{LineFit} > 0.19$ and $\theta_{Zenith} > 88^\circ$ in order to further reject the muonic background. Furthermore, we reduced the number of attributes entering our final attribute selection by hand exluding attributes that were known to be useless, redundant or a source of a potential bias. This preselection of attributes reduced the number of attributes entering the final selection to 477.

A Maximum Relevance Minimum Redundancy (MRMR) [3,9] algorithm embedded within the Feature Selection Extension [10] for RapidMiner [7] was used for feature selection. Simulated events from Corsika [4] were used as background. Simulated events from the IceCube neutrino generator Nugen were used as signal. The machine learning environment RapidMiner [7] was used throughout the study.

Figure 1 depicts the stability of MRMR as a function of the number of attributes considered. The Feature Selection Stability Validation, also included in the Feature Selection Extension for RapidMiner, was used to estimate the stability.

The Jaccard index is depicted by triangles, whereas squares represent Kuncheva's index [6].

Figure 1 clearly shows that MRMR can be considered stable on IceCube Monte Carlo simulations if the considered number of attributes in the selection is $n_{Attributes} \geq 20$.

Figure 2 shows the output of the Random Forest after a 5-fold cross validation. Within this cross validation 3.8×10^5 simulated background and 7×10^4 simulated signal events

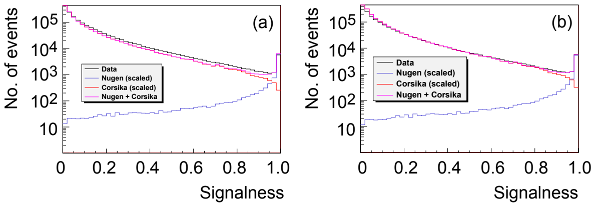


Figure 2: Random Forest score (signalness) for simulated signal and background events as well as for data. When the total number of MC events is scaled to match data in the absolute number of events a data/MC mismatch is observed for signalness > 0.2 (a). When the total number of MC however, is scaled to match the data for signalness> 0.2 a mismatch is only observed for small signalness values.

were used. In order to avoid overtraining a sampling was performed prior to each training of the forest in the cross validation which limited the number of events used for training to 2.8×10^4 for each class. The number of trees in the forest was chosen to $n_{trees} = 500$. From figure 2 a data/MC mismatch for a signalness $s \geq 0.2$ is observed, which would in turn lead to an underestimation of the remaining muonic background. To achieve a realistic background estimate the Corsika events are rescaled by a factor of 1.23 such that they match the distribution of data for $s \geq 0.2$. This leads to a data/MC mismatch only in the low signalness region.

Due to the small errorbars on the expected number of signal and background events for individual cuts the performance of the forest can be considered stable (see table 1). No indications of overtraining were observed within the cross validation. Note however that the large errorbars on the number of expected background events are due to small statistics when cuts in the high signalness regions are applied.

The performance of the forest on data lies within the range expected from the cross validation. Only for a signalness cut of $s = 1.0$ one finds an underfluctuation of 96 % of the expected number of events.

The last two columns of table 1 show the expected purity of the final neutrino sample under the assumption that the number of background events is as expected from the mean of the cross validation (column 6) and as a worst case scenario (column 7). For the worst case scenario the purity was computed using the upper limit of the errorbar for the expected number of background events. One finds that in both cases a purity well above 95 % can be achieved for the cuts listed in table 1.

References

[1] J. Ahrens *et al.* Sensitivity of the IceCube detector to astrophysical sources of high energy muon neutrinos, Astropart. Phys. **20** (2004)

Cut	Est. Back. Ev.	Est. Sig. Ev.	Sum	Data	Est. Pur.[%]	Pur. Pess. [%]
0.990	114 ± 57	4817 ± 44	4931 ± 64	4988	97.7	96.7
0.992	98 ± 37	4633 ± 43	4731 ± 57	4757	97.9	97.1
0.994	71 ± 37	4414 ± 41	4485 ± 55	4476	98.4	97.6
0.996	60 ± 32	4122 ± 32	4182 ± 45	4134	98.5	97.8
0.998	22 ± 20	3695 ± 44	3717 ± 50	3638	99.4	98.8
1.000	5 ± 11	2932 ± 33	2937 ± 35	2833	99.8	99.4

Table 1: Estimated number of signal and background as well as the estimated purity after an application of cuts on the signalness. The number of data events yielded for individual cuts is shown as well.

[2] T. DeYoung, Neutrino Astronomy with IceCube, Modern Physics Letters A, Vol. 24, Iss. 20 (2009)

[3] C. H. Q. Ding and Hanchuan Peng, Minimum Redundancy Feature Selection from Microarray Gene Expression Data, 2nd IEEE Computer Society Bioninformatics Conference (CSB 2003) (2003)

[4] D. Heck, CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers, Forschungszentrum Karlsruhe Report RZKA 6019 (1998)

[5] M. Honda *et al.*, Calculation of the flux of atmospheric neutrinos, Phys. Rev. D **52**,9 (1995)

[6] L.I. Kuncheva, A stability index for feature selection, Proceedings of the 25th IASTED International Multi-Conference (2007)

[7] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, YALE: Rapid Prototyping for Complex Data Mining Tasks, KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (2006)

[8] E. Resconi, Status and prospects of the IceCube neutrino telescope, Nucl. Instr. and Meth. A **602**, 7 (2009)

[9] B. Schowe and K. Morik, Fast-Ensembles of Minimum Redundancy Feature Selection, Workshop on Supervised and Unsupervised Ensemble Methods and their Applications - SUEMA 2010 (2010)

[10] B. Schowe, http://sourceforge.net/projects/rm-featselext (2011)

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

Katja Ickstadt Christian Sohler

Ideas on efficient Bayesian analysis

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

The aim of this work is to find efficient MCMC algorithms for data sets with a very large number of observations by combining ideas from the methods *merge and reduce* and *meta-analysis*. This report contains short descriptions of *merge and reduce* and *meta-analysis* as well as the outline of two approaches to solve the problem.

Project C4 deals with regression analysis for very large datasets. Very large can mean a very large number of observations n, a very large number of variables p or both. Here we focus on the case where the number of observations is large while the number of variables is small to moderate.

We want to analyse data sets with large n using Bayesian regression analysis. Bayesian statistics have seen a rise in popularity in recent years, mainly due to increased computer power. Bayesian methods allow to incorporate prior knowledge about some or all of the parameters in the model. The absence of prior knowledge can also be dealt with. In addition, Bayesian methods are suitable for modelling complex hierarchical systems.

The aim of Bayesian analyses is to obtain the posterior distribution of the parameters of interest. This is a computationally demanding task in all but the simplest cases. Usually, Markov Chain Monte Carlo methods are employed. 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. On large data sets the computational cost and the necessary memory become prohibitive.

We aim to solve this problem using a method called *merge and reduce*. *Merge and reduce* is well-known to Computer Scientists [1]. It has been used for a number of applications with very large amounts of data including clustering and classification problems. The data set is divided into b different blocks. Ideally, $b = 2^x$ with $x \in \mathbb{N}$. We go through every block, starting with the first. In most cases the data are deleted immediately after the model is ready. Two blocks on the same hierarchical level are combined as soon as possible. This way, the number of models that need to be saved is minimal. Figure 1 illustrates the principle of *merge and reduce* using 8 blocks as an example.

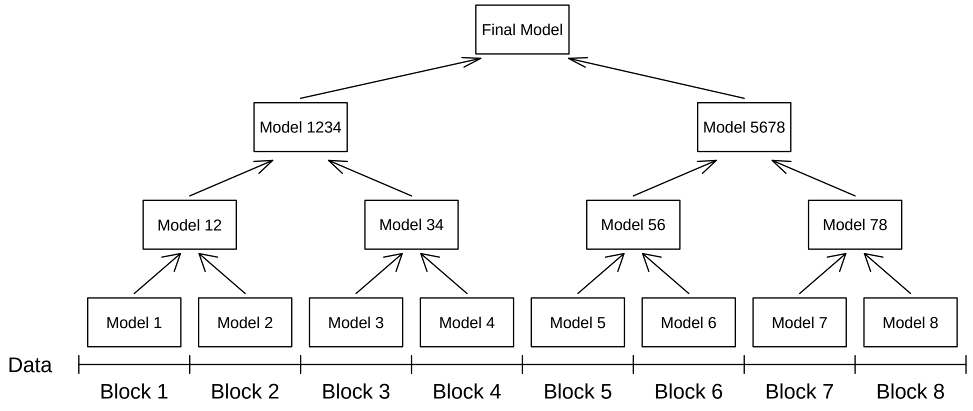


Figure 1: Illustration of *merge and reduce*

When using *merge and reduce* it is crucial to ensure that models do not become more complex as they are merged – hence the name. This is even more important when dealing with resource-constrained data analysis and/or very large data sets. Merging the models is not trivial and depends heavily on the goal of the analysis and the question one is interested in.

At this point we want to distinguish between two different cases, a “summary value approach” and a “sample approach”. In both cases we obtain a sample from the posterior distribution using Markov Chain Monte Carlo-methods. In the first case the posterior distribution can (at least in principle) be characterised using only a few values such as the mean, the median, the standard deviation or certain quantiles. It is important to note that estimating those summary values does not allow conclusions about the probability density function. We do not know e.g. what kind of distribution the data come from. However, we do assume a few values suffice to carry out the analysis. For that reason, we only calculate certain values using the MCMC sample and delete it afterwards.

As a starting point in the “summary value approach” we only consider one variable m ($m \in \{1, \dots, n\}$). Let $X_1, \dots, X_n$ be independent and identically distributed random variables with expected value μ and variance σ^2. Let $x_1, \dots, x_n$ be the corresponding observed values. We split the random variables into b blocks $X_{1,1}, \dots, X_{1,n_1}, \dots, X_{b,1}, \dots, X_{b,n_b}$

where $j = 1, \dots, n_j$ is the size of block j. The blocks do not have to be of the same size. This split can also take place on the level of the observed values.

Our general idea is similar to ideas in *meta-analysis*. *Meta-analysis* is used in statistics to compare the results from different surveys and combine them to get an overall result. The classical random-effects and fixed-effects models (confer [2, 4]) utilise the observed variance per study to estimate the overall effect size. In our setting the variance may well be an interesting summary value, so it may not be possible to follow those approaches. However, in contrast to typical *meta-analysis* settings, we can assume that the variance is the same across all blocks. For that reason it seems plausible that we do not need to adjust for different variances between different studies.

We have tested summarising each block using the mean, the variance, the lower quartile, and the upper quartile, whose formulae are given below. The lower and upper quantile are $Q_{0.25}$ and $Q_{0.75}$ respectively.

$$\begin{aligned}
\bar{X}_j =& \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji} \\
S_j^2 =& \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ji} - \bar{X}_j) \\
Q_p(j) =& \begin{cases} \frac{1}{2}\left(X_{(pn_j)} + X_{(pn_j+1)}\right), & \text{if } pn_j \in \mathbb{N} \\ X_{(\lfloor pn_j \rfloor + 1)}, & \text{if } pn_j \notin \mathbb{N} \end{cases}
\end{aligned}$$

where $j = 1, \dots, b$. Other summary statistics can be chosen without problems. On the first layer this approach gives us $4 \times b$ summary values, 4 for each block. To combine blocks 1 and 2, say, we take the mean value of the respective summary value. This returns $V_{1,2} = \frac{1}{2}(V_1 + V_2)$ for any summary value V. On upper layers we combine the combined summary values in the same way.

This approach returns unbiased estimators for μ and σ^2 when every observation is given the same weight. If the number of observations is not equal in all the blocks, care must be taken to adjust the weights accordingly. The expected value for the quartiles is not included here as their probability density function is hard to handle analytically.

To be able to detect deviations from the assumption of independent and identically distributed random variables we calculate an additional value every time we merge two blocks. At the first layer we take the sum of the squared differences between each of the corresponding summary statistics. When dealing with summary statistics that result from merging two blocks (second and later layers) we also add the two sums of squared differences. This avoids errors cancelling each other out unnoticed and is in some sense a replacement for the classical effects-models. We use simulation studies to evaluate whether a sum of squared differences should be considered “high”. Preliminary results are very promising. With increasing sample sizes, deviations from the mean are identified easily. Higher or lower variances in some blocks are harder to detect, but do not pose a problem for sample sizes of 1'000 or more. We plan to find "critical values" for the sum of squared differences, which then can be interpreted similarly to a hypothesis test.

Future work needs to be done in the case of multivariate analyses. [3] contains suggestions for multivariate *meta-analysis*. We plan to adapt some of them for our problem.

In the second case – the "sample approach" – it is assumed that the distribution cannot be characterised by a few summary values only. When this occurs we plan to use a whole sample from the posterior distribution instead. The first step is still the same: Do an MCMC run on every block and keep everything after the burn-in period. After the first two blocks this gives us two samples of size m each. In the second step, we combine the two samples to a new one, giving us a total sample size of $2m$. This is an estimate of the underlying probability density function f.

It is not acceptable to just stop here as this would increase the complexity of the model exponentially. For that reason, we reduce the complexity in the third step. We sample m sample values from the total of $2m$ without replacement. The probability for every value to get sampled is according to f. We have to be careful not to unduly reduce the natural variation in the sample while doing this. Therefore the main idea is to remove values that are outliers or in other senses unusual.

Contrary to the "summary value approach" we also want to keep a proportion of the data. The number of observations should be substantially reduced, ideally the number of kept observations is proportional to the logarithm of the number originally in the block $n_{j,\text{keep}} \propto \ln(n_j)$. We plan to sample the observations according to the $2m$ sample values. The kept observations will help sampling good sample values in upper layers.

Major further work will concentrate on finding good ways of sampling the sample values and the observations to be kept in higher layers.

References

[1] Nicolas Bruno and Surajit Chaudhuri. Physical design refinement: The "merge-reduce" approach. In *Advances in Database Technology - EDBT 2006*, volume 3896 of *Lecture Notes in Computer Science*, pages 386–404. 2006.

[2] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[3] In-Sun Nam, Kerrie Mengersen, and Paul Garthwaite. Multivariate meta-analysis. *Statistics in Medicine*, 22:2309–2333, 2003.

[4] Barnet Woolf. On estimating the relationship between blood group and disease. *Annals of Human Genetics*, 19:251–253, 1955.

Approximations for Logistic Regression

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

The aim of this work is to develop algorithms for efficient training of logistic regression classifiers for large datasets. This report contains a description of the logistic regression model, an overview of techniques and algorithms used for similar problems in ℓ_2 and ℓ_1 regression, and finally a summary of our current state of research.

Logistic Regression

In classification problems, we deduce the most likely class of an object based on its properties. The simplest classification problem only considers two states (cancer – no cancer, spam – ham) and determines the most likely class for future data based on a set of objects each with attributes $\mathbf{X} \in \mathbb{R}^d$ and predetermined class $t \in \{0, 1\}$. A classifier consists of two probability functions $P(Y = 1|\mathbf{X})$ and $P(Y = 0|\mathbf{X})$ telling how likely it is that the class Y of a new object belongs to class 0 or 1. In order to train such a classifier, we need some kind of assumption on its structure. *Logistic regression* assumes that the logarithm of the so-called *odds* is a linear function of $\mathbf{X}$. The *odds* of a two-class probability function are defined as $\frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$ or accordingly $\frac{P(Y=1|X)}{1-P(Y=1|X)}$. They lie in between 0 and ∞, with a number less than 1 favoring $Y = 0$ and a number greater than 1 favoring $Y = 1$. In the context of logistic regression, we consider the logarithm of the odds or *logits* with $\ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \mathbf{w}^\top\mathbf{X}$ for certain parameters $\mathbf{w} \in \mathbb{R}^d$. The logits are point symmetric at $(0.5, 0)$, that is, if $\ln\left(\frac{P(Y_i=1|X)}{1-P(Y_i=1|X)}\right) = y$ for an object i and $\ln\left(\frac{P(Y_j=1|X)}{1-P(Y_j=1|X)}\right) = -y$ for an object j then $P(Y_i = 1|\mathbf{X}) = 1 - P(Y_j = 1|\mathbf{X})$. For convenience, we write $P(Y)$ instead of $P(Y = 1)$ from now on. Thus our classifier corresponds to the *logistic sigmoid* function:

$$\begin{aligned}
P(Y|\mathbf{X}) &= P(Y|\mathbf{X})\frac{\frac{1}{1-P(Y|X)}}{\frac{1}{1-P(Y|X)}} = \frac{\frac{P(Y|X)}{1-P(Y|X)}}{\frac{1-P(Y|X)+P(Y|X)}{1-P(Y|X)}} \\
&= \frac{\frac{P(Y|X)}{1-P(Y|X)}}{1+\frac{P(Y|X)}{1-P(Y|X)}} = \frac{\exp(\mathbf{w}^\top\mathbf{X})}{1+\exp(\mathbf{w}^\top\mathbf{X})}.
\end{aligned}$$

A good choice of the parameters leads to a small classification error. A typical approach is to find parameters $\mathbf{w}$ such that the *likelihood* for the known data set is maximized [1]. The likelihood function of a data set with N elements $\{\mathbf{X}, \mathbf{t}\}$ is given by $P(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} P(Y = t_n|\mathbf{X}_n)$, that is, it represents the probability for the observed data $\mathbf{X}$ given the assumption that the parameters are chosen to be $\mathbf{w}$. Since we are only considering two-class logistic regression for now, we may also partition the factors of the likelihood into two disjoint sets $\prod_{t_n=0}(1 - P(Y = 1|\mathbf{X}_n)) = \prod_{n=1}^{N}(1 - P(Y = 1|\mathbf{X}_n))^{1-t_n}$ and $\prod_{t_n=1} P(Y = 1|\mathbf{X}_n) = \prod_{n=1}^{N} P(Y = 1|\mathbf{X}_n)^{t_n}$. Thus, we are able to combine both products into the mathematically convenient form $P(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} P(Y|\mathbf{X}_n)^{t_n} \cdot (1 - P(Y|\mathbf{X}_n))^{1-t_n}$. Now, instead of maximizing this function, we minimize the negative logarithm of it, leading to the *cross entropy error function*:

$$\begin{aligned}
E(\mathbf{w}) &:= -\ln P(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} t_n \cdot \ln\ (P(Y|\mathbf{X}_n)) + (1-t_n)\cdot\ln(1-P(Y|\mathbf{X}_n)) \\
&= -\sum_{n=1}^{N} t_n \cdot \ln\ \left(\frac{\exp(\mathbf{w}^\top\mathbf{X}_n)}{1+\exp(\mathbf{w}^\top\mathbf{X}_n)}\right) + (1-t_n)\cdot\ln\left(1-\frac{\exp(\mathbf{w}^\top\mathbf{X}_n)}{1+\exp(\mathbf{w}^\top\mathbf{X}_n)}\right) \\
&= \sum_{n=1}^{N} t_n \cdot (-\mathbf{w}^\top\mathbf{X}_n + \ln\ (1+\exp(\mathbf{w}^\top\mathbf{X}_n)) + (1-t_n)\cdot\ln(1+\exp(\mathbf{w}^\top\mathbf{X}_n)) \\
&= \sum_{n=1}^{N} \mathbf{w}^\top\mathbf{X}_n \cdot (-t_n) + \ln(1+\exp(\mathbf{w}^\top\mathbf{X}_n)) \qquad (1)
\end{aligned}$$

ℓ_2 and ℓ_1 Regression

The ℓ_p norm of a d-dimensional vector $\mathbf{x}$ is defined as $||\mathbf{x}||_p = \sqrt[p]{x_1^p + x_2^p + \ldots x_d^p}$. The linear ℓ_2 regression for an N-sized data set is given an input matrix $\mathbf{A} \in \mathbb{R}^{N\times d}$, a target vector $\mathbf{b} \in \mathbb{R}^N$ and adjustable parameters $\mathbf{x} \in \mathbb{R}^d$ and consists of computing $\min_{\mathbf{x}\in\mathbb{R}^d} ||\mathbf{b} - \mathbf{Ax}||_2$. When facing large datasets, a possible approach to determine approximate solutions is via random sign matrices $\mathbf{S} \in \mathbb{R}^{N\times k}$, where k is the "new", reduced size of the dataset. Therefore the problem now consists of determining bounds to k such that $||\mathbf{S}^\top(\mathbf{b} - \mathbf{Ax}')||_2 \approx ||\mathbf{b} - \mathbf{Ax}||_2$. Generally, such algorithms are also given constants $0 < \epsilon, \delta < 1$ and bound k such that $\min_{\mathbf{x}\in\mathbb{R}^d} ||\mathbf{b} - \mathbf{Ax}||_2 \leq (1+\epsilon)\min_{\mathbf{x}\in\mathbb{R}^d} ||\mathbf{S}^\top(\mathbf{b} - \mathbf{Ax})||_2$

with probability at least $1-\delta$. A very successful approach described by Sarlós [4] and Clarkson and Woodruff [2] uses Johnson-Lindenstrauss type embeddings [3], that is, they project $\mathbf{A}$ into a subspace $\mathbf{S}^\mathsf{T}\mathbf{A}$ with small error. It should be noted that the above definition of the cross entropy error function corresponds to the ℓ_1 norm, rather than ℓ_2, though it can be argued that the choice between norms is a modeling question. Nevertheless, similar results have recently been published for ℓ_1 regression by Sohler and Woodruff [5].

Our Approach

The ℓ_2 and ℓ_1 regression results are heavily dependent on linear algebra, which is not immediately applicable for logistic regression. Therefore, we linearize the cross entropy error function for subsequent embeddings. Specifically, we construct a function G consisting of k linear functions such that for any choice of $\mathbf{w}$ we have $G(\mathbf{w}) \leq (1+\epsilon)E(\mathbf{w})$ with $k \in O(\sqrt{\epsilon^{-1}}\ln(1/\epsilon))$.

The only non-linear term in Eq. (1) is $f(\mathbf{w}) = \ln(1+\exp(\mathbf{w}^\mathsf{T}\mathbf{X}_n))$. With l'Hôpital's rule we have $\lim_{\mathrm{w}^\mathsf{T}\mathrm{X}_n \to \infty} f(\mathbf{w}) - \mathbf{w}^\mathsf{T}\mathbf{X}_n = 0$ and $\lim_{\mathrm{w}^\mathsf{T}\mathrm{X}_n \to -\infty} f(\mathbf{w}) = 0$. The interval where $\ln(1+\exp(\mathbf{w}^\mathsf{T}\mathbf{X}_n))$ is not within ϵ-distance of either $\mathbf{w}^\mathsf{T}\mathbf{X}_n$ or 0 is $]\ln(\exp(\epsilon)-1), \ln(\frac{1}{\exp(\epsilon)-1})[$. Since $\exp(\epsilon)-1 \geq \epsilon$ we can bound the size of the interval by $\ln(\frac{1}{\exp(\epsilon)-1}) - \ln(\exp(\epsilon)-1) = 2\ln(\frac{1}{\exp(\epsilon)-1}) = \Theta(\ln(1/\epsilon))$. The first Taylor remainder at x, that is, the difference between the first Taylor expansion (tangent) at $(a, f(a))$ and the original function $f(x) = \ln(1+\exp(x))$ in Lagrange form is given by

$$R_1(x) = \frac{f^{(1+1)}(\xi)}{(1+1)!}(x-a)^{1+1} = \frac{f''(\xi)}{2}(x-a)^2,$$

for ξ in between x and a. If $\epsilon \leq R_1(x)$, then $(x-a) \in \Omega(\sqrt{\epsilon})$ since $f''(\xi) = \frac{\exp(\xi)}{(1+\exp(\xi))^2} \leq \frac{1}{4}$. We therefore require at most $\Theta(\ln(1/\epsilon))/\Omega(\sqrt{\epsilon}) = O(\sqrt{\epsilon^{-1}}\ln(1/\epsilon))$ functions to linearize f for an $(1+\epsilon)$ approximation. We now give a first lower bound for k. Since $f''(x) = f''(-x)$, $f''(x)$ is monotonically increasing up for $x \leq 0$, and $\xi \geq \ln(\exp(x)-1)$ we have

$$\begin{aligned}
\epsilon = R_1(x) &= \frac{f''(\xi)}{2}(x-a)^2 \geq \frac{\exp((\epsilon)-1)\cdot(x-a)^2}{\exp(\epsilon)^2\cdot 2} \\
\Rightarrow \sqrt{\frac{\epsilon}{\exp(\epsilon)-1}}\cdot\sqrt{2}\exp(\epsilon) &\geq (x-a) \\
\Rightarrow \sqrt{\frac{\epsilon}{\epsilon}}\cdot\sqrt{2}\exp(\epsilon) &\geq (x-a) \\
\Rightarrow O(\exp(\epsilon)) &= (x-a)
\end{aligned}$$

This amounts to at least $\Theta(\ln(1/\epsilon))/O(\exp(\epsilon)) = \Omega(\ln(1/\epsilon))$ functions necessary for an for an $(1+\epsilon)$ approximation.

Open Questions

The immediate question is whether the upper bound for the space required for linearization is tight for instance by giving better characterizations of ξ. Also, until now we considered the cross-entropy error function (log likelihood) in our approximations. Though mathematically more convenient, any additive approximation for the log likelihood results in a multiplicative approximation for the likelihood. Whether we can do any better for the likelihood remains to be seen.

Thereafter, we will focus on more general models such as multiclass logistic regression, probit regression and Markov random fields [1] [6].

Furthermore, the parameters of logistic regression are usually determined via the method of steepest decent [1]. In this general (non-streaming) context, we will explore the possibilities of calculating approximate solutions with faster methods.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[2] K. L. Clarkeson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41th ACM Symposium on Theory of Computing (STOC)*, pages 341–350, 2010.

[3] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

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

[5] C. Sohler and D. P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *Proceedings of the 43th ACM Symposium on Theory of Computing (STOC)*, pages 755–764, 2011.

[6] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. *Introduction to Statistical Relational Learning*, 2006.

Efficient Algorithms for Bayesian Regression

Alexander Munteanu
Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
alexander.munteanu@tu-dortmund.de

The aim of our work is to develop algorithms for efficient computation of Bayesian regression models on large datasets with emphasis on the underdetermined high-dimensional case. In this technical report we describe the Bayesian regression model, provide an overview of techniques and algorithms used for similar regression problems and give a description of our approach.

Preliminaries

Let $||A|| = \left(\sum_{i,j} A_{ij}^2\right)^{1/2}$ denote the Frobenius norm for any real-valued matrix A. Note that it coincides with the well-known Euclidean ℓ_2-norm in the special case of A being a vector. Let I denote the identity matrix. Let $\mathcal{N}(x|\mu, \Sigma)$ denote the (multivariate) Gaussian distribution over the vector x with mean μ and covariance matrix Σ.

Bayesian Regression

Suppose we are given n high-dimensional observations $x_i \in \mathbb{R}^d$, $d \gg n$ and their corresponding target values t_i. Following [2] we assume that each t_i is given by a linear mapping of x_i with additive Gaussian noise, i.e. $t_i = x_i^T w + \epsilon$, where ϵ is a zero mean Gaussian distributed value with constant variance σ^2 and $w \in \mathbb{R}^d$ denotes a parameter vector. Then we can express the distribution of t_i as $p(t_i|w, x_i, \sigma^2) = \mathcal{N}(t_i|x_i^T w, \sigma^2)$. Subsequently the Likelihood of the observed data, which is assumed to be drawn independently from an unknown distribution, is given by $\mathcal{L}(t|X, w, \sigma^2) = \prod_{i=1}^n \mathcal{N}(t_i|x_i^T w, \sigma^2) = \mathcal{N}(t|Xw, \sigma^2 I)$. Note that the only unknown parameter in this distribution is the parameter vector w.

The Bayesian treatment of linear regression consists of computing a distribution over the parameter space rather than a single parameter vector, as it is done in the traditional maximum likelihood and in similar approaches. Given an initial informative or uninformative prior distribution $\pi_{pre}(w)$ over the weight vectors, say, for convenience of presentation, $\pi_{pre}(w) = \mathcal{N}(w|0, \sigma^2 I)$, the posterior distribution (after the observations have been considered) is computed by the following formula:

$$\pi_{post}(w) = \frac{\mathcal{L}(t|X, w, \sigma^2)\, \pi_{pre}(w)}{\int \mathcal{L}(t|X, v, \sigma^2)\, \pi_{pre}(v)\, \mathrm{d}v}.$$

The evaluation of the integral in the denominator is obviously the computationally challenging part. In many cases the integral can not even be determined analytically. Markov-Chain-Monte-Carlo algorithms solve this problem by defining a Markov-Chain over the parameter space and simulating a random walk which converges to the desired posterior distribution, such that it enables us to sample points directly from the posterior without actually computing the integral. The integral itself can be approximated numerically with these methods [2]. But still a huge number of evaluations of the likelihood function are needed to perform a proper approximation, which is computationally expensive and not practically treatable for large-scale data. Ignoring some constant factors (which include the variance of the underlying distributions) the problem can be simplified to evaluating for every w sampled by the algorithm

$$||Xw - t||^2 + ||w||^2, \tag{1}$$

i.e. the (simplified) negative natural logarithm of $\mathcal{L}(t|X, w, \sigma^2)\, \pi_{pre}(w)$, for every w sampled by the algorithm. Clearly the exact computation needs $O(Nnd)$ time for N samples due to the matrix-vector product Xw.

Bayesian Regression for Massive Data Sets

There are several works which cope with huge data sets with so called sequential Markov-Chain-Monte-Carlo algorithms [1, 5]. They load as much data into memory as possible and compute some samples on the reduced data set. Then they iterate over the remaining data and keep track of their importance to the whole model. When it is detected that some weights do not fit any more, a resampling step is done from the whole data seen so far. Balakrishnan and Madigan avoid reconsidering already seen data and thus make their algorithm a so called one-pass algorithm [1]. Note that the notion of a massive data set refers to the presence of many observed data vectors rather than to the case of very high-dimensional data. To our knowledge, the latter has not been studied so far for Bayesian regression.

Random Linear Embeddings and Streaming Algorithms for Matrix Multiplication

Johnson and Lindenstrauss [4] have shown that any finite set of n points from $\mathbb{R}^d$ can be mapped into $\mathbb{R}^k$ for some $k = O(\epsilon^{-2}\log(n))$, where $\epsilon > 0$ denotes an approximation parameter, such that the pairwise ℓ_2-distances are preserved up to a factor of $1 \pm \epsilon$. Thus, computations which involve the ℓ_2-norm of some vectors can be done approximately in a low-dimensional space with arbitrary precision.

As shown by Sarlós [6] and recently improved by Clarkeson and Woodruff [3], similar techniques can be used to approximate matrix multiplication in the turnstile streaming model, where the matrices A and B are given in a data stream of updates to single entries and each entry can be changed several times and in arbitrary order throughout the stream. Additionally the update stream may only be read once.

Let $A \in \mathbb{R}^{r\times c_1}$, $B \in \mathbb{R}^{r\times c_2}$ be given in the previously described model and let $c = c_1 + c_2$. The problem is to compute a matrix $C \in \mathbb{R}^{c_1\times c_2}$ such that

$$||A^T B - C|| \leq \epsilon \, ||A|| \, ||B||. \tag{2}$$

They use random sign matrices $S \in \{-1, 1\}^{r\times m}$ whose entries are drawn independently from the Rademacher distribution, i.e. each entry is chosen uniformly from $\{-1, 1\}$. These can be seen as Johnson-Lindenstrauss type of embeddings. It can be shown that

$$\frac{1}{m}\mathbf{E}(A^T S S^T B) = \frac{1}{m} A^T \mathbf{E}(S S^T) B = \frac{1}{m} A^T (mI) B = A^T B$$

and appropriate variance bounds imply that choosing $C = \frac{1}{m} A^T S S^T B$ and $m = O(\epsilon^{-2})$ yields the desired approximation bound given by (2) with high (constant) probability [3].

In the streaming context it is a common approach to maintain a sketch of the seen data of at most polylogarithmic size and do the needed computations on the sketch and output the result as a solution to the original data. For matrix multiplication the low-dimensional sketches are $A^T S \in \mathbb{R}^{c_1\times m}$ and $S^T B \in \mathbb{R}^{m\times c_2}$ of total size of $O(c\epsilon^{-2}\log(rc))$ bits. These can be updated in time $O(\epsilon^{-2})$ for every single update coming from the stream.

Our Approach

We intend to consider the parameter vectors w sampled by the Markov-Chain-Monte-Carlo algorithm as a stream of updates to w. Then we can maintain a low-dimensional sketch $S^T w$ of the vector and another sketch of X, given by XS which only has to be computed once as a preprocessing step since X never changes throughout a run of the algorithm. Instead of evaluating (1) exactly, we compute

$$||XSS^T w - t||^2 + ||S^T w||^2,$$

which will give an appropriate approximation to the exact value of (1). Since the dimensions of the sketches are $n \times m$ and m respectively, only $O(nm)$ time is needed instead of $O(nd)$, where $m = O(\epsilon^{-2})$. Note that in general the sampling of a new parameter vector can affect every single component which, in our streaming context with single component updates unfortunately leads to an update time of $\Omega(d)$ in each round. But there are classes of Markov-Chain-Monte-Carlo algorithms which change only a few, say k, of the entries in every round. If we can assume $k = O(1)$ like it can be done in the case of the well-known component-wise Metropolis-Hastings algorithm [2] or in the case of Gibbs algorithm [2], then the update time can still be bounded by a constant.

Open Questions

We are exploring the possibility of running Markov-Chain-Monte-Carlo sampling directly in the low-dimensional space instead of projecting the high-dimensional samples. It seems that the sampling distribution must be preserved under the projection, in the sense that the probability to choose a certain vector v from the low-dimensional space is (approximately) the same as the probability that v is the image of a sampled vector x from the original, high-dimensional space.

Further we investigate random embeddings for other (families of) distributions than the strictly Gaussian model described in this report.

References

[1] S Balakrishnan and D Madigan. A one-pass sequential monte carlo method for bayesian analysis of massive datasets. *Bayesian Analysis*, 2006.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[3] K. L. Clarkeson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41th ACM Symposium on Theory of Computing (STOC)*, pages 341–350, 2010.

[4] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26::189–206, 1984.

[5] G Ridgeway and D Madigan. A sequential monte carlo method for bayesian analysis of massive datasets. *Data Mining and Knowledge Discovery*, 2003.

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